18e Conférence en Recherche d'Information et Applications
16e Rencontres Jeunes Chercheurs en RI
30e Conférence sur le Traitement Automatique des Langues Naturelles
25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues
(CORIA-TALN) [1]

Actes de CORIA-TALN 2023.

Actes des 16e Rencontres Jeunes Chercheurs en RI (RJCRI) et 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL)

Marie Candito, Thomas Gerald, José G Moreno (Éds.)

Paris, France, 5 au 9 juin 2023

1. https://coria-taln-2023.sciencesconf.org/

Avec le soutien de

Préface

Organisée conjointement par les laboratoires franciliens sous l'égide de l'Association francophone de Recherche d'Information et Applications (ARIA) et l'Association pour le Traitement Automatique des Langues (ATALA), la conférence CORIA-TALN-RJCRI-RECITAL 2023 regroupe :

— la 18e Conférence en Recherche d'Information et Applications (CORIA)
— la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN) ;

ainsi que les deux conférences associées, destinées aux jeunes chercheuses et chercheurs :
— Les 16e Rencontres Jeunes Chercheurs en RI (RJCRI)
— la 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL)

Ces deux derniers évènements ont été nommés cette année les "Rencontres Jeunes Chercheurs", et ont fait l'objet d'un processus de relecture et de sélection unique. Nous avons reçu 13 soumissions, dont 11 ont été acceptées (à comparer pour RECITAL en 2022, 15 soumissions et 9 articles acceptés).

Lors de la conférence, qui se tient à Paris, du 6 au 9 juin 2023, les articles RJC sont présentés dans des sessions communes aux conférences CORIA et/ou TALN. Quatre des onze articles RJC sont présentés oralement, et les sept autres sous forme de poster.

Comme c'est le propre des Rencontres Jeunes Chercheurs, cette édition conjointe RJCRI/RECITAL a de nouveau été l'occasion pour de jeunes chercheuses et chercheurs, de se former à l'écriture d'articles scientifiques, de diffuser leurs travaux en cours et d'obtenir des avis de la communauté scientifique, tant lors du processus de relecture que, pour les articles acceptés, lors de leur présentation pendant la conférence.

Nous tenons à remercier chaleureusement les membres du comité de programme des RJC, pour leurs relectures exigeantes mais constructives. Nous espérons que celles-ci ont permis aux autrices et auteurs ayant soumis aux RJC d'avancer dans leurs travaux.

Un grand merci également à toute l'équipe d'organisation de l'édition 2023 ainsi qu'aux sponsors, pour tout le travail réalisé et le soutien financier afin que ces rencontres soient une réussite.

Marie Candito, Thomas Gérald et José G. Moreno, co-présidents des RJC 2023

Comités

Comité de programme

Présidence

— Marie Candito
— Thomas Gerald
— José Moreno

Membres

— Catherine Berrut
— Sylvie Calabretto
— Thierry Charnois
— Maximin Coavoux
— Kata Gabor
— Aina Garí Soler
— Gaël Guibon
— Paul Lerner
— Jesus Lovon
— Aurélie Névéol
— Yannick Parmentier
— Karen Pinel-Sauvagnat
— Carlos Ramisch
— Arnaud Soulet
— Julien Velcin

Table des matières

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Les jeux de données en compréhension du langage naturel et parlé : paradigmes d'annotation et représentations sémantiques

Rim Abrougui[1, 2]
(1) Orange Innovation, Lannion, France
(2) Aix-Marseille Université, LIS UMR 7020, Marseille, France
rim.abrougui@orange.com

RÉSUMÉ

La compréhension du langage naturel et parlé (NLU/SLU) couvre le problème d'extraire et d'annoter la structure sémantique, à partir des énoncés des utilisateurs dans le contexte des interactions humain/machine, telles que les systèmes de dialogue. Elle se compose souvent de deux tâches principales : la détection des intentions et la classification des concepts. Dans cet article, différents corpora SLU sont étudiés au niveau formel et sémantique : leurs différents formats d'annotations (à plat et structuré) et leurs ontologies ont été comparés et discutés. Avec leur pouvoir expressif gardant la hiérarchie sémantique entre les intentions et les concepts, les représentations sémantiques structurées sous forme de graphe ont été mises en exergue. En se positionnant vis à vis de la littérature et pour les futures études, une projection sémantique et une modification au niveau de l'ontologie du corpus MultiWOZ ont été proposées.

ABSTRACT

The Challenges of Spoken Language Understanding Datasets : A Study on Annotations and Semantic Representations

Natural and Spoken Language Understanding (NLU/SLU) covers the problem of extracting and annotating the meaning structure from user utterances in the context of human/machine interaction, such as dialogue systems, consisting oftenly of two main tasks : intent detection and slot filling. In this paper, different SLU corpora were studied at a formal and semantic level : their different annotation formats (flat and structured) and ontologies were compared and discussed. With their expressive power maintaining the semantic hierarchy between intents and slots, graph semantic representations were highlighted. In line with the literature and for future studies, a semantic projection and a modification of the ontology of the MultiWOZ corpus were proposed.

MOTS-CLÉS : Compréhension du langage, ontologies, représentation sémantique à plat (BIO), représentation sémantique structurée (graphe).

KEYWORDS: Language Understanding, ontologies, flat semantic representation (BIO), structured semantic representation (graph).

1 Introduction

La compréhension du langage naturel et parlé est un sujet d'étude important dans le cadre des interactions homme-machine. Le domaine comprend plusieurs niveaux d'étude, mais actuellement la tâche de compréhension est principalement axée sur la compréhension de la sémantique globale

Actes des 16e Rencontres Jeunes Chercheurs en RI (RJCRI) et 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL), pages 1–20.

des requêtes des utilisateurs et sur l'identification des concepts génériques des mots-clés, à savoir, la détection des intentions et l'identification des concepts (Tur & De Mori, 2011).

Bien qu'il existe plusieurs approches de représentations sémantiques, la majorité des méthodes se base sur la représentation à base de frames sémantiques en utilisant des modèles supervisés pour la classification et l'étiquetage de séquence. Ces modèles qui sont basés sur des réseaux de neurones utilisant des modèles de langage pré-entraînés, ont obtenu des performances élevées sur plusieurs jeux de données SLU qui sont représentés avec un schéma à plat (Béchet & Raymond, 2019) (cf. figure 3).

Cependant, les interactions dans une conversation en conditions réelles sont beaucoup plus complexes. Afin de relever ces défis, il est nécessaire d'avoir d'un côté des corpus d'apprentissage dotés de représentations sémantiques complexes et contextuelles, et de l'autre côté des schémas d'annotation capables de prendre en compte les représentations sémantiques hiérarchiques. De plus, pour construire des modèles de SLU plus robustes et pour les comparer de manière plus précise, il est primordial d'unifier les ensembles de données existants. Cette unification permettra de diversifier les domaines et de fournir plus de connaissances aux systèmes de compréhension du langage qui pourront par conséquent apprendre plusieurs structures sémantiques.

Il existe plusieurs études complètes sur la compréhension du langage naturel et parlé, telles que (Weld *et al.*, 2022) et (Qin *et al.*, 2021), mais cet article explorera plus la question des jeux de données en étudiant leurs ontologies et formats d'annotation et de représentation sémantique. Dans le cadre de la problématique d'unification de toutes les ressources publiques d'apprentissage et d'évaluation des systèmes NLU/SLU, nous avons comparé les différentes ontologies et schémas d'annotation. Nous mettons ainsi en exergue le potentiel des représentations sémantiques structurées qui peuvent être utilisées plus facilement avec le développement des modèles de génération du langage. En s'intéressant particulièrement aux représentations en graphe qui préservent la hiérarchie et le lien sémantique entre les différents labels, nous proposerons une projection sémantique et une modification au niveau de l'ontologie du corpus MultiWOZ2.3.

La présentation des jeux de données et la comparaison de leurs ontologies sont présentées dans la section 2, alors que l'étude des schémas d'annotation et des représentations sémantiques structurées ainsi que nos perspectives sont exposées dans la section 3.

2 Exploration des jeux de données en compréhension du langage

Les jeux de données jouent un rôle crucial dans l'avancement de la recherche dans le domaine de la compréhension du langage. Ils permettent en effet d'entraîner, d'évaluer et de comparer les systèmes SLU. Les corpora disponibles sont variés et peuvent couvrir plusieurs domaines. La façon dont les annotations sont représentées et le choix des labels reflètent une certaine variation au niveau des ontologies ce qui rend difficile l'unification de ces jeux de données. Les schémas de projection et des représentations sémantiques choisis affectant la qualité d'annotation peuvent être aussi différents. Certains corpora sont des conversations complètes multi-domaines et/ou multi-intentions, tandis que d'autres ne contiennent que des simples requêtes. Malgré cette diversité, un corpus large avec des conversations complètes entièrement annotées en logique SLU selon une ontologie générique applicable à toutes les données est difficile à trouver, ce qui rend les travaux sur l'exploitation de l'histoire conversationnelle et du contexte plus difficile.

Nous pouvons trouver également plusieurs types de collectes de ces corpus. L'une de ces méthodes est appelée la méthode «`Wizard-Of-OZ`» où les participants interagissent en temps réel avec un système qu'ils croient être autonome, mais il est contrôlé en réalité par un opérateur humain invisible. Cette méthode est plus fréquente puisqu'elle permet de collecter des données de manière contrôlée, avec une grande variété de scénario de dialogues. Les transcriptions de la parole à l'aide de la reconnaissance automatique de la parole (ASR), sont un autre moyen de collecter des données. La qualité des transcriptions va dépendre des systèmes ASR mais cette méthode permet de collecter d'une façon plus rapide les énoncés dans des conditions réelles et qui vont être vérifiés et annotés manuellement ou d'une manière semi-automatique. Nous trouvons enfin des méthodes plus automatisées qui consistent à synthétiser les conversations à l'aide de modèles de langage ou de modèles de génération de texte. Cette méthode est la plus rapide mais elle est artificielle car elle ne peut pas refléter toutes les variations linguistiques dans la parole humaine et peut avoir des problèmes d'hallucination. Cette section présentera certains jeux de données publiés pour la tâche SLU, ainsi que leurs ontologies et leurs méthodes d'annotation.

2.1 Les ensembles de données pour les tâches de compréhension du langage naturel et parlé

Jeux de données	Langues	Modes	#énoncés/dialogues	#labels
ATIS (**Hemphill *et al.*, 1990**)	en	requêtes	4978	17 intentions 84 slots
Frames (**Asri *et al.*, 2017**)	en	dialogues	1369	20 actes 16 slots
Massive (**FitzGerald *et al.*, 2022**)	multilingues 51 langues	requêtes	19521 par langue	60 intentions 55 slots
MEDIA (**Devillers *et al.*, 2004**)	fr	dialogues	1250	83 slots 19 spécifieurs
mTOD (**Schuster *et al.*, 2019**)	multilingues 3 langues	requêtes	43000	12 intentions 11 slots
mTOP (**Li *et al.*, 2020**)	multilingues 6 langues	requêtes	10000	117 intentions 78 slots
MultiDoGo (**Peskov *et al.*, 2019**)	en	dialogues	15000 annotés	85 intentions 73 slots
MultiWOZ (**Budzianowski *et al.*, 2018**)	en	dialogues	10438	32 actes 27 slots
M2M (**Shah *et al.*, 2018**)	en	dialogue	3000	15 actes 12 slots
SNIPS (**Coucke *et al.*, 2018**)	en	requêtes	14484	7 intentions 39 slots
TOP (**Gupta *et al.*, 2018**)	en	requêtes	44783	25 intentions 36 slots
The restaurant-8K dataset (**Coope *et al.*, 2020**)	en	requêtes	8198	5 slots
VocaDOM (**Portet *et al.*, 2019**)	fr	requêtes	4610	7 intentions 12 slots

TABLE 1 – Tableau récapitulatif des ensembles de données en NLU/SLU

Il existe de nombreux corpora SLU disponibles publiquement, chacun ayant ses propres caractéristiques et domaines d'application. Les tableaux 1 et 2 synthétisent les caractéristiques de ces données et nous détaillons ci-dessous chaque corpus.

Jeux de données	Multi-Domaines	Multi-Intents	Inter-Domaines	Annot. contextuelles	Annot. à plat	Annot. structurée ou semi-structurée
ATIS (Hemphill *et al.*, 1990)					X	
Frames (Asri *et al.*, 2017)				X		X
Massive (FitzGerald *et al.*, 2022)	X				X	
MEDIA (Devillers *et al.*, 2004)					X	
mTOD (Schuster *et al.*, 2019)	X				X	
mTOP (Li *et al.*, 2020)	X				X	X
MultiDoGo (Peskov *et al.*, 2019)	X	X			X	
MultiWOZ (Budzianowski *et al.*, 2018)	X	X	X	X		X
M2M (Shah *et al.*, 2018)	X					X
SNIPS (Coucke *et al.*, 2018)	X				X	
TOP (Gupta *et al.*, 2018)	X					X
The restaurant-8K dataset (Coope *et al.*, 2020)				X	X	
VocaDOM (Portet *et al.*, 2019)					X	

TABLE 2 – Tableau récapitulatif des caractéristiques des jeux de données en NLU/SLU

2.1.1 Jeux de données avec des requêtes simples

1. Ressources mono-lingues :
   (a) **ATIS** (Hemphill *et al.*, 1990) : Le corpus ATIS (Air Travel Information System) est l'un des corpus SLU les plus utilisés. Il contient des informations sur des compagnies aériennes et des commandes pour réserver des vols. La première version a été collectée suivant l'approche «`Wizard-Of-OZ`», mais les auteurs ont utilisé des transcriptions ASR pour les autres versions (Dahl *et al.*, 1994). Les premières annotations de ce corpus ont été effectuées à l'aide d'une requête SQL, puis ont été transférées au niveau global sous forme d'intentions, ainsi qu'au niveau mot sous forme des concepts (Béchet & Raymond, 2018). Ce corpus est en anglais et contient 4978 énoncés annotés en frame avec 17 intentions et 84 concepts.
   (b) **Frames** (Asri *et al.*, 2017) : Il s'agit d'un corpus avec des interactions humain-humain en anglais et qui contient des informations sur les réservations d'hôtel. Il a été publié pour encourager les recherches sur les systèmes conversationnels textuels. La notion de «mémoire» et l'exploitation de l'histoire conversationnelle ont été les premières questions abordées, où les auteurs ont rajouté à la tâche NLU, la tâche de "suivi des frames sémantiques". Des références et des identifiants des frames sémantiques ont été donc rajoutés aux annotations en acte de dialogue et en slot-valeur. Ce corpus peut être utilisé dans les tâches de compréhension et dans les tâches de suivi d'état de dialogue. 1369 dialogues ont été collectés suivant l'approche «`Wizard-Of-OZ`» . Les énoncés au niveau utilisateurs et systèmes ont été annotés avec 20 types d'acte de dialogue et 16 types de slot.
   (c) **MEDIA** (Devillers *et al.*, 2004) : Le corpus MEDIA contient des informations touristiques en français. Il a été collecté suivant l'approche «`Wizard-Of-OZ`». 1250 dialogues

ont été transcrits et annotés manuellement suivant une ontologie très riche au niveau des énoncés de l'utilisateur. Nous retrouvons 83 concepts de base regroupés dans un dictionnaire sémantique et 19 "spécifieurs" pouvant leur être associés. En lien avec le corpus MEDIA, PORT-MEDIA (Lefevre *et al.*, 2012) a été publié en français et en italien. Les méthodes de collecte étaient similaires en rajoutant des scénarios de dialogue variés. Les annotations étaient semi-automatiques à partir des systèmes de compréhension entraînés sur MEDIA suivies d'une vérification et correction manuelle. La version italienne a été générée par des traductions automatiques de la version française. Les questions de la complexité sémantique et les différents niveaux hiérarchiques ont été creusées et traitées sur une base linguistique solide.

(d) **SNIPS** (Coucke *et al.*, 2018) : SNIPS est un corpus en anglais qui contient plusieurs domaines différents. Les données ont été collectées par des transcriptions ASR suivi d'une annotation et vérification manuelles. Il contient 7 intentions et 39 concepts. La même version de ce corpus a été publiée en d'autres langues comme le français et l'allemand. SNIPS a été l'origine d'un autre corpus, **Almawave SLU** (Bellomaria *et al.*, 2019), le premier ensemble de données en italien pour les expériences SLU. Il a été généré d'une manière semi-automatique en traduisant les énoncés et les labels et en remplaçant les entités ouverts (comme les noms de restaurants et des livres) par des références italiennes. La vérification et la correction manuelles a été effectuée pour les énoncés.

(e) **TOP** (Gupta *et al.*, 2018) : Ce corpus a été publié pour étudier les problématiques sémantiques plus complexes. Les auteurs ont introduit une représentation hiérarchique appelée "*Task Oriented Parsing*" (TOP) pour les systèmes de dialogue basés sur des intentions et des concepts. Les énoncés ont été collectés par *crowd-sourcing* et annotés par deux annotateurs, un troisième annotateur peut intervenir en cas de désaccord. 44783 annotations avec 25 intentions et 36 slots ont été obtenues. Une version étendue du corpus avec 6 domaines supplémentaires a été publié dans **TOPv2** (Chen *et al.*, 2020).

The restaurant-8K dataset (Coope *et al.*, 2020) : Pour renforcer le travail d'extraction de concepts dans le cadre de dialogues, ce jeu de donnée qui comprend des conversations d'un système de réservation de restaurant, a été introduit. 8198 énoncés d'utilisateurs réels interagissant avec un système de dialogue déployé dans le domaine de la réservation de restaurants ont été annotés d'une manière contextuelle indiquant quels concepts ont été demandés par le système. Les réponses de système ne sont pas incluses dans l'ensemble des données, et il n'y a que 5 concepts.

VocaDOM (Portet *et al.*, 2019) : Afin de soutenir les tâches dans le cadre des systèmes *"Smart Home*" comme l'identification du locutaire, la reconnaissance de la parole et les tâches SLU, ce corpus a été publié en rassemblant des interactions dans des conditions réelles de 11 participants dans une maison intelligente, la méthode «`Wizard-Of-OZ`» était la base de ce protocole. Les enregistrements ont été transcrits et annotés manuellement par des intentions et des concepts. Au total, le corpus contient 4610 énoncés en français étiquetés par 7 intentions et 12 concepts. Dans (Desot *et al.*, 2018), un jeu de donnée synthétiques du même domaine a été généré automatiquement à partir du corpus VocaDOM.

2. Ressources multi-lingues :

(a) **Massive** (FitzGerald *et al.*, 2022) : Massive est un corpus multilingue qui contient des requêtes appartenant à 18 domaines. Sa publication a été motivée par le manque des jeux de données en plusieurs langues pour évaluer les modèles multilingues. Le corpus

SLURP (Bastianelli *et al.*, 2020), publié pour développer un assistant robotique personnel à domicile et pour des expériences SLU *End-to-end*, est la version d'origine du Massive. La version du corpus SLURP disponible publiquement est textuelle en anglais, elle a été collectée par les travailleurs de «`Mechanical Turk`» (AMT) et annotée manuellement au niveau "scénario" (domaine), "action" (Intention) et "entités" (concepts). Des traducteurs professionnels ont traduit les énoncés du corpus en 51 langues et ont également vérifié les frontières des concepts sur les tokens, aboutissant à la création du corpus Massive. Ce dernier est considéré comme une grande source des intentions (60) et de concepts (55) vu la diversité des domaines.

(b) **mTOD** (Schuster *et al.*, 2019) : Il s'agit d'un ensemble de données multilingue qui permet d'étudier les méthodes d'apprentissage par transfert inter-linguistique. Ce corpus offre l'opportunité d'étudier les modèles sémantiques inter-langues et constitue le premier ensemble de données parallèles pour une tâche d'étiquetage de mots qui a été annoté selon les mêmes guides d'annotation dans plusieurs langues. Les auteurs ont collecté 43000 énoncés en anglais dans les domaines *ALARM*, *REMINDER*, et *WEATHER*, et ils ont demandé à des anglophones natifs de proposer des labels d'intentions utilisées par deux annotateurs pour étiqueter les énoncés et les valeurs par des concepts. Cette annotation a été vérifiée ensuite par un troisième annotateur. Des locuteurs natifs en espagnol et en thaï ont traduits les énoncés qui ont été aussi annotés par deux annotateurs. Il contient au total 12 types d'intentions et 11 concepts

(c) **mTOP** (Li *et al.*, 2020) : En creusant la même problématique de la sémantique compositionnelle mise en évidence dans le corpus TOP (Gupta *et al.*, 2018) et en suivant la même logique de représentation hiérarchique, le corpus mTOP, a été publié. Cet ensemble de données est le premier qui contient des représentations sémantiques compositionnelles qui permettent l'annotation des requêtes imbriquées. Il a été publié avec les deux versions d'annotation : une plate et une autre compositionnelle. Les auteurs ont commencé par collecter une version en anglais des données, suivant la même approche dans (Gupta *et al.*, 2018), qui est traduite ensuite par des traducteurs professionnels. Le corpus mTOP est plus grand que TOP où nous avons 100.000 exemples avec 6 langues différentes, 11 domaines, 117 intentions et 78 concepts. En outre, une version parlée (**STOP**) a été publiée dans (Tomasello *et al.*, 2023) à partir de **TOPv2** pour encourager les recherches sur les approches end-to-end tout en focalisant sur les problématiques des requêtes compositionnelles.

2.1.2 Jeux de données conversationnels

1. Ressources mono-lingues :

(a) **MultiDoGo** (Peskov *et al.*, 2019) : Cet ensemble de données a été collecté par «`crowd-sourcing`» dans le cadre du progrès des assistants virtuels et du manque des données pour leur développement. Ce corpus est composé par 81000 conversations, dont 15000 ont été annotées avec 6 domaines différents, 85 intentions et 73 slots. Dans le corpus disponible publiquement, les énoncés systèmes ne sont pas annotés. L'article présentant ces données a mis en valeur la possibilité d'avoir des multi-intentions en montrant que les annotations ont été réalisées par des experts selon deux niveaux : au niveau des tours des dialogues et au niveau des phrases, afin de garder l'ordre entre les énoncés coordonnées et leurs intentions. Toutefois, dans la version publiée, nous ne trouvons pas souvent cette illustration et nous pouvons même perdre le lien entre les différentes

intentions et leurs concepts.

(b) **MultiWOZ** (Budzianowski *et al.*, 2018) : Il s'agit d'un corpus de dialogue multi-domaines en anglais, à grande échelle, souvent utilisé pour plusieurs tâches, notamment le suivi de l'état du dialogue, la politique de dialogue et les tâches de génération de dialogue. Il a été collecté à partir de la méthode «`Wizard-Of-OZ`» via un «`crowd-sourcing`». La première version de ce corpus a été publiée dans le but de faciliter la construction de systèmes de dialogue supervisés. Chaque énoncé dans les dialogues est annoté avec une séquence d'acte de dialogue. Cependant, la première version comporte des erreurs d'annotations, surtout au niveau de l'utilisateur, puisqu'elles ont été effectuées automatiquement à partir des annotations système (Eric *et al.*, 2019). Plusieurs versions ont été produites pour corriger ces erreurs et simplifier le format des annotations. Les versions MultiWOZ2.2 (Zang *et al.*, 2020) et MultiWOZ2.4 (Ye *et al.*, 2021) sont mieux adaptées aux tâches de suivi de l'état du dialogue, tandis que la version MultiWOZ2.3 (Han *et al.*, 2020) a des annotations utilisateur plus précises. Le corpus contient 7 domaines, 32 actes de dialogue et 27 slots.

(c) **M2M** (Shah *et al.*, 2018) : M2M est une fusion de 2 données contenant des dialogues en anglais pour la réservation des restaurants et des tickets de cinémas. Les méthodes de collecte et de l'annotation ont été réalisées d'une manière automatique où un développeur de dialogue fournit un scénario et les chatbots génèrent des tours de conversations en les annotant par des actes de dialogue et par des slots. Ce processus était répété jusqu'à la fin des tours de dialogue soit par un acte "bye" soit en atteignant un seuil maximum de tours. Au total, nous avons 3000 dialogues annotés avec 15 actes de dialogues et 12 slots.

Dans la partie suivante, nous allons nous focaliser sur l'analyse de certains ensembles de données au niveau de leurs ontologies et schémas sémantiques.

2.2 Ontologies et Annotations

L'annotation sémantique repose généralement sur des ontologies basées sur des connaissances linguistiques permettant de définir des liens hiérarchiques entre les entités (Ma *et al.*, 2009). Dans les tâches de compréhension du langage, les ontologies permettent de décrire le lien sémantique entre les domaines, les intentions ou les actes de dialogue et leurs concepts (Loos, 2006). Ces ontologies se varient selon les schémas et les règles d'annotation des corpus, mais parfois, nous pouvons trouver la même logique surtout lorsque le schéma d'annotation est limité à un cadre sémantique simple. En d'autres termes, dans des jeux de données comme le cas d'ATIS (Hemphill *et al.*, 1990), de SNIPS (Coucke *et al.*, 2018), de Massive (FitzGerald *et al.*, 2022) et de mTOD (Schuster *et al.*, 2019), chaque énoncé est labélisé par une seule intention, la notion de multi-intentions ou le croisement entre les domaines (inter-domaines) sont donc absents pour ces corpora.

En plus, les données citées peuvent avoir des domaines en commun et donc des similarités au niveau des concepts. Par exemple, le corpus ATIS n'a qu'un seul domaine (réservation de vol) qui peut se croiser avec le domaine "airline" dans le corpus MultiDoGo (Peskov *et al.*, 2019). SNIPS a aussi plusieurs domaines en commun avec TOP (Gupta *et al.*, 2018) («`Restaurant, Weather, Music`»...). Cependant, les différentes façons d'exprimer les intentions et le choix des concepts entraînent une grande diversité au niveau des ontologies, ce qui rend leur unification assez problématique. Les intentions dans ATIS sont des noms simples, par contre dans mTOD et SNIPS elles sont composées

d'un acte de dialogue avec le domaine («Show_alarms», «BookRestaurant»). Les actes de dialogue dans MultiWOZ sont aussi composés par le domaine associé à l'acte («Hotel-Inform»), mais le choix des actes est associé à la sémantique derrière les concepts plus qu'au sens global des énoncés. Autrement dit, les concepts comme «Phone» et «Car» pour le domaine «Taxi» sont toujours associés à l'acte «Request».

Les concepts peuvent être aussi variés au niveau de leur composition, ils peuvent se représenter comme une seule entité («country») ou une entité composée («party_size_description»). Nous avons remarqué aussi que dans toutes ces données les concepts sont plus compositionnels et reliés à leurs domaines. Dans le corpus Massive, on a par exemple les slots «sport_type», «drink_type» et «alarm_type», à l'antipode de MultiWOZ qui n'a que le concept «Type» partagé par les domaines «Attraction» et «Hotel». Dans le corpus ATIS, les concepts présentent un niveau plus haut au niveau de sa composition où les prépositions peuvent faire partie des slots (comme «from» dans «fromloc.city_name»), ou nous pouvons même trouver des concepts imbriqués (le slot «depart_time.period_of_day» est composé par le slot «time» et le slot «period_of_day»).

Le corpus MEDIA (Devillers *et al.*, 2004) est par ailleurs assez particulier au niveau de son schéma d'annotation. L'ontologie a été basée sur un niveau sémantique haut qui essaie de relever le lien hiérarchique des labels sémantiques. Les frames sémantiques ne se composent pas que par des paires de slot-valeur, mais aussi par un spécifieur qui définit les relations entre les entités. Une annotation des modes a été aussi effectuée et attachée aux concepts. Ainsi, la représentation hiérarchique a été recomposée par la combinaison des "spécifieurs" et des concepts.

Il convient également de noter que ces ensembles de données varient considérablement en termes de complexité linguistique. Dans l'article (Bechet *et al.*, 2022), divers phénomènes linguistiques qui peuvent impacter les performances des systèmes SLU sont observés. Certains corpora ne reflètent pas les caractéristiques des interactions dans des conditions réelles, tandis que d'autres sont plus difficiles pour les modèles. L'approche proposée pour évaluer la qualité et comparer les corpora, comme décrite dans (Bechet *et al.*, 2022) et (Béchet & Raymond, 2019), peut contribuer à la question d'unification des données.

En ce qui concerne la méthode de projection des frames sémantiques, le format d'annotation "BIO" à plat (cf. tableau 3) est souvent le paradigme le plus utilisé, notamment pour les données avec de simples requêtes, afin de faire un étiquetage de séquence facilement avec les modèles pré-entraînés à base de Transformers de type Bert (Devlin *et al.*, 2018). Néanmoins, ce paradigme est limité si nous voulons passer à la compréhension des énoncés plus complexes en exploitant le contexte de la conversation. Nous avons remarqué les limites de ce paradigme avec le corpus MultiWOZ où un seul énoncé peut avoir plusieurs domaines ou plusieurs intentions (cf. 4). Les annotations d'origine de ce corpus se représentent sous un format plus structuré en format `json` où nous pouvons trouver le lien entre les différents concepts et leurs intentions. Une suggestion d'une annotation à plat a été proposée dans (Lee *et al.*, 2019), mais les difficultés des annotations en contexte, notamment pour les concepts sans informations d'empans, n'ont pas été entièrement résolues. Ces entités sont justement définies comme des "slots de catégories", où leurs valeurs se trouvent dans les énoncés précédents, ou bien implicitement dans le sens global de l'énoncé. Les slots `"Parking"` et `"Post"` dans la table 4 sont un exemple des concepts de catégories.

Les différents paradigmes d'annotation et les différentes hiérarchies sémantiques des ontologies mettent en valeur la difficulté de représenter les données d'une manière structurée où le contexte peut-être bien exploité. Dans la section suivante, cette question sera creusée où nous allons nous

Enoncés	[CLS]	I'm	traveling	to	dallas	from	philadelphia
Annotation	Flight	O	O	O	B-toloc.city_name	O	B-fromloc.city_name

TABLE 3 – Annotation à plat en BIO du corpus ATIS :"B" pour «Begining», "I" pour «Inside» et "O" pour «Outside»

Enoncés	**Annotation en acte de dialogue**
◇ I won't have a car, so parking isn't important	`"Hotel-Inform": [["Parking", "no"]]`
◇ Can I have the postcode for the attraction, I also need a Taxi	`"Attraction-Inform": [["Post", "?"]], "Taxi-Inform": [["none", "none"]]`

TABLE 4 – Exemples d'énoncés annotés dans le corpus MultiWOZ2.3

Enoncés	take grandma Jane off the call
Annotation	`[IN:update_call [SL:contact_removed [IN:get_contact [SL:type_relation grandma ] [SL:contact Jane ]]]]`

TABLE 5 – Exemple d'annotation dans le corpus mTOP

intéresser aux différentes motivations qui sous-tendent l'utilisation de représentations structurées des étiquettes sémantiques des données pouvant être une solution de certaines limites des annotations à plat.

3 Projections des annotations et représentations sémantiques structurées dans le NLU/SLU

Nous avons présenté dans la section précédente les différences et les similitudes entre les jeux de données utilisés pour les tâches de compréhension du langage au niveau de leurs ontologies et leurs schémas d'annotation. Dans cette section, nous allons retracer d'une manière générale l'historique des différentes représentations sémantiques utilisées en compréhension du langage, ainsi que les enjeux liés aux représentations actuelles dans le contexte des systèmes de dialogue. Nous étudierons ensuite les projections structurées en illustrant les problématiques liées aux différents schémas à plat, ainsi que le potentiel des formats structurés, en particulier ceux basés sur la méthode des graphes.

3.1 Représentations sémantiques

Les interprétations sémantiques peuvent être considérées comme un processus de traduction, réalisé par un parseur sémantique, entre les mots d'une phrase et les représentations sémantiques du langage, comme montré dans (Dinarelli, 2010). Les représentations sémantiques permettent la modélisation des énoncés et de leurs interprétations sémantiques pour les machines. Elles peuvent être sous forme de logique formelle (Zettlemoyer & Collins, 2012), des frames sémantiques (Dinarelli *et al.*, 2009) ou encore des graphes sémantiques (Banarescu *et al.*, 2013).

Avec le progrès des systèmes de dialogue et des modèles modernes basés sur des approches de

statistiques et de probabilités, les interprétations sémantiques en SLU sont principalement basées le plus souvent sur l'identification des intentions ou des actes de dialogue et des slots. En outre, ces annotations notamment au niveau des slots sont très similaires aux annotations par frame sémantique, comme dans FrameNet (Baker *et al.*, 1998), dans la mesure où la représentation SLU est basée sur des attributs, qui sont des unités sémantiques, instanciées par séquences de mots. Contrairement aux frames, les attributs en SLU n'ont pas besoin d'expliciter les relations sémantiques entre les éléments de la phrase (Dinarelli, 2010). Au niveau concepts, les étiquettes consistent à identifier les éléments clés de l'énoncé, comme les entités nommées, les dates ou les destinations. Quant aux intentions, les annotations sont utilisées pour aider les systèmes SLU à comprendre le but global de l'utilisateur. Nous pouvons trouver ainsi des annotations en acte de dialogue pour mieux représenter les intentions. Par ailleurs, des recherches ont été menées pour représenter les attributs sémantiques du dialogue sous forme d'une ontologie abstraite, générique et structurée en exploitant les représentations AMR pour l'analyse sémantique du dialogue. (Bonial *et al.*, 2020).

Il est important ainsi de réfléchir à la question de la projection des informations sémantiques pour les traduire en entrées exploitables par les modèles. L'approche la plus courante pour la tâche de compréhension du langage repose sur les approches de l'étiquetage de séquence. La projection à plat en format BIO facilite cette tâche notamment pour les approches jointes pour prédire les intentions et les concepts simultanément. Comme il est montré dans l'exemple 3, ces approches associent généralement l'intention au token de classification globale [CLS] et détectent les concepts sur chaque token concerné avec une étiquette B ou I lorsque cela est possible.

Bien que la projection basée sur le format BIO soit utile pour l'utilisation des modèles à base de Transformers, elle ne permet pas de tirer parti des structures sémantiques hiérarchiques et imbriquées. Ainsi, des recherches sur des annotations plus structurées ont été étudiées, notamment avec l'utilisation des approches de séquence à séquence (*seq2seq*). Les chercheurs dans (Li *et al.*, 2020) ont proposé une représentation composée découplée (cf. 5) pour représenter des intentions imbriquées dans les slots. De même les expériences dans (Hu *et al.*, 2022) présentent le NLU comme une tâche de génération des graphes composés par des nœuds pour les labels. Dans la section suivante, nous allons montrer les différentes motivations et illustrations des schémas de représentation structurée.

3.2 Les représentations structurées des frames sémantiques

Selon (Devillers *et al.*, 2004), la représentation sémantique des annotations d'un corpus a été définie comme un moyen pour représenter les frames sémantiques d'une manière générique et complète selon la tâche mais qui permet aussi l'annotation des corpora larges d'une manière simple. Par conséquent, les schémas de représentation à plat ont été le centre des annotations dans la majorité des copora publics. Cependant, les représentations hiérarchiques sont plus expressives et permettent le lien entre les sous-structures (Tur & De Mori, 2011).

Dans la section 2.2 nous avons présenté quelques tentatives de projection des annotations d'origine du corpus multi-domaines MultiWOZ au paradigme à plat dans (Lee *et al.*, 2019). Comme nous remarquons dans l'exemple 6, le schéma de représentation a repris l'idée de compositionnalité des labels notamment pour les concepts de catégorie, où l'ensemble de l'acte de dialogue, le concept et sa valeur ont été projetés au niveau global `[CLS]`. Il est important de souligner cependant que cette projection demeure limitée si l'on veut prédire des représentations plus complexes.

En outre dans (Gupta *et al.*, 2018) le paradigme à plat a été remis en question pour les requêtes

Enoncés	[CLS]	I	need	parking
Annotation	Hotel-Inform+Parking*yes	O	O	O

TABLE 6 – Annotation à plat en BIO du corpus MultiWOZ

plus complexes. En effet le corpus TOP est modélisé d'une manière compositionnelle qui autorise les intentions imbriquées, il s'agit en effet d'une représentation hiérarchique similaire aux arbres syntaxiques. Vu la complexité de la représentation et pour faciliter l'utilisation du même schéma en plusieurs langues, (Aghajanyan *et al.*, 2020) ont proposé une extension de la représentation compositionnelle en représentation découplée qui a été utilisée dans (Li *et al.*, 2020). La figure 1 illustre la projection : le premier niveau de l'arbre correspond à l'intention, qui peut inclure un ou plusieurs concepts. Ces derniers peuvent à leur tour comporter des intentions ou une séquence de mot comme une valeur. En somme, les auteurs ont démontré que cette projection est un compromis entre le paradigme traditionnel à plat et la représentation en logique formelle. De même, les expériences faites dans (Cheng *et al.*, 2020) s'inscrivent dans le même cadre de la sémantique compositionnelle en affirmant que la compositionnalité peut simplifier la compréhension pour faciliter la tâche de suivi d'état de dialogue.

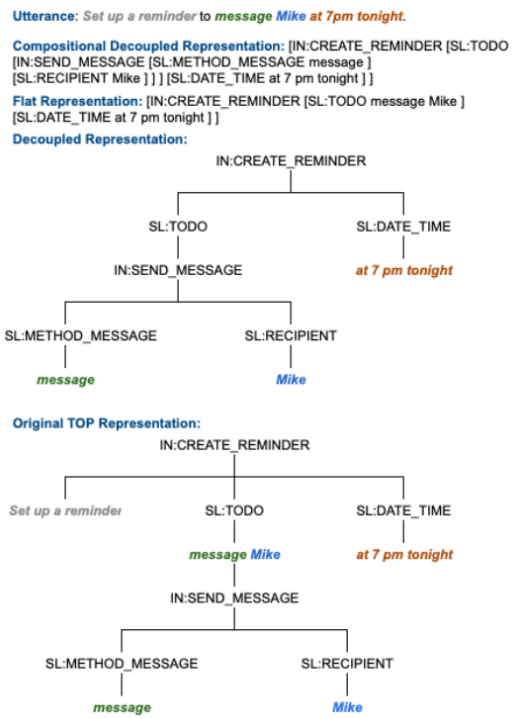


FIGURE 1 – Représentation compositionnelle découplée vs représentation plate dans mTOP (Li *et al.*, 2020)

De surcroît, les limitations du paradigme à plat et les motivations d'une représentation plus structurée ont été discutées dans (Hu *et al.*, 2022). L'article a proposé «DMR» une représentation en graphe qui comprend des nœuds d'Intention, de slots, des opérateurs indiquant les coréférences et la conjonction, et des mots-clés pour quelques éléments spéciaux en sémantique comme la négation. Une définition d'une nouvelle ontologie du domaine «Fast Food» du corpus MultiDoGo et une ré-annotation structurée qui lie les tours de dialogue permettant les annotations en contexte ont été l'objet de cet article. Des notions comme la "quantification" et "les adjectifs modificateurs" ont été ainsi soulignées.

Un exemple de leur représentation est dans la figure 2.

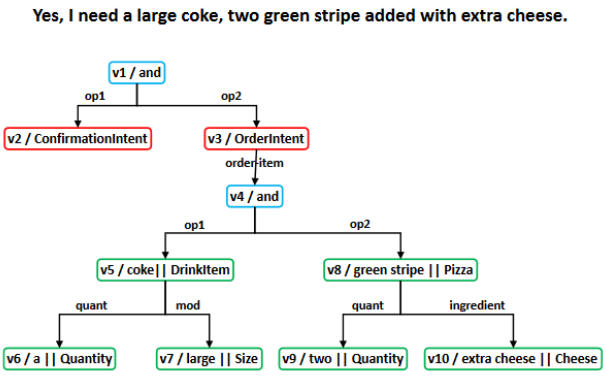


FIGURE 2 – Représentation en graphe dans DMR (Hu *et al.*, 2022)

Une proposition similaire à (Hu *et al.*, 2022) a été suggérée dans (Abrougui *et al.*, 2022). Cette proposition illustre des projections effectuées sur les annotations d'origine du corpus MultiWOZ2.3 sans qu'il soit nécessaire de modifier l'ontologie. Tel qu'indiqué dans la figure 3 Les actes de dialogue et les slots-valeurs sont transposés comme des nœuds, tandis que les slots sont représentés sous forme d'arcs reflétant la hiérarchie entre les labels. Les cas complexes, tels que les multi-intentions ou les intentions imbriquées peuvent être projetés dans ce format grâce à l'encodage Penman (Kasper, 1989), également utilisé dans les analyses AMR (Banarescu *et al.*, 2013).

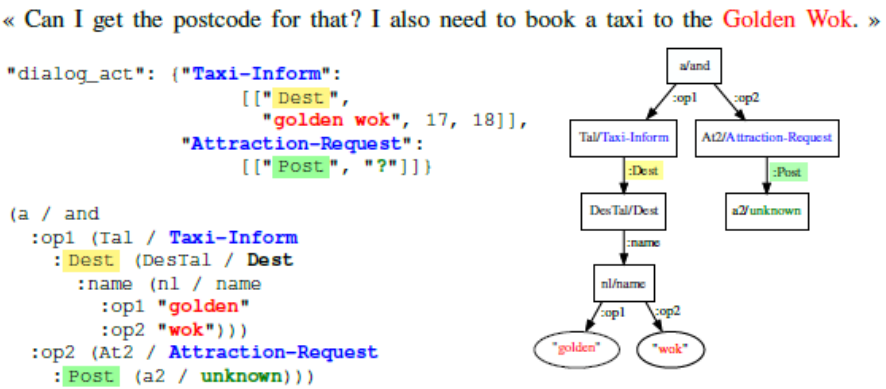


FIGURE 3 – Représentation en graphe et encodage Penman dans (Abrougui *et al.*, 2022)

Avec le développement des modèles de réseaux de neurones et l'essor des approches *seq2seq*, les représentations structurées pour les annotations sont devenues facilement exploitables par les systèmes NLU. Dans le même contexte, nous présentons nos perspectives et projets de recherche dans la partie suivante.

3.3 NLU avec une annotation structurée sur le corpus MultiWOZ

Les schémas d'annotation structurés et hiérarchiques offrent la possibilité d'étudier d'une façon plus approfondie une sémantique complexe et compositionnelle. Nous avons choisi donc de travailler sur le corpus MultiWOZ présentant des défis intéressants en raison de sa complexité (multi-domaines,

croisement entre les différentes intentions et concepts). Nous avons choisi de travailler en particulier sur la version 2.3, car elle offre des annotations utilisateur plus précises que les autres versions. Cependant, nous avons constaté que les annotations de cette version comportent des erreurs selon les règles de la logique NLU. Pour remédier à cela, notre objectif est de corriger et d'enrichir l'ontologie du corpus et de proposer un modèle de représentation qui permet une annotation contextuelle plus précise liant les tours de dialogue entre eux.

3.3.1 Corrections des annotations

Nous avons constaté dans la section 2 que les premières annotations au niveau utilisateurs de MultiWOZ ont été générée automatiquement. Bien que les auteurs dans MultiWOZ2.3 aient apporté des corrections, des précisions manquent encore. Nous avons donc commencé à examiner les conversations en identifiant une liste d'expressions, tels que «`I want to travel from`» ou «`I want to arrive by`», qui permettent de sélectionner un ensemble d'énoncés à vérifier et à corriger. Notre objectif dans cette étape est de garantir la cohérence entre les actes de dialogue, les concepts et les valeurs sans modifier l'ontologie de base.

En d'autres termes, l'exemple dans le tableau 7 doit être corrigé comme `"Hotel-Request": [["Internet", "free"]]` puisque l'utilisateur demande une information spécifique sur l'accès gratuit à Internet. Toutefois, la combinaison de la valeur `"free"` avec l'acte `"Request"` n'existe pas dans l'ontologie du corpus. Les actes dans MultiWOZ sont choisis en fonction du type des concepts, (slots-valeur en cas des concepts de catégorie), plutôt que du sens global de l'énoncé. Ainsi dans cette ontologie, la valeur de catégorie `"free"` est associée au concept `"Price"`, qui est lui-même associé à l'acte `"Inform"`. Cependant, le slot `"Price"` n'est utilisé que pour définir les prix des hôtels et des restaurants. Dans le cas du slot `"Internet"`, on a 4 valeurs principales : `"yes"` `"no"` `"dontcare"` associées à l'acte `"Inform"` et la valeur `"?"` associée à l'acte `"Request"`. Etant donné que l'énoncé dans l'exemple 7 est une simple demande d'information, nous nous limitons donc à corriger l'annotation en `"Hotel-Request": [["Internet", "?"]]`. Les travaux de correction sont actuellement en cours et leur évaluation fera l'objet d'une publication future.

Stratégie	Recherche de l'expression `"do they offer free wifi ?"`
Type de correction	semi-automatique
Annotation d'origine	`"Hotel-Inform": [["Internet", "yes"]]`
Correction	`"Hotel-Request": [["Internet", "?"]]`

TABLE 7 – Exemple de correction dans MultiWOZ2.3

3.3.2 Projection structurée

Dans (Abrougui *et al.*, 2022) nous trouvons une projection structurée des intentions, concepts et valeurs du MultiWOZ sans effectuer aucune modification (cf. figure 3). L'avantage de cette représentation est sa capacité de projeter les jeux de données standards comme ATIS, et les requêtes les plus complexes comme dans TOP et dans MultiWOZ. Dans cette étape nous avons deux étapes. Tout d'abord, nous visons à rajouter de nouveaux labels à l'ontologie du corpus comme l'acte `"Confirmation"`, ou comme les adverbes représentés par un concept de catégorie, comme il est

montré dans l'exemple 8. Nous réfléchissons aussi à la question si l'acte doit être associé ou dissocié du domaine.

Enoncé	Can you also give me some information about Finches Bed and Breakfast ? We 're thinking of staying there .
Annotation d'origine	`"Hotel-Inform": [["Name", "finches bed and breakfast"]]`
Proposition d'annotation	`"Hotel": [["Inform-Name", "finches bed and breakfast"],["Request-Info", "?"], ["Modifier", "maybe"]]`

TABLE 8 – Proposition d'une nouvelle annotation du corpus MultiWOZ2.3

Notre objectif dans un second temps est de reprendre cette structure en rajoutant des annotations de coréférence en liant les antécédents et les anaphores avec les variables utilisées dans la notation Penman. La figure 4 ci-dessous qui présente un exemple fictif illustre ce projet. En examinant l'énoncé, nous remarquons la présence de deux actes de dialogues distincts qui partagent le slot `"Area"` clairement indiqué par l'utilisation de l'adverbe `"there"`. La représentation sémantique de cette structure avec un format à plat serait particulièrement difficile. En outre, il convient de souligner la problématique de l'implicite soulevée par l'expression `"I have a car"`, qui fait référence au concept de catégorie `"Parking"` et à sa valeur normalisée `"yes"`. Afin de faciliter la représentation de ces entités et de leurs liens, l'utilisation du format penman basé sur les variables (comme `"a1"` dans la figure) s'avère être un choix judicieux.

Notre objectif est en effet de faire une tâche NLU mais en couvrant toutes les informations possibles, comme l'annotation de la coréférence.

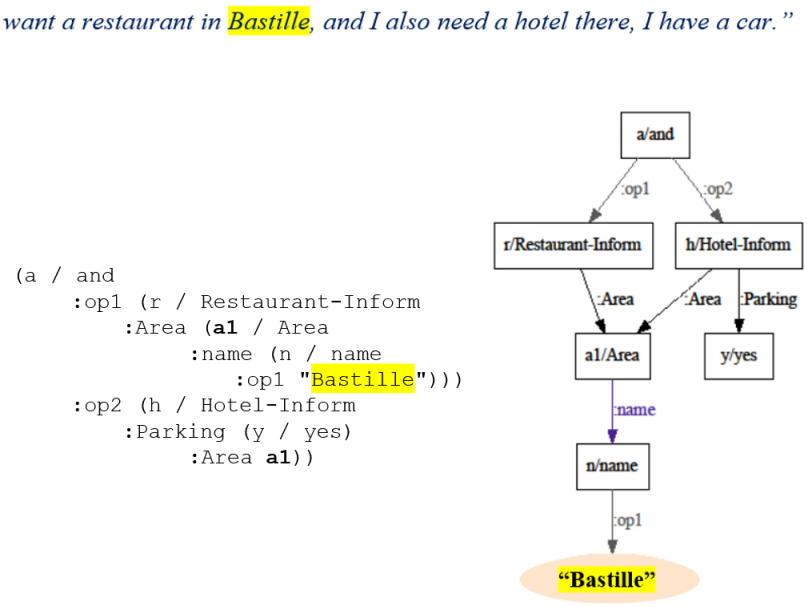


FIGURE 4 – schéma structuré pour la coréférence

3.3.3 Sémantique des labels

Il est vrai que les schémas de représentation structurés sont considérés comme des outils efficaces pour représenter les données avec des informations sémantiques complexes, mais la sémantique

sous-jacente des étiquettes est tout aussi importante. Les auteurs dans (Athiwaratkun *et al.*, 2020) ont souligné cette question en expliquant que les modèles de langage génératifs offrent un moyen naturel d'incorporer le sens des étiquettes dans les tâches de compréhension. Ils ont représenté la sortie en une séquence augmentée qui contient la séquence d'entrée avec leurs labels. Ils ont nettoyé les labels de tous les symboles et les ont représentés sous leur forme de langage naturel (par exemple, "AddToPlaylist" est devenu "Add To Playlist") afin de mieux exploiter les capacités des modèles de langage génératifs à comprendre le langage naturel.

Par ailleurs, l'unification des ensembles de données SLU reste un défi dans ce domaine, car les ontologies et les noms des concepts varient considérablement. Néanmoins, si des noms de concepts sont communs et que les modèles génératifs sont capables de comprendre le langage naturel, il est possible d'unifier ces jeux de données plus facilement sans effectuer de nombreuses modifications au niveau de leurs ontologies. Nous avons testé cette hypothèse en examinant les données SNIPS et MultiWOZ.

En effet, SNIPS a un domaine en commun avec MultiWOZ, qui est "Restaurant". Il existe également le concept "Name" associé à cinq catégories différentes (tels que `"restaurant_name"`, `"movie_name"` et `"location_name"`). Nous avons fusionné les deux en deux étapes : **(1)** la première consiste à fusionner les deux données sans changer les ontologies. **(2)** La deuxième consiste à changer les concepts de SNIPS en une seule entité (`"movie_name"` devient `"name"`) et les intentions en une forme `"Domaine-Acte"` (`"SearchScreeningEvent"` devient `"Event-Search"`). En ce qui concerne MultiWOZ, nous avons étendu les concepts en leur forme de langage naturel (`"Dest"` devient `"Destination"`), nous les avons mis tous en minuscules et nous avons appris un modèle mT5 (Xue *et al.*, 2020) qui prend en entrée les énoncés et génère les représentations structurées comme indiqué dans la figure 3, codées en format Penman.

Les performances globales dans le tableau 9 montrent que la fusion des deux corpora n'a pas engendré de résultats significatifs. Bien qu'une légère baisse ait été observée pour SNIPS au niveau de l'accuracy globale, MultiWOZ n'a pas vraiment changé. Nous avons ensuite examiné les performances au niveau des concepts composés comportant le mot `"name"` dans leur nom et simplifiés au format MultiWOZ.

Le tableau 10 présente les performances des F1-mesures au niveau Intention(slot,valeur) correspondant aux slots modifiés et leurs domaines respectifs.

Nous constatons une amélioration au niveau des résultats pour le slot `"movie_name"` et en particulier pour `"restaurant_name"` suite à la modification de l'ontologie de SNIPS (MS-S^O), avec une augmentation de 16%. Ce label partage en effet le domaine et le nom du concept avec MultiWOZ, et il semble que le modèle génératif a bien capturé la sémantique derrière.

En revanche, les performances pour `"location_name"` ont diminué. Les résultats pour `"object_name"` ont également diminué avec l'intention `"SearchScreeningEvent"` mais ils s'améliorent avec 1% avec l'intention `"RateBook"`. Tout cela montre que l'association entre les différents labels affectent leurs significations, et il est possible pour les modèles génératifs de les prédire si on peut les représenter d'une manière cohérente. L'augmentation significative de certains concepts met en évidence l'importance de l'unification des ontologies. En effet, lorsque les énoncés et leurs labels partagent une sémantique commune, et sont bien définis et unifiés sous le même label, cela peut contribuer à renforcer les performances des systèmes N/SLU.

Nous envisageons d'approfondir cette approche et étudier précisément les ontologies en exploitant à la fois les représentations structurées des frames sémantiques et le potentiel des modèles génératifs

pour l'unification des jeux de données en compréhension du langage naturel et parlé.

SNIPS			
	S-S	**MS-S**	**MS-S^O**
F1 intention	98,1	97,8	98,6
F1 (concept,valeur)	95,0	94,8	94.9
F1 Intent(concept,valeur)	94,7	94,6	94,6
Accuracy global	88,7	87,8	88,0
MultiWOZ 2.3			
	M-M	**MS-M**	**MS-M^O**
F1 intention	96,2	96,2	96,3
F1 (concept,valeur)	94,7	94,9	94,9
F1 Intent(concept,valeur)	94,1	94,2	94,3
Accuracy global	87,6	87,7	87,5

TABLE 9 – Résultats des expériences sur la sémantique des labels : apprentissage et test sur SNIPS (**S-S**), apprentissage et test sur MultiWOZ (**M-M**), apprentissage sur MultiWOZ et SNIPS sans modifications de l'ontologie et test sur SNIPS **(MS-S**), apprentissage sur MultiWOZ et SNIPS sans modifications de l'ontologie et test sur MultiWOZ (**MS-M**), apprentissage sur MultiWOZ et SNIPS avec modifications de l'ontologie et test sur SNIPS (**MS-S**^{O}), apprentissage sur MultiWOZ et SNIPS avec modifications de l'ontologie et test sur MultiWOZ (**MS-M**^{O})

	S-S	**MS-S**	**MS-S^O**
`SearchScreeningEvent+movie_name (event-search+name)`	86,7	77,1	89,6
`SearchScreeningEvent+location_name (event-search+name)`	97,9	97,9	91,7
`SearchCreativeWork+object_name (work-search+name)`	85,9	85,6	82,9
`AddToPlaylist+entity_name (music-add+name)`	74,6	80,0	71,9
`RateBook+object_name (book-rate+name)`	95,0	97,5	96,3
`BookRestaurant+restaurant_name (restaurant-book+name)`	73,3	83,9	89,7

TABLE 10 – F1 mesure niveau Inent(concept,valeur) pour les concepts composés par le slot "name" : format d'origine (format modifié)

4 Conclusion

La compréhension du langage naturel et parlé est une tâche fondamentale dans les systèmes de dialogue. Les deux tâches connues visent à comprendre les commandes de l'utilisateur et ses interactions avec un agent robotique. Dans cet article, nous avons présenté les différents jeux de données utilisés dans ce domaine et nous avons comparé leurs ontologies et leurs schémas d'annotation. Les représentations sémantiques structurées et la méthode de graphe ont été mises en valeur pour leur potentiel à refléter la hiérarchie sémantique entre les frames sémantiques et à exploiter le contexte. La sémantique des labels a également fait l'objet d'une discussion dans nos projets de recherche basés fondamentalement sur des expériences sur le corpus conversationnel MultiWOZ. Dans le but d'unifier les données et d'exploiter mieux le contexte pour construire des systèmes robustes, nous envisageons d'approfondir nos études des ontologies et des représentations structurées tout en exploitant l'histoire conversationnelle.

Références

ABROUGUI R., DAMNATI G., HEINECKE J. & BÉCHET F. (2022). Étiquetage ou génération de séquences pour la compréhension automatique du langage en contexte d'interaction ? In *Traitement Automatique des Langues Naturelles (TALN 2022)*, p. 64–73 : ATALA.

AGHAJANYAN A., MAILLARD J., SHRIVASTAVA A., DIEDRICK K., HAEGER M., LI H., MEHDAD Y., STOYANOV V., KUMAR A., LEWIS M. *et al.* (2020). Conversational semantic parsing. *arXiv preprint arXiv :2009.13655*.

ASRI L. E., SCHULZ H., SHARMA S., ZUMER J., HARRIS J., FINE E., MEHROTRA R. & SULEMAN K. (2017). Frames : a corpus for adding memory to goal-oriented dialogue systems. *arXiv preprint arXiv :1704.00057*.

ATHIWARATKUN B., SANTOS C. N. D., KRONE J. & XIANG B. (2020). Augmented natural language for generative sequence labeling. *arXiv preprint arXiv :2009.13272*.

BAKER C. F., FILLMORE C. J. & LOWE J. B. (1998). The berkeley framenet project. In *COLING 1998 Volume 1 : The 17th International Conference on Computational Linguistics*.

BANARESCU L., BONIAL C., CAI S., GEORGESCU M., GRIFFITT K., HERMJAKOB U., KNIGHT K., KOEHN P., PALMER M. & SCHNEIDER N. (2013). Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, p. 178–186.

BASTIANELLI E., VANZO A., SWIETOJANSKI P. & RIESER V. (2020). Slurp : A spoken language understanding resource package. *arXiv preprint arXiv :2011.13205*.

BÉCHET F. & RAYMOND C. (2018). Is atis too shallow to go deeper for benchmarking spoken language understanding models ? In *InterSpeech 2018*, p. 1–5.

BÉCHET F. & RAYMOND C. (2019). Benchmarking benchmarks : introducing new automatic indicators for benchmarking spoken language understanding corpora. In *Interspeech*.

BECHET F., RAYMOND C., HAMANE A., ABROUGUI R., MARZINOTTO G. & DAMNATI G. (2022). Can we predict how challenging spoken language understanding corpora are across sources, languages, and domains ? In *Conversational AI for Natural Human-Centric Interaction : 12th International Workshop on Spoken Dialogue System Technology, IWSDS 2021, Singapore*, p. 33–45 : Springer.

BELLOMARIA V., CASTELLUCCI G., FAVALLI A. & ROMAGNOLI R. (2019). Almawave-slu : A new dataset for slu in italian. *arXiv preprint arXiv :1907.07526*.

BONIAL C., DONATELLI L., ABRAMS M., LUKIN S., TRATZ S., MARGE M., ARTSTEIN R., TRAUM D. & VOSS C. (2020). Dialogue-amr : abstract meaning representation for dialogue. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 684–695.

BUDZIANOWSKI P., WEN T.-H., TSENG B.-H., CASANUEVA I., ULTES S., RAMADAN O. & GAŠIĆ M. (2018). Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv :1810.00278*.

CHEN X., GHOSHAL A., MEHDAD Y., ZETTLEMOYER L. & GUPTA S. (2020). Low-resource domain adaptation for compositional task-oriented semantic parsing. *arXiv preprint arXiv :2010.03546*.

CHENG J., AGRAWAL D., ALONSO H. M., BHARGAVA S., DRIESEN J., FLEGO F., GHOSH S., KAPLAN D., KARTSAKLIS D., LI L. *et al.* (2020). Conversational semantic parsing for dialog state tracking. *arXiv preprint arXiv :2010.12770*.

COOPE S., FARGHLY T., GERZ D., VULIĆ I. & HENDERSON M. (2020). Span-convert : Few-shot span extraction for dialog with pretrained conversational representations. *arXiv preprint arXiv :2005.08866*.

COUCKE A., SAADE A., BALL A., BLUCHE T., CAULIER A., LEROY D., DOUMOURO C., GISSELBRECHT T., CALTAGIRONE F., LAVRIL T. *et al.* (2018). Snips voice platform : an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv :1805.10190*.

DAHL D. A., BATES M., BROWN M. K., FISHER W. M., HUNICKE-SMITH K., PALLETT D. S., PAO C., RUDNICKY A. & SHRIBERG E. (1994). Expanding the scope of the atis task : The atis-3 corpus. In *Human Language Technology : Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

DESOT T., RAIMONDO S., MISHAKOVA A., PORTET F. & VACHER M. (2018). Towards a french smart-home voice command corpus : Design and nlu experiments. In *Text, Speech, and Dialogue : 21st International Conference, TSD 2018, Brno, Czech Republic, September 11-14, 2018, Proceedings 21*, p. 509–517 : Springer.

DEVILLERS L., MAYNARD H., ROSSET S., PAROUBEK P., MCTAIT K., MOSTEFA D., CHOUKRI K., CHARNAY L., BOUSQUET C., VIGOUROUX N. *et al.* (2004). The french media/evalda project : the evaluation of the understanding capability of spoken language dialogue systems. In *LREC*.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2018). Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*.

DINARELLI M. (2010). *Spoken language understanding : from spoken utterances to semantic structures*. Thèse de doctorat, University of Trento.

DINARELLI M., QUARTERONI S., TONELLI S., MOSCHITTI A. & RICCARDI G. (2009). Annotating spoken dialogs : from speech segments to dialog acts and frame semantics. In *Proceedings of SRSL 2009, the 2nd Workshop on Semantic Representation of Spoken Language*, p. 34–41.

ERIC M., GOEL R., PAUL S., SETHI A., AGARWAL S., GAO S. & HAKKANI-TUR D. (2019). Multiwoz 2.1 : Multi-domain dialogue state corrections and state tracking baselines. *arXiv :1907.01669*.

FITZGERALD J., HENCH C., PERIS C., MACKIE S., ROTTMANN K., SANCHEZ A., NASH A., URBACH L., KAKARALA V., SINGH R. *et al.* (2022). Massive : A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages. *arXiv preprint arXiv :2204.08582*.

GUPTA S., SHAH R., MOHIT M., KUMAR A. & LEWIS M. (2018). Semantic parsing for task oriented dialog using hierarchical representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 2787–2792, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/D18-1300.

HAN T., LIU X., TAKANOBU R., LIAN Y., HUANG C., WAN D., PENG W. & HUANG M. (2020). Multiwoz 2.3 : A multi-domain task-oriented dialogue dataset enhanced with annotation corrections and co-reference annotation. *arXiv :2010.05594*.

HEMPHILL C. T., GODFREY J. J. & DODDINGTON G. R. (1990). The atis spoken language systems pilot corpus. In *Speech and Natural Language : Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.

HU X., DAI J., YAN H., ZHANG Y., GUO Q., QIU X. & ZHANG Z. (2022). Dialogue meaning representation for task-oriented dialogue systems. *arXiv preprint arXiv :2204.10989*.

KASPER R. T. (1989). A flexible interface for linking applications to penman's sentence generator. In *Speech and Natural Language : Proceedings of a Workshop Held at Philadelphia, Pennsylvania, February 21-23, 1989*.

LEE S., ZHU Q., TAKANOBU R., LI X., ZHANG Y., ZHANG Z., LI J., PENG B., LI X., HUANG M. *et al.* (2019). Convlab : Multi-domain end-to-end dialog system platform. *arXiv preprint arXiv :1904.08637*.

LEFEVRE F., MOSTEFA D., BESACIER L., QUIGNARD M., CAMELIN N., FAVRE B., JABAIAN B., BARAHONA L. M. R. *et al.* (2012). Leveraging study of robustness and portability of spoken language understanding systems across languages and domains : the portmedia corpora. In *The International Conference on Language Resources and Evaluation*.

LI H., ARORA A., CHEN S., GUPTA A., GUPTA S. & MEHDAD Y. (2020). Mtop : A comprehensive multilingual task-oriented semantic parsing benchmark. *arXiv preprint arXiv :2008.09335*.

LOOS B. (2006). Scaling natural language understanding via user-driven ontology learning. In *Proceedings of the Third Workshop on Scalable Natural Language Understanding*, p. 33–40.

MA Y., AUDIBERT L. & NAZARENKO A. (2009). Ontologies étendues pour l'annotation sémantique. In *20es Journées Francophones d'Ingénierie des Connaissances*, p. 205–216.

PESKOV D., CLARKE N., KRONE J., FODOR B., ZHANG Y., YOUSSEF A. & DIAB M. (2019). Multi-domain goal-oriented dialogues (multidogo) : Strategies toward curating and annotating large scale dialogue data. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 4526–4536.

PORTET F., CAFFIAU S., RINGEVAL F., VACHER M., BONNEFOND N., ROSSATO S., LECOUTEUX B. & DESOT T. (2019). Context-aware voice-based interaction in smart home-vocadom@ a4h corpus collection and empirical assessment of its usefulness. In *2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, p. 811–818 : IEEE.

QIN L., XIE T., CHE W. & LIU T. (2021). A survey on spoken language understanding : Recent advances and new frontiers. *arXiv preprint arXiv :2103.03095*.

SCHUSTER S., GUPTA S., SHAH R. & LEWIS M. (2019). Cross-lingual transfer learning for multilingual task oriented dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 3795–3805, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1380.

SHAH P., HAKKANI-TÜR D., TÜR G., RASTOGI A., BAPNA A., NAYAK N. & HECK L. (2018). Building a conversational agent overnight with dialogue self-play. *arXiv preprint arXiv :1801.04871*.

TOMASELLO P., SHRIVASTAVA A., LAZAR D., HSU P.-C., LE D., SAGAR A., ELKAHKY A., COPET J., HSU W.-N., ADI Y. *et al.* (2023). Stop : A dataset for spoken task oriented semantic parsing. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, p. 991–998 : IEEE.

TUR G. & DE MORI R. (2011). *Spoken language understanding : Systems for extracting semantic information from speech*. John Wiley & Sons.

WELD H., HUANG X., LONG S., POON J. & HAN S. C. (2022). A survey of joint intent detection and slot filling models in natural language understanding. *ACM Computing Surveys*, **55**(8), 1–38.

XUE L., CONSTANT N., ROBERTS A., KALE M., AL-RFOU R., SIDDHANT A., BARUA A. & RAFFEL C. (2020). mt5 : A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv :2010.11934*.

YE F., MANOTUMRUKSA J. & YILMAZ E. (2021). Multiwoz 2.4 : A multi-domain task-oriented dialogue dataset with essential annotation corrections to improve state tracking evaluation.

ZANG X., RASTOGI A., SUNKARA S., GUPTA R., ZHANG J. & CHEN J. (2020). Multiwoz 2.2 : A dialogue dataset with additional annotation corrections and state tracking baselines. In *WNLPC AI, ACL'20*.

ZETTLEMOYER L. S. & COLLINS M. (2012). Learning to map sentences to logical form : Structured classification with probabilistic categorial grammars. *arXiv preprint arXiv :1207.1420*.

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Étude de la fidélité des entités dans les résumés par abstraction

Eunice Akani[1, 2]
(1) Aix-Marseille Univ, CNRS, LIS, Marseille, France
(2) Enedis, Marseille, France
eunice.akani@lis-lab.fr

RÉSUMÉ

L'un des problèmes majeurs dans le résumé automatique de texte par abstraction est la fidélité du résumé généré vis-à-vis du document. Les systèmes peuvent produire des informations incohérentes vis-à-vis du document. Ici, nous mettons l'accent sur ce phénomène en restant focalisé sur les entités nommées. L'objectif est de réduire les hallucinations sur celles-ci. Ainsi, nous avons généré des résumés par sampling et avons sélectionné, à l'aide d'un critère basé sur le risque d'hallucination sur les entités et les performances du modèle, ceux qui minimisent les hallucinations sur les entités. Une étude empirique du critère montre son adaptabilité pour la sélection de résumé. Nous avons proposé des heuristiques pour la détection des entités qui sont des variations ou flexions d'autres entités. Les résultats obtenus montrent que le critère réduit les hallucinations sur les entités nommées en gardant un score ROUGE comparable pour CNN/DM.

ABSTRACT

Named Entities Faithfulness in Abstractive Text Summarization.

One of the major problems in abstraction text summarization is the faithfulness of the generated summary regard to the source document. Systems may produce inconsistent information reading the document. Here, we emphasize this phenomenon by remaining focused on named entities. The goal is to reduce hallucinations on these entities. Thus, we generated summaries by sampling and selected, using a criterion based on the risk of hallucination on entities and the performance of the model, those which minimize the hallucinations on the entities. An empirical study of the criterion shows its adaptability for summary selection. We also proposed heuristics for the detection of entities which are variations or inflections of other entities. The results obtained show that the criterion reduces hallucinations on named entities by keeping a comparable ROUGE score for CNN/DM.

MOTS-CLÉS : Résumé automatique de texte, hallucination, entité nommée.

KEYWORDS: Automatic text summarization, hallucination, named entity.

1 Introduction

Le résumé automatique de texte consiste à faire une synthèse d'un document en gardant les informations pertinentes. Il existe deux types de résumé automatique : le résumé par extraction et le résumé par abstraction. Le résumé par extraction qui consiste à extraire dans un premier temps les informations importantes du document puis dans un second temps à en faire un résumé. Le résumé par abstraction consiste quant à lui à faire la synthèse d'un document en utilisant des paraphrases et des nouveaux mots. La tâche de résumé automatique a beaucoup évolué depuis l'apparition des modèles à

Actes des 16e Rencontres Jeunes Chercheurs en RI (RJCRI) et 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL), pages 21–36.

base de transformers et de modèles de langue pré-entraînés (Vaswani *et al.*, 2017; Devlin *et al.*, 2019; Lewis *et al.*, 2020). Pendant que la tâche de résumé par extraction devient de plus en plus atteignable, celle du résumé par abstraction reste un grand défi. En effet, Kryscinski *et al.*, 2019 a montré que les systèmes produisent des informations qui ne sont pas toujours fidèles au document source et cela n'est pas vérifiable par l'utilisation de mesures d'évaluation du résumé telles que ROUGE (Lin, 2004), METEOR (Banerjee & Lavie, 2005) ou encore le BERTScore (Zhang* *et al.*, 2020). 30% des résumés générés par des systèmes abstractifs contiennent des incohérences vis-à-vis du document source selon Cao *et al.*, 2018. Maynez *et al.*, 2020 a utilisé le terme « hallucination » pour qualifier ces incohérences. Elles peuvent être dues à une mauvaise association d'informations provenant du document ou à l'utilisation d'informations hors du document. Plusieurs études ont été menées pour l'évaluation factuelle des résumés générés (Durmus *et al.*, 2020; Wang *et al.*, 2020) ainsi que pour la réduction d'hallucinations (Pagnoni *et al.*, 2021; Chen *et al.*, 2021; Fan *et al.*, 2018a). Pour notre part, nous étudions les hallucinations au niveau des entités nommées, ce qui s'est avéré assez fréquent dans les résumés générés (Chen *et al.*, 2021). Plutôt que d'intervenir en amont, pendant l'entraînement du modèle, nous avons généré plusieurs résumés dans un espace afin de sélectionner le meilleur résumé suivant un critère qui pourra réduire les hallucinations et donc corriger les problèmes de fidélité du résumé généré par rapport au document source. L'idée étant de réduire le nombre d'entités hallucinées, nous avons introduit un critère de sélection de résumé basé sur le « risque » d'avoir des entités qui n'apparaissent pas dans le document source (entités hallucinées). Ce critère nous permet ainsi de choisir un résumé avec le moins d'entités hallucinées dans l'ensemble des résumés générés. Nous avons sélectionné, conjointement, les résumés ayant le meilleur score ROUGE avec la référence et avions procédé à une évaluation humaine de leurs entités pour vérifier que les entités du résumé sont correctement utilisées (leur emploi ne contredit pas le document source), et que les entités dites hors du document le sont vraiment. Cette évaluation, permet ainsi d'étudier de manière empirique le critère de sélection. Ce critère étant basé sur les entités nommées, nous avons donc mis sur place des heuristiques afin de détecter les variations d'entités nommées (ex. les nombres écrits en lettres ou en chiffres, les erreurs d'orthographe, les flexions). Nos contributions se résument donc à :

— La génération de divers résumés avec la méthode de sampling afin d'avoir un résumé qui minimise les erreurs factuelles.
— L'introduction d'un critère de sélection de résumé avec le moins d'entités hors du document.
— La mise en place d'heuristique pour la détection de variation d'entités.

Document :
She said she will seek a judicial review against Mark H Durkan because he took the decision to adopt the planning policy without the agreement of the full Northern Ireland Executive. Helen Jones reports for BBC Newsline.

Résumé :
Helen Jones seek a judicial review against Mark H Durkan, the Northern Ireland's environment minister reports BBC.

FIGURE 1 – Exemple d'hallucination intrinsèque (en bleu) et extrinsèque (en rouge).

2 Contexte et motivation

L'évaluation des résumés automatiques en terme de fidélité est de plus en plus importante. En effet, un résumé n'étant pas fidèle au document source n'est pas exploitable car rempli d'incohérences. Cela suscite l'attention des chercheurs en trois questions : pourquoi se focaliser sur la fidélité, comment évaluer la fidélité d'un résumé, et comment réduire les hallucinations ?

Certains travaux présentent l'insuffisance des mesures d'évaluation actuelles par catégorisation des erreurs fréquentes dans le résumé suivant des typologies d'erreurs qu'ils ont défini (Ji *et al.*, 2023). Maynez *et al.*, 2020 ont nommés les informations incohérentes vis-à-vis du document des hallucinations. Ils ont défini deux types d'hallucinations, les hallucinations intrinsèques qui sont les informations provenant du document qui ne sont pas cohérentes vis-à-vis de celui-ci et les hallucinations extrinsèques qui sont définies comme des informations du résumé qui sont hors du document source. Un exemple est donné sur la figure 1. Pagnoni *et al.*, 2021 propose une typologie plus détaillée prenant en compte la vérification de contenu, les erreurs liées au discours et la sémantique de surface. Ils ont mené une évaluation humaine à grande échelle sur des résumés de divers systèmes en utilisant le datatset CNN/Daily Mail (Hermann *et al.*, 2015; Nallapati *et al.*, 2016) et XSum (Narayan *et al.*, 2018) afin d'identifier les erreurs fréquentes dans les résumés de référence. L'analyse des hallucinations extrinsèques par Chen *et al.*, 2021 montre que, pour la majorité, cela se produit sur des entités et quantités nommées. Akani *et al.*, 2022 propose une typologie d'erreurs pour les résumés candidats et une typologie d'abstraction pour les résumés de référence. Cela leur a permis de montrer que les erreurs les plus courantes étaient les informations provenant hors du document source et les informations non inférables à partir du document source.

Pour tenter de résoudre le problème des hallucinations, certains proposent l'utilisation d'entailment textuel (Maynez *et al.*, 2020) ou encore de modèle de question-réponse (Durmus *et al.*, 2020; Wang *et al.*, 2020) pour l'évaluation de la fidélité du résumé par rapport au document source. Chen *et al.*, 2021 propose de corriger les erreurs sur les entités nommées en modifiant les entités des résumés générés par des entités du document source. Puis à l'aide des modèles qu'ils ont entraînés, ils choisissent les résumés factuellement cohérents. Cela leur permet d'éviter les hallucinations extrinsèques, mais la méthode crée des hallucinations intrinsèques dans les résumés. (Nan *et al.*, 2021) propose une méthode basée sur le filtrage des données d'entraînement et l'apprentissage multi-tâches. Fan *et al.*, 2018a propose de contrôler la génération du résumé avec une liste d'entités nommées désirées en entrée, et Narayan *et al.*, 2021 propose, pendant l'entraînement du modèle, de conditionner la génération des résumés par les entités nommées en générant d'abord la liste des entités à utiliser dans le résumé, puis le résumé lui-même.

Notre étude est proche des différentes études sur les entités nommées tant l'objectif est de réduire les hallucinations sur ceux-ci. L'étude la plus proche de la nôtre est celle de Chen *et al.*, 2021 car il s'agit de choisir un résumé parmi un ensemble de résumés créé ou généré. La différence réside dans la génération des résumés et dans la sélection du meilleur résumé. En effet, plutôt que de modifier les résumés générés après leur génération par le modèle, nous générons un ensemble de résumé par la méthode de sampling afin de couvrir un grand espace de génération multiple. La section suivante présente la mise en place de cette génération.

Dans la suite de ce document, nous appellerons hallucination, de manière générale, uniquement les informations en dehors du document source. Le terme « entités hallucinées » correspondra donc aux entités qui ne sont pas dans le document source. Aussi, il est à noter que nous utiliserons le terme « hallucination factuelle » pour parler des informations hors du document qui sont factuellement

correctes.

3 Génération de résumés par échantillonnage (sampling)

La plupart des méthodes pour la génération automatique de texte se font par l'utilisation du « beam search » pour parcourir l'espace de recherche de manière efficace. Le « beam search » conserve les num_beams ayant la probabilité élevée à chaque étape. Il minimise la possibilité de manquer des séquences cachées à forte probabilité. Cela ne permet pas d'avoir un ensemble de résumés variés dans l'espace de génération. De cette façon, la méthode de sampling est la plus adaptée pour générer plusieurs résumés. Nous avons donc généré des résumés en l'utilisant pour la sélection du prochain token à générer. Ainsi, les résumés obtenus proviennent de l'utilisation de « greedy search », « beam search », la température, du top-P et du top-K. Tandis que le « greedy search » consiste à sélectionner le token ayant la probabilité la plus élevée, l'échantillonnage avec la température consiste à remettre à l'échelle les logits avant d'appliquer la fonction softmax. Le Top-K (Fan *et al.*, 2018b) consiste à prendre les K mots suivants les plus probables et à redistribuer la probabilité entre ces K mots. Le Top-P quant à lui ou échantillonnage Nucleus (Holtzman *et al.*, 2019) consiste, sachant une probabilité p, à prendre le plus petit ensemble possible de mots suivants dont la probabilité cumulée est supérieure à la probabilité p. Il y a également une redistribution de la probabilité entre les mots de l'ensemble. Pour notre part, nous avons fait varier les différents paramètres comme suit :

— Pour la température : de 0.5 à 1 avec un pas de 0.1.
$T = [0.5, 0.6, 0.7, 0.8, 0.9]$;
— Pour le Top-p : de 0.75 à 0.96 avec un pas de 0.05
$Top - p = [0.75, 0.80, 0.85, 0.90, 0.95]$;
— Pour le Top-k de 40 à 70 avec un pas de 10
$Top - k = [40, 50, 60]$

Nous avons généré ainsi 77 résumés [1] pour chaque exemple en faisant varier les différentes valeurs enumérées plus haut à chaque étape. Nous avons pris 1024 comme le nombre de tokens maximum pour l'entrée du système et 128 tokens maximum pour la sortie du système (les résumés).

Dans la suite du papier, nous introduisons un critère de sélection de résumé basé sur les entités nommées afin de réduire les hallucinations sur celles-ci.

4 Sélection de résumé par critère

Pour réduire le problème de fidélité des entités du résumé généré par rapport au document source, il est important d'étudier le risque d'hallucinations sur les entités nommées. Nous avons donc introduit « NEHR Named Entities Hallucination Risk » une mesure sur les entités nommées comme (Nan *et al.*, 2021). Cette mesure peut être utilisée pour construire un modèle de sélection de résumés minimisant les hallucinations sur les entités nommées. Certains ont proposé d'utiliser l'entailment textuel (Falke *et al.*, 2019; Maynez *et al.*, 2020) ou encore un système de question-réponse (Durmus *et al.*, 2020) afin d'évaluer la fidélité du résumé par rapport au document source. Ce sont des méthodes qui sont plutôt gourmandes en termes de ressources. Pour notre part, nous avons décidé de mettre en place une heuristique qui considère que si une entité est en dehors du document source alors elle est hallucinée.

1. 75 avec les méthodes de d'échantillonnage + la génération greedy + la génération avec le beam de taille 4

Elle rend donc le résumé non fidèle au document source. Ceci nous permet de définir le critère comme suit.

4.1 Définition

Pour un document d et un résumé s :

$$\textit{NEHR}(d, s) = (1 - \frac{|\textit{entities} \in d \wedge s|}{|entities \in s|}) \times 100 \quad (1)$$

Pour la détection des entités nommées, nous avons utilisé un système automatique de reconnaissance d'entités nommées. Ce critère n'inclut pas la référence de telle sorte qu'il puisse être utilisé en prédiction. Cependant, n'ayant pas de moyen de dire si une entité considérée comme risquée est correcte ou pas, nous avons mené une étude empirique pour savoir si NEHR est corrélé aux hallucinations dans les résumés générés. Dans la sous-section suivante, nous présenterons cette étude.

4.2 Dataset et modèle de résumé automatique

Pour notre étude, nous avons utilisé le corpus CNN/Daily Mail (Hermann *et al.*, 2015; Nallapati *et al.*, 2016) et le corpus XSum (Narayan *et al.*, 2018). CNN/Daily Mail est un corpus populaire pour la tâche de résumé automatique. Il est composé d'articles provenant des sites de CNN et Daily Mail. XSum est un corpus composé de 226 711 articles provenant de BBC de 2010 à 2017. Les différents articles traitent de plusieurs sujets notamment, l'actualité, l'éducation, le business, la météo, la technologie, la santé, la politique, le sport, etc. La particularité de ce corpus est que les résumés ont été écrits par des professionnels. Les résumés de XSum sont faits en une seule phrase. Ce corpus a été introduit comme corpus pour la tâche de résumé automatique par abstraction, car il contient 36% de nouveaux unigrams. Comme modèle, nous avons BART (Lewis *et al.*, 2020), c'est une architecture basée sur les Transformers (Vaswani *et al.*, 2017) qui est utilisée pour la tâche de résumé automatique de texte. Il existe différentes tailles du modèle. En ce qui nous concerne, nous avons utilisé BART-large qui se compose de 12 couches de Transformers aussi bien dans l'encodeur que dans le décodeur. Nous avons initialisé le modèle avec les poids du modèle se trouvant sur Hugging Face (Wolf *et al.*, 2020) aussi bien pour CNN/DM [2] que pour XSum [3].

4.3 Étude empirique du risque

Vérification de la pertinence du critère. L'étude empirique utilisée pour vérifier la pertinence du critère est la suivante :

1. Sélectionner un corpus C de résumé automatique contenant des documents sources et les résumés associés. Entraîner plusieurs systèmes de résumé automatique sur l'ensemble d'entraînement. Puis à l'aide du modèle entraîné, générer un ensemble de résumés S_d alternatifs pour le document d du jeu de données de test. Ensuite, calculer le ROUGE Score ainsi que le NEHR de chaque résumé $s \in S_d$.

2. https://huggingface.co/facebook/bart-large-cnn
3. https://huggingface.co/facebook/bart-large-xsum

2. Pour chaque document $d \in C$, sélectionner $\hat{s_d}$:

$$\hat{s_d} = \underset{s \in S_d}{\operatorname{argmax}} ROUGE(s, s_{ref})$$

3. Exécuter le système de reconnaissance automatique d'entités sur chaque résumé $\hat{s_d}$ pour y extraire les entités nommées.
4. Annoter manuellement chaque entité e détectée dans le résumé $\hat{s_d}$ suivant deux dimensions : d'abord, dans le document (in) ou en dehors du document (out) puis bien ou mal utilisé (utilisation correct/incorrect) de e dans $\hat{s_d}$.

Cette étude nous permettra de savoir si le critère du risque est corrélé ou pas avec la fidélité du résumé et si le pourcentage d'entités incorrectes est plus élevé pour les entités en dehors du document source que pour les entités dans le document source. Nous avons sélectionné parmi les 77 résumés générés dans la section 3 les résumés qui ont un ROUGE score (Lin, 2004) maximal avec la référence pour ne pas être dépendant du système de résumé automatique et ainsi d'avoir un meilleur ROUGE score que tous les systèmes de l'état de l'art actuel.

Pour l'extraction des entités nommées, nous avons utilisé le système FLERT [4] (Schweter & Akbik, 2020) qui a été entraîné sur OntoNotes, un large corpus pour la tâche qui a 18 tags différents. Les 18 tags sont : nombre cardinal, date, événement, installation, entité géopolitique, langue, loi, lieu, argent, groupes, nombre ordinal, organisation, pourcentage, personne, produit, quantité, temps et œuvre d'art.

	summary	ROUGE	NEHR
CNN	ROUGE max	57.45 / 32.59 /41.63	4.6
	ROUGE min	30.04 / 09.33 /19.47	6.0
	Logit	41.99 / 18.96 / 28.01	5.6
XSUM	ROUGE max	60.14 / 35.68 / 51.20	45.91
	ROUGE min	27.43 / 07.51 / 21.46	47.39
	Logit	40.26 / 16.79 / 31.29	47.55

TABLE 1 – ROUGE (R-1/R-2/R-L) et NEHR sur les différents résumés produits par notre méthode d'échantillonnage. ROUGE Max, ROUGE min et Logit correspondent respectivement à la sélection du résumé avec le score ROUGE maximal par rapport à la référence la référence, du résumé avec le score ROUGE minimal avec la référence et du résumé ayant le logit maximal produit par le modèle.

Le tableau 1 montre les résultats obtenus en calculant le ROUGE Score ainsi que le critère NEHR sur CNN/DM et XSum. On voit dans le tableau 1 une variation considérable du ROUGE Score qui ne traduit pas une variation du critère NEHR. Cela peut s'expliquer par le fait que le ROUGE score n'est pas un bon indicateur de la fidélité du résumé par rapport au document source. Par contre on a une variation légère du NEHR. Ainsi, nous avons décidé d'annoter les résumés sélectionnés qui ont un ROUGE score maximal par rapport à la référence afin de savoir si le critère NEHR est un bon indicateur de la fidélité du résumé par rapport au document source.

Annotation des entités du meilleur résumé en terme de ROUGE score. Nous avons effectué une évaluation humaine pour savoir si les entités détectées comme risquées par le système d'entités nommées le sont vraiment et si les entités sont utilisées dans le bon contexte ou non.

4. https://huggingface.co/flair/ner-english-ontonotes

Ainsi, nous avons choisi de manière arbitraire 50 exemples du jeu de test de CNN/DM pour les annotations. Pour chaque résumé, nous avons choisi de manière aléatoire le même nombre d'entités hors du document que d'entités dans le document pour l'évaluation manuelle. Nous avons obtenu 145 entités hors du document source et 145 entités dans le document source soit un total de 290 entités à annoter.

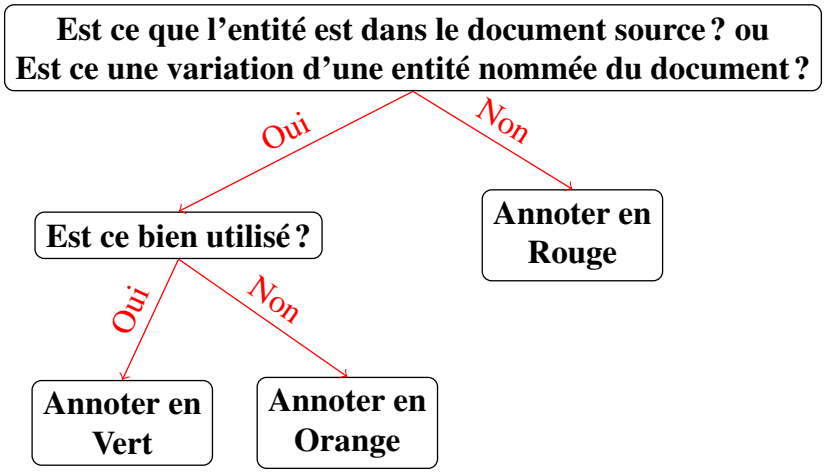


FIGURE 2 – Phase d'annotation des entités nommées. Vert : Les entités nommées qui sont dans le document ou des variations d'entités dans le document qui sont bien utilisées (ne contredisent pas le document) ; Orange : Les entités qui sont dans le document ou des variations d'entités du document qui sont mal utilisé ; Rouge : Les entités hors du document source.

Nous avons ensuite demandé aux annotateurs d'annoter les entités comme étant correctement ou mal utilisées lorsqu'elles sont dans le document source sinon si elles sont en dehors du document source, il s'agissait de vérifier qu'elles ne soient pas des variations d'entités dans le document. En présence de variation, il fallait annoter leur utilisation dans le résumé. Les variations sont les entités du document source qui ont des erreurs de frappe dans le résumé ou encore des dates écrites d'une autre manière. Un exemple de variation peut être *England* pour *Britain* ou encore *trentre trois* pour *33* ou *Balloteli* à la place de *Balotelli*... On definit une entité nommée du résumé comme étant bien utilisée si son emploi ne contredit pas le document source. La figure 2 nous présente la procédure d'annotation qui a été donnée à chaque annotateur. En reprenant l'exemple de la figure 1, on peut annoter les entités nommées du résumé en suivant la procédure de la figure 2. On obtient ainsi cette annotation sur la partie gauche de la figure 3. En vert, nous avons les entités qui ne contredisent pas le document source comme « Mark H Durkan » et « BBC ». En effet, c'est BBC qui a rapporté la news et c'est contre Mark H Durkan que le contrôle judiciaire est demandé mais ce n'est pas Helen Jones qui a demandé ce contrôle. Ainsi, l'entité « Helen Jones » est mal utilisée dans le résumé ; elle est donc mise en orange. On met l'entité « the Northern Ireland's environment minister » en rouge car elle ne provient pas du document. En effet, il n'est pas marqué dans le document qu'il s'agit du ministre de l'environnement ou pas.

Pour une évaluation plus poussée, nous avons décidé d'annoter les 145 entités hors du document pour évaluer leur factualité. Ainsi, les entités ayant été marquées comme hors du document source ont été annotées comme *variation*, *acceptable* ou *hallucination*. *Variation* correspond aux entités détectées comme hors du document mais qui sont en réalité des variation d'entités du document, *acceptable* correspond aux entités qui sont hors du document, mais factuellement correctes et *hallucination* aux entités qui ne sont pas factuellement correctes. Si l'on reprend l'exemple précédent de la figure 3,

FIGURE 3 – Exemple d'annotation des entités nommées suivant le guide de la figure 2.

l'entité **the Northern Ireland's environment minister** n'est pas dans le document source. Mais en se basant sur des connaissances générales, on sait que Mark H Durkan était le ministre de l'environnement de l'Irlande du Nord. Ainsi, on peut la marquer comme acceptable.

Cela nous permet d'avoir deux étages d'annotation, le premier pour savoir le type d'entités (si une entité est dans ou hors du document source) et le second pour savoir si elle est bien utilisée ou non. Nous avons collecté les annotations manuelles de 3 annotateurs. De plus, nous avons calculé l'accord inter-annotateur pour les annotations. Le tableau 2 présente le coefficient Cohen Kappa entre les 3 annotateurs. Le coefficient étant supérieur à 0.63 dans les 3 cas, cela est un indicateur d'un accord important entre les annotateurs.

A1 - A2	A1 - A3	A2 - A3
0.6833	0.6468	0.6302

TABLE 2 – Accord inter-annotateurs Cohen kappa (Cohen, 1960) entre les 3 annotateurs.

Les résultats obtenus pour un annotateur choisi aléatoirement sont consignés dans les tableaux 3a et 3b. Pour les entités hors du document, nous avons permis aux annotateurs de se référer à leur connaissance générale, mais également à Internet pour le second étage d'annotation.

Le tableau 3a montre que pour les entités en dehors du document, 90% sont bien utilisées (ne contredisent pas le document source). Cela signifie que quand une entité du résumé généré est dans le document source, la plupart du temps, elle est bien utilisée. Pour ce qui est des entités en dehors du document source, nous avons 71% qui sont bien utilisées. Le tableau 3b contient la distribution des entités hors du document selon les 3 labels présentés plus haut (*variation, acceptable et hallucination*). Parmi les entités hors du document, 59% sont des variations d'entités du document, 17% sont acceptables (inférables) et 29% sont des hallucinations (n'étant pas dans le document source). Pour ce qui est des variations et des entités acceptables, environ 90% sont bien utilisées dans les résumés (ligne %correct). Ce qui est aligné aux résultats obtenus quand une entité est dans le document source. Ainsi, la proportion qui vient des hallucinations est de 30%. Donc, en minimisant le nombre d'entités en dehors du document source, nous réduisons le risque d'hallucination sur les entités nommées et cela permet d'augmenter la fidélité du résumé par rapport au document source.

Dans la section suivante, nous avons utilisé le critère du risque pour sélectionner des résumés. Cela nous permettra de voir l'impact du critère sur la qualité des résumés.

	%correct
in-doc	90.3
out-doc	71.0

(a) % d'entités correctement utilisées dans les résumés générés selon l'annotation manuelle.

	variation	accept.	hallucination
distribution	59.3%	11.7%	29%
%correct	90	88.0	0

(b) Distribution des entités hors du document selon les 3 partitions avec le pourcentage d'entités correctement utilisées pour chaque ensemble.

TABLE 3 – Résultats des annotations des entités nommées du résumé avec le score ROUGE maximal avec la référence.

4.4 Sélection de résumé en utilisant le NEHR

L'idée est d'évaluer l'impact de l'utilisation du critère NEHR pour la sélection de résumés parmi plusieurs résumés pendant l'inférence. Ainsi, dans cette première partie, nous présentons le processus de sélection de résumés automatiques basé sur le NEHR mis au point. Pour évaluer l'efficacité de ce critère, nous avons évalué les résumés sélectionnés à l'aide des mesures d'évaluation habituelles pour le résumé telles que le ROUGE Score et le BERT Score (Lin, 2004; Zhang* *et al.*, 2020) mais également en suivant la valeur du NEHR (équation 1).

Description Les résumés sélectionnés en utilisant le critère NEHR ont été comparés avec deux baselines (le résumé avec le plus grand logit et le meilleur résumé avec un beam de taille 4). Cette comparaison a été faite en termes de ROUGE score et BERT score.

Pour la sélection de résumés, nous avons proposé un critère basé sur le critère NEHR mais également sur la performance du modèle. D'abord, nous sélectionnons les résumés qui ont un NEHR minimal. Puis, le résumé sélectionné est celui qui a le plus grand logit. Supposons H l'ensemble des résumés obtenus par échantillonnage du modèle, V l'ensemble des résumés avec un risque minimal, $P(\cdot|model)$ est la probabilité donnée par le modèle a un résumé et $\hat{s}$ est le résumé sélectionné :

$$V = \left\{ x \in H | risk(x) = \min_{s' \in H} NEHR(s') \right\} \tag{2}$$

$$\hat{s} = \underset{s \in V}{\operatorname{argmax}} P(s|model) \tag{3}$$

Mésure automatique Nous avons utilisé le ROUGE score et le BERTScore (Lin, 2004; Zhang* *et al.*, 2020) pour évaluer le résumé. Les résultats sont consignés dans le tableau 4. Ce tableau montre que BART-Large avec un *beam = 4* donne le meilleur résultat selon les différentes métriques aussi bien pour CNN/DM que pour XSum. Notre approche minimise bien le risque d'hallucination en ayant des scores ROUGE et BERTScore équivalents aux autres approches pour CNN/DM. Cependant pour XSum, on perd en ROUGE Score.

Le NEHR dépend de la capacité à retrouver les entités dans le document source. Ainsi, il est donc important de détecter efficacement les entités nommées qui peuvent être des variations d'entités nommées du document. La section suivante présente les heuristiques introduites pour la détection de variation.

		R-1	R-2	R-L	BERTScore	NEHR
CNN	BEAM 4	**43.74**	**20.84**	**30.44**	**32.00 / 88.52**	1.53
	BEST LOGIT	41.99	18.96	28.01	29.89 / 88.17	5.60
	MIN NEHR (ours)	42.31	19.21	28.41	30.36 / 88.25	**0.02**
XSUM	BEAM 4	**45.32**	**22.20**	**37.10**	**51.56 / 91.82**	39.84
	BEST LOGIT	40.26	16.79	31.29	45.23 / 90.76	47.55
	MIN NEHR (OUR)	40.05	16.41	31.33	45.79 / 90.85	**18.78**

TABLE 4 – Evaluation des résumés de CNN/DM et XSum en terme de (R-1, R-2, R-L), BERTScore et NEHR. Les deux valeurs du BERTScore correspondent au score avec et sans le paramètre rescale. *NEHR* est le pourcentage d'entités en dehors du document source. BEAM 4 correspond à la génération avec beam égale à 4. BEST LOGIT correspond à la sélection des résumés avec le meilleur logit. Et, MIN NEHR est notre méthode de sélection de résumé décrite dans la section 4.4.

5 Détection de variation

59% des entités considérées comme hors du document étant des variations d'entités dans le document source (voir tableau 3b). Nous avons mis en place une détection automatique de variations afin d'avoir une meilleure précision sur les entités qui sont hors du document.

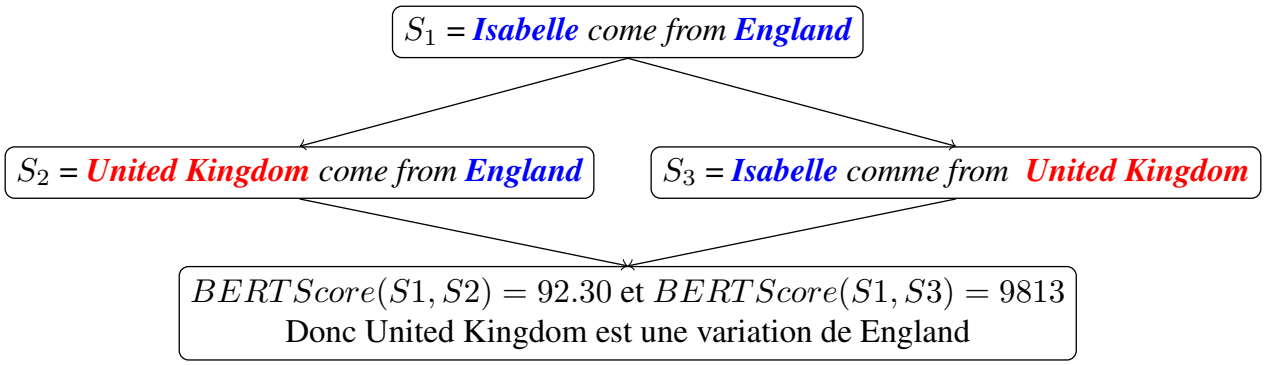


FIGURE 4 – BERTScore pour la détection de variation

Heuristiques de détection de variation Nous avons introduit plusieurs heuristiques pour la détection des variations :

— *Processing* : Consiste à supprimer dans le texte les caractères indésirables. Dans le cas d'un corpus anglais vérifié si ce n'est pas l'indicatif de la possession qui est gênant.
— *Stemming* : Pour la transformation des flexions en leur radical ou racine.
— *Distance de levenshtein* [5] : Elle permet la correction orthographique des mots. Pour un mot et un texte donné, il s'agit de retrouver dans le texte, des mots qui sont proches du mot en entrée lorsqu'on applique une suppression de caractère, un remplacement, une transposition ou encore une insertion. Ainsi, on regarde la distance entre les mots.
— *Alpha2Digit* [6] : On utilise un package python qui convertit les nombres écrits en lettre dans un texte en chiffre.

5. https://norvig.com/spell-correct.html
6. https://pypi.org/project/text2num/

— *SplitWord* : L'idée est de traiter les entités qui viennent d'un groupe de mots. De cette façon, on regarde de manière individuelle si chaque mot du groupe appartient au document source.
— *BERTScoreCheck* : Le BERTScore (Zhang* *et al.*, 2020) étant basé sur des représentations contextuelles, l'idée est de retrouver une entité du document source proche de l'entité qu'on veut tester. Le BERTScore étant une mesure entre des phrases, nous avons mis au point un moyen de l'utiliser pour la détection de variation. La figure 4 donne un exemple. Supposons que la phrase $S1$ provient du document, et « United Kingdom » une variation d'une entité du document. Ici, on la remplace par chaque entité pour créer de nouvelles phrases puis on calcule le BERTScore des phrases obtenues vis-à-vis de la phrase de départ $S1$. Le score le plus élevé entre $S1$ et $S2$ revient à dire que United Kingdom est une variation de l'entité « England ».

Pour créer les règles de détection de variation, nous avons conçu un corpus de variation à partir de CNN/DM et avons sélectionné 91 variations. Le tableau 5 montre le pourcentage de variations détectées par chaque heuristique puis par la combinaison de différentes heuristiques. Cette combinaison se fait en utilisant successivement chaque heuristique pour la détection des variations dans l'ordre suivant : Processing Processing + Stemming + Alpha2digit > Levenshtein > SplitWord > BertScoreCheck.

Méthodes	% **de variations détectées**
Processing	14
Levenshtein	24
SplitWord	30
BertScoreCheck	**39**
Processing + Stemming + Alpha2digit	44
Levenshtein + SplitWord	42
BertScoreCheck + Processing + Stemming + Alphadigit	53
Levenshtein + SplitWord + Processing + Stemming	**59**
All method combined	**60**

TABLE 5 – Pourcentage de variations détectées pour chaque heuristique.

Nous avons ensuite créé un autre corpus contenant 43 variations et 26 entités hors du document source pour savoir si les heuristiques ne détectent pas des entités comme des variations pourtant en réalité elles n'appartiennent pas au document source. On constate que, sans l'utilisation du BERT score (5), on est à 59% du pourcentage d'entités détectées pourtant avec nous sommes à 60%. Pour vérifier ce résultat, nous avons utilisé la combinaison de méthodes pour détecter les entités dans ou en dehors du document source avec ou sans l'utilisation de l'heuristique BERTScoreCheck.

Le tableau 6 présente les résultats. On peut voir que l'utilisation du BERTScore introduit des erreurs dans la détection car certaines entités qui sont hors du document ont été qualifiées comme étant dans le document (voir le score de précision sur les variations). Pour le coût de la méthode, le gain obtenu n'est pas celui attendu. Ainsi, nous avons utilisé uniquement la combinaison des autres heuristiques basées sur les règles.

Ayant obtenu un système pour la détection de variations, nous avons sélectionné à nouveau les résumés générés par la méthode de sampling avec le NEHR comme critère de sélection en prenant en compte le fait que les variations d'entité dans le document sont des entités du document. Le tableau 7 présente les résultats obtenus. Aux différentes méthodes présentées dans le tableau 4, nous avons

	Sans BertScore			Avec BertScore		
	précision	rappel	f1 score	précision	rappel	f1 score
Variation	**1.00**	0.60	**0.75**	**0.90**	0.65	**0.76**
Hors du document	0.60	**1.00**	**0.75**	0.61	**0.88**	0.72

TABLE 6 – Précision, Rappel et F1 score pour la détection de variation et d'entités hors document

également utilisé une méthode basée sur de l'entailment. C'est-à-dire le résumé sélectionné est celui qui à la similarité la plus élevée avec le document source.

		R-1	R-2	R-L	NEHR	%HallDoc
CNN/DM	BEAM 4	**43.74**	**20.84**	**30.44**	0.5	3.86
	BEST LOGIT	41.99	18.96	28.01	2.6	20.57
	ENTAILMENT	43.61	19.69	29.26	1.62	12.92
	MIN NEHR + VAR	42.19	19.12	28.24	**0.003**	**0.035**
XSUM	BEAM 4	**45.32**	**22.20**	**37.10**	27.67	52.48
	BEST LOGIT	40.26	16.79	31.29	31.05	61.24
	ENTAILMENT	40.92	17.14	31.96	27.08	54.98
	MIN NEHR + VAR	40.16	16.54	31.31	**6.92**	**21.49**

TABLE 7 – Évaluation du résumé généré sur les jeux de données CNN/DM et XSum. ROUGE (R-1, R-2, R-L) pour les différents critères de sélection. *NEHR* est le pourcentage d'entités en dehors du document source. Pour BEAM 4, BEST LOGIT, ENTAILMENT la valeur de *NEHR* est avant et après détection de variation BEST LOGIT correspond au moment où nous sélectionnons le meilleur logit de toutes synthèses générées. Et MIN NEHR + VAR est notre méthode de sélection de résumé proposée après l'utilisation de la détection de variation . %HallDoc correspond au pourcentage d'exemples avec au moins une entité hors du document.

Bien que les heuristiques de détection de variations ont été testées sur CNN/DM, on voit l'efficacité sur XSUM en comparant les résultats obtenus dans le tableau 4 aux résultats du tableau 7. Pour les deux datasets, MIN NEHR + VAR donne de meilleurs résultats et réduit bien la quantité d'entités hors du document source. Au niveau du ROUGE, nous avons des scores comparables pour CNN/DM mais un écart en ce qui concerne XSUM. Pour comprendre l'impact du critère NEHR sur les différents datasets, nous avons calculé le pourcentage d'exemples avec au moins une entité hors du document (colonne %HallDoc). Pour XSUM on divise de plus de 2 fois le nombre d'exemples avec des entités hors du document. De même, nous avons calculé le pourcentage d'exemples avec au moins une entité hors du document pour les résumés de référence de chaque corpus afin de comprendre pourquoi le pourcentage d'exemple avec au moins une entité hors du document est élevé. Les résultats sont consignés dans le tableau 8. Ces résultats nous montrent que plus de la moitié des références du corpus XSum ont au moins une entité hors du document soit 2 fois plus que le corpus CNN/DM. Cela montre l'abstractivité du corpus XSum.

Les résultats obtenus dans cette partie montrent que les systèmes tentent de reproduire le caractère abstrait du corpus XSum qui a 63% des exemples du jeu de test qui ont une entité en dehors du document. On constate qu'en utilisant notre critère de sélection de résumé, ce pourcentage est réduit à 21% ce qui peut augmenter la fidélité des entités nommées au regard du document source.

Dataset	% HallDoc	#lines
Cnn/DM	29.33	11490
Xsum	63.65	11333

TABLE 8 – Le pourcentage de résumé de référence avec au moins une entité hors du document source

6 Conclusion et Discussion

En somme, nous avons proposé d'utiliser la méthode de sampling pour générer des résumés couvrant un large espace de recherche afin de sélectionner des résumés ayant le moins d'hallucination au niveau des entités. Cette problématique est très importante du fait que pour une utilisation industrielle, il est important que les entités nommées ne soient pas hallucinées. Notre étude empirique du risque nous a montré que les entités dans le document sont à 90% du temps bien utilisées. Cette étude nous a permis de comprendre que plusieurs entités taguées comme hors du document sont à 59% des variations, et de nous pencher sur la détection de variations au moyen d'heuristique. L'utilisation du critère NEHR combiné aux prédictions du modèle pour la sélection du résumé réduit le risque d'hallucination sur les entités nommées.

Limites Cette étude présente plusieurs limites notamment le fait que le NEHR soit dépendant de la qualité du système de reconnaissance automatique d'entités nommées. Nous avons fait l'annotation humaine des entités uniquement sur 50 paires de données de résumé-document et quelques entités sélectionnées dans le résumé. Les résumés obtenus par l'utilisation du critère NEHR combiné à la prédiction du modèle n'ont pas été évalués manuellement. Il peut être intéressant d'évaluer les différentes sorties.

Remerciements

Remerciement à Frédéric Bechet et Benoit Favre pour l'encadrement et Romain Gemignani, responsable innovation à Enedis.

Ces travaux ont bénéficié d'un accès aux moyens de calcul de l'IDRIS au travers de l'allocation de ressources 2021-AD011012525R2 attribuée par GENCI.

Références

AKANI E., FAVRE B. & BECHET F. (2022). Abstraction ou hallucination ? état des lieux et évaluation du risque pour les modèles de génération de résumés automatiques de type séquence-à-séquence (abstraction or hallucination ? status and risk assessment for sequence-to-sequence automatic). In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, p. 2–11, Avignon, France : ATALA.

BANERJEE S. & LAVIE A. (2005). METEOR : An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic*

and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, p. 65–72, Ann Arbor, Michigan : Association for Computational Linguistics.

CAO Z., WEI F., LI W. & LI S. (2018). Faithful to the original : Fact aware neural abstractive summarization. *ArXiv*, **abs/1711.04434**.

CHEN S., ZHANG F., SONE K. & ROTH D. (2021). Improving faithfulness in abstractive summarization with contrast candidate generation and selection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 5935–5941, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-main.475.

COHEN J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**(1), 37–46. DOI : 10.1177/001316446002000104.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

DURMUS E., HE H. & DIAB M. (2020). FEQA : A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 5055–5070, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.454.

FALKE T., RIBEIRO L. F. R., UTAMA P. A., DAGAN I. & GUREVYCH I. (2019). Ranking generated summaries by correctness : An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 2214–2220, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1213.

FAN A., GRANGIER D. & AULI M. (2018a). Controllable abstractive summarization. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, p. 45–54, Melbourne, Australia : Association for Computational Linguistics. DOI : 10.18653/v1/W18-2706.

FAN A., LEWIS M. & DAUPHIN Y. (2018b). Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 889–898, Melbourne, Australia : Association for Computational Linguistics. DOI : 10.18653/v1/P18-1082.

HERMANN K. M., KOČISKÝ T., GREFENSTETTE E., ESPEHOLT L., KAY W., SULEYMAN M. & BLUNSOM P. (2015). Teaching machines to read and comprehend. DOI : 10.48550/ARXIV.1506.03340.

HOLTZMAN A., BUYS J., DU L., FORBES M. & CHOI Y. (2019). The curious case of neural text degeneration. DOI : 10.48550/ARXIV.1904.09751.

JI Z., LEE N., FRIESKE R., YU T., SU D., XU Y., ISHII E., BANG Y. J., MADOTTO A. & FUNG P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, **55**(12), 1–38. DOI : 10.1145/3571730.

KRYSCINSKI W., KESKAR N. S., MCCANN B., XIONG C. & SOCHER R. (2019). Neural text summarization : A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 540–551, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1051.

LEWIS M., LIU Y., GOYAL N., GHAZVININEJAD M., MOHAMED A., LEVY O., STOYANOV V. & ZETTLEMOYER L. (2020). BART : Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7871–7880, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.703.

LIN C.-Y. (2004). ROUGE : A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, p. 74–81, Barcelona, Spain : Association for Computational Linguistics.

MAYNEZ J., NARAYAN S., BOHNET B. & MCDONALD R. (2020). On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 1906–1919, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.173.

NALLAPATI R., ZHOU B., DOS SANTOS C., GULCEHRE C. & XIANG B. (2016). Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, p. 280–290, Berlin, Germany : Association for Computational Linguistics. DOI : 10.18653/v1/K16-1028.

NAN F., NALLAPATI R., WANG Z., NOGUEIRA DOS SANTOS C., ZHU H., ZHANG D., MCKEOWN K. & XIANG B. (2021). Entity-level factual consistency of abstractive text summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : Main Volume*, p. 2727–2733, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.eacl-main.235.

NARAYAN S., COHEN S. B. & LAPATA M. (2018). Don't give me the details, just the summary ! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 1797–1807, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/D18-1206.

NARAYAN S., ZHAO Y., MAYNEZ J., SIMÕES G., NIKOLAEV V. & MCDONALD R. (2021). Planning with learned entity prompts for abstractive summarization. *Transactions of the Association for Computational Linguistics*, **9**, 1475–1492. DOI : 10.1162/tacl_a_00438.

PAGNONI A., BALACHANDRAN V. & TSVETKOV Y. (2021). Understanding factuality in abstractive summarization with FRANK : A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 4812–4829, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-main.383.

SCHWETER S. & AKBIK A. (2020). Flert : Document-level features for named entity recognition.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. & POLOSUKHIN I. (2017). Attention is all you need.

WANG A., CHO K. & LEWIS M. (2020). Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 5008–5020, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.450.

WOLF T., DEBUT L., SANH V., CHAUMOND J., DELANGUE C., MOI A., CISTAC P., RAULT T., LOUF R., FUNTOWICZ M., DAVISON J., SHLEIFER S., VON PLATEN P., MA C., JERNITE Y., PLU J., XU C., LE SCAO T., GUGGER S., DRAME M., LHOEST Q. & RUSH A. (2020). Transformers : State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, p. 38–45, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-demos.6.

ZHANG* T., KISHORE* V., WU* F., WEINBERGER K. Q. & ARTZI Y. (2020). Bertscore : Evaluating text generation with bert. In *International Conference on Learning Representations*.

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Mise en place d'un modèle compact à architecture Transformer pour la détection jointe des intentions et des concepts dans le cadre d'un système interactif de questions-réponses

Nadège Alavoine[1, 2] Arthur Babin[1, 3]
(1) Université Paris-Saclay, LISN, Campus Universitaire bâtiment 507, Rue du Belvédère, 91400 Orsay, France
(2) École 42, 96 boulevard Bessières, 75017 Paris, France
(3) ENSIIE, 1 square de la Résistance, 91000 Évry-Courcouronnes
prénom.nom@lisn.fr, prénom.nom@ensiie.fr

RÉSUMÉ

Les tâches de détection d'intention et d'identification des concepts sont toutes deux des éléments importants de la compréhension du langage naturel. Elles sont souvent réalisées par deux modules différents au sein d'un *pipeline*. L'apparition de modèles réalisant conjointement ces deux tâches a permis d'exploiter les dépendances entre elles et d'améliorer les performances obtenues. Plus récemment, des modèles de détection jointe reposant sur des architectures *Transformer* ont été décrits dans la littérature. Par ailleurs, avec la popularité et taille croissante des modèles *Transformer* ainsi que les inquiétudes ergonomiques et écologiques grandissantes, des modèles compacts ont été proposés. Dans cet article, nous présentons la mise en place et l'évaluation d'un modèle compact pour la détection jointe de l'intention et des concepts. Notre contexte applicatif est celui d'un système interactif de questions-réponses français.

ABSTRACT

Implementation of a light model with Transformer architecture for joint intent classification and slot filling aimed at an interactive question-answers system

Intent classification and slot filling are important tasks of Natural Language Understanding. They're usually performed by two distinct modules inserted in one pipeline. Models conducting both tasks emerged in literature and improved previous performances by exploiting the dependencies between them. More recently, models performing intent detection and slot filling based on *Transformer*'s architecture were described. On another note, with the growing popularity and size of *Transformer*'s models as well as increasing ecological and ergonomic concerns, light versions were proposed. This article presents the implementation and evaluation of a joint detection light model for a French interactive question answering system.

MOTS-CLÉS : Détection jointe, Détection d'intention, Identification de concepts, Transformer, BERT joint, Classifieur DIET, CamemBERT, FrALBERT.

KEYWORDS: Joint detection, Intent classification, Slot filling, Transformer, Joint BERT, DIET classifier, CamemBERT, FrALBERT.

Actes des 16e Rencontres Jeunes Chercheurs en RI (RJCRI) et 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL), pages 37–56.

1 Introduction

La compréhension du langage naturel ou NLU (*natural language understanding)* est un domaine classique du traitement de la parole transcrite ainsi qu'un élément essentiel aux systèmes de dialogues. Son but est d'extraire les concepts sémantiques du discours. Elle consiste notamment en la détection de l'intention et l'identification des concepts d'une phrase (ou tâche de *slot filling*). Ces deux tâches sont illustrées en Tableau 1 avec l'exemple *"Trouve les horaires de la bibliothèque municipale demain"*. Dans cet exemple, l'intention est d'obtenir les horaires d'ouverture d'un lieu précis à une date précise : deux concepts devant être identifiés pour formuler la réponse attendue.

Mots	trouve	les	horaires	de	la	bibliothèque	municipale	demain
C	O	O	O	O	O	B-lieu	I-lieu	B-date
I	obtenir_les_horaires							

FIGURE 1 – Exemple de discours avec étiquettes d'intention (**I**) et de concepts (**C**). Les concepts sont étiquetés selon la norme BIO.

Ces deux tâches sont fréquemment réalisées par des modules indépendants insérés dans un même *pipeline* et ne partageant pas directement d'informations entre-eux (Hakkani-Tür *et al.*, 2016; Goo *et al.*, 2018). En conséquence, le *pipeline* peut souffrir de propagation et d'accumulation d'erreurs. Depuis une quinzaine d'années (Weld *et al.*, 2022), des modèles réalisant conjointement ces tâches ont été proposés. Ils reposent sur différentes stratégies impliquant notamment des champs aléatoires conditionnels (Jeong & Lee, 2008), des réseaux neuronaux convolutifs (Xu & Sarikaya, 2013), des réseaux neuronaux récurrents (Guo *et al.*, 2014; Hakkani-Tür *et al.*, 2016; Liu & Lane, 2016), des modèles avec *slot-gate* (Goo *et al.*, 2018) ou des mécanismes d'attention (Chen *et al.*, 2016; Liu & Lane, 2016). Cette détection jointe permet d'exploiter les dépendances entre les deux tâches et d'améliorer les performances obtenues. Certains de ces travaux, réalisés principalement sur des données en langue anglaise, ont pu être adaptés pour d'autres langues en conservant les mêmes architectures (Weld *et al.*, 2022).

Plus récemment, l'apparition des modèles *Transformer* (Vaswani *et al.*, 2017) utilisant des mécanismes d'attention a permis de réaliser de nombreux progrès dans le domaine du Traitement Automatique des Langues (TAL). Des modèles pré-entraînés reposant sur des variations de cette architecture, tel que le modèle BERT (Devlin *et al.*, 2019), ont permis d'atteindre de nouveaux états de l'art pour de multiples tâches du TAL. Des modèles s'appuyant sur ces architectures *Transformer* pour réaliser une détection jointe de l'intention et des concepts ont été proposés dans la littérature (Chen *et al.*, 2019; Castellucci *et al.*, 2019; Bunk *et al.*, 2020). Ces modèles permettent d'obtenir de meilleures performances que les précédents modèles présentés pour la détection jointe.

Parallèlement, une tendance à la création de modèles de langues de tailles croissantes, en termes de données et de nombre de paramètres, est constatée. Si cette augmentation permet d'obtenir des performances grandissantes, elle s'associe à un coût financier ainsi qu'écologique (Strubell *et al.*, 2019; Moosavi *et al.*, 2020; Bender *et al.*, 2021). De plus, de trop grands modèles ne sont pas toujours utilisables en fonction de limites matérielles. Afin de répondre à ces problématiques, des modèles de langues compacts aux performances similaires à ceux de plus grandes tailles sont proposés (Cattan *et al.*, 2022). C'est notamment le cas d'ALBERT (Lan *et al.*, 2020), version allégée de BERT, et de son équivalent pré-entraîné sur un corpus français FrALBERT (Cattan *et al.*, 2021). Tous deux sont

optimisés grâce à des méthodes de réduction et de partage des poids.

Nos contributions principales sont l'élaboration d'un corpus en langue française de questions destinées à un système de dialogue fournissant des renseignements généraux sur des bibliothèques universitaires, son annotation en intentions et concepts pour des tâches de NLU, la mise en place d'un modèle compact réalisant une détection jointe des intentions et concepts évalué sur ce corpus et sa comparaison à un modèle de plus grande taille. Le contexte applicatif est décrit en section 2. Les modèles présentés reposent sur des architectures existantes, présentées en section 3, avec l'utilisation de modèles pré-entraînés sur des corpus français. Les données utilisées, les optimisations des modèles et les protocoles expérimentaux, ainsi que les résultats et leurs analyses sont exposés en section 4.

2 Contexte

Notre problématique initiale concerne la création d'un système de compréhension du langage lorsque aucune donnée n'est disponible. Notre contexte applicatif est le projet DIBISO. Ce projet est issu d'une collaboration entre la DIBISO (DIRECTION DES BIBLIOTHÈQUES, DE L'INFORMATION ET DE LA SCIENCE OUVERTE) et le LISN (LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE). Il consiste en l'élaboration d'un système interactif de questions-réponses (SQR) destiné à fournir des renseignements relatifs aux BIBLIOTHÈQUES UNIVERSITAIRES PARIS-SACLAY. Le LISN et la DIBISO étant engagés sur le plan écologique et l'impact de l'intelligence artificielle sur l'environnement n'étant pas négligeable (Strubell *et al.*, 2019; Bender *et al.*, 2021), il fut décidé que les solutions adoptées pour ce projet s'inscriraient dans une utilisation raisonnée des ressources.

Une première étude de faisabilité révéla l'absence de corpus disponible répondant précisément à ce besoin. Nous avons réalisé une campagne de collecte de données annotées. Dans ce but, un prototype fut créé et mis à disposition d'une vingtaine d'annotateurs sélectionnés par la DIBISO. Les données utilisées pour la conception de ce prototype ont été fournies par la plateforme UBIB [1], permettant de mettre en relation les usagers de certaines bibliothèques universitaires avec des bibliothécaires à travers une interface de messagerie instantanée. Afin d'en permettre un déploiement rapide, la partie agent de dialogue de ce prototype fut construite avec la librairie RASA [2] (Bocklisch *et al.*, 2017) destinée à la mise en place aisée de systèmes conversationnels (*chatbot*). Les données initiales furent découpées en corpus d'entraînement et d'évaluation puis augmentées par des données générées. Techniquement, le prototype réalise d'abord une identification jointe de l'intention et des concepts des questions des utilisateurs à travers un premier modèle reposant sur un *Transformer*. Ce modèle repose sur une architecture DIET (Bunk *et al.*, 2020). Selon l'intention détectée, la réponse fournie est soit générique, soit générée en fonction des concepts identifiés ou fait appel à d'autres modèles pour identifier le passage le plus pertinent dans une base de données issues des pages internet des bibliothèques universitaires. Grâce à ce prototype, 471 questions ont été récupérées. L'analyse des erreurs obtenues nous laisse suggérer qu'une amélioration du système d'identification jointe de l'intention et des concepts est nécessaire pour permettre au SQR de fournir des réponses plus pertinentes. Par ailleurs, bien que modulable, la librairie RASA présente un effet « boîte noire » limitant les manipulations possibles.

Dans l'optique de construire un nouveau SQR plus manipulable et permettant toujours une détection jointe de l'intention et des concepts d'une question, un modèle reposant sur une architecture jointe BERT (Chen *et al.*, 2019) est envisagé.

1. https://ubib.fr
2. https://rasa.com

3 Établissement de modèles compactes pour la détection jointe

Dans cette section, nous aborderons l'architecture et les modalités de mise en place du modèle utilisé dans le prototype, ainsi que du modèle visant à le remplacer.

3.1 Préliminaire : Détection jointe à l'aide du classifieur DIET

Le classifieur DIET (pour *Dual Intent and Entity Transformer*) est une architecture *Transformer* légère adaptée à la compréhension du langage. Ce classifieur, présenté par l'équipe de recherche de la plateforme RASA (Bunk *et al.*, 2020) et intégré à leur librairie, réalise la détection jointe de l'intention et des concepts d'une phrase. Son architecture, telle que décrite dans l'article de Bunk *et al.* (2020), est présentée ci-dessous.

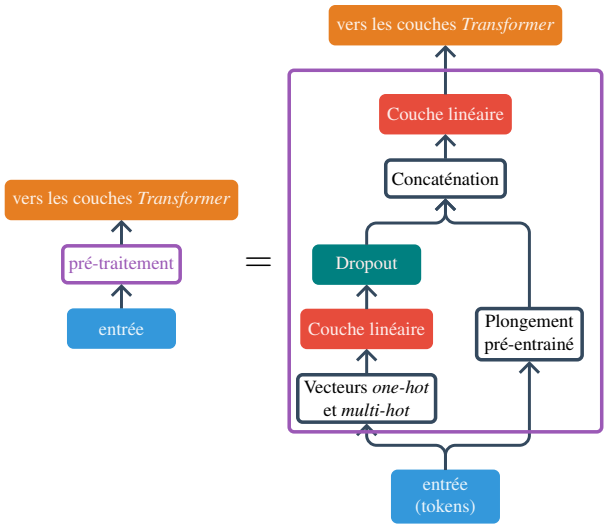


FIGURE 2 – Description du pré-traitement des tokens dans le modèle DIET

Les données sont d'abord étiquetées selon le format BILOU (Ramshaw & Marcus, 1995). Ainsi, le premier mot constitutif d'un concept est identifié par B (pour *Beginning*), les suivants par I (pour *Inside*) et le dernier par L (pour *Last*), suivis de l'étiquette associée. Les mots non constitutifs d'un concept sont identifiés par O (pour *Outside*). Les concepts d'une seule unité sont aussi identifiés, par la lettre U (pour *Unit-length*). Cet étiquetage est plus complet que BIO, plus fréquemment utilisé, où la première lettre d'un concept sont identifiées par B et les suivantes par I. Un pré-traitement des données, représenté en Figure 2, est ensuite réalisé : les phrases présentées en entrée au modèle sont transformées en représentations denses et éparses. Les premières sont obtenues à partir de plongements lexicaux pré-entraînés tels que ceux issus de la dernière couche de BERT (Devlin *et al.*, 2019). Les secondes sont obtenues par un processus de tokenisation suivi d'un encodage vectoriel *one-hot* et *multi-hot* de n-grammes de caractères (avec $n \leq 5$). Les informations des représentations éparses pouvant être redondantes, un *dropout* leur est appliqué pour éviter un phénomène de sur-apprentissage. Les représentations éparses sont alignées aux dimensions des denses par une couche entièrement connectée. Un token **[CLS]** représentant la classification d'une phrase (Devlin *et al.*, 2019) est ajouté à la fin de chaque entrée. Les différentes représentations sont ensuite concaténées. Ces représentations peuvent être affinées par apprentissage du modèle, ou être gelées. Si dans l'article de Bunk *et al.* (2020), un système de masquage des tokens associé à sa fonction de coût est décrit, les résultats de leur étude d'ablation montrent que ce système peut légèrement diminuer les performances des modèles. Le calcul de ce coût est donc inactivé par défaut dans l'outil proposé par RASA.

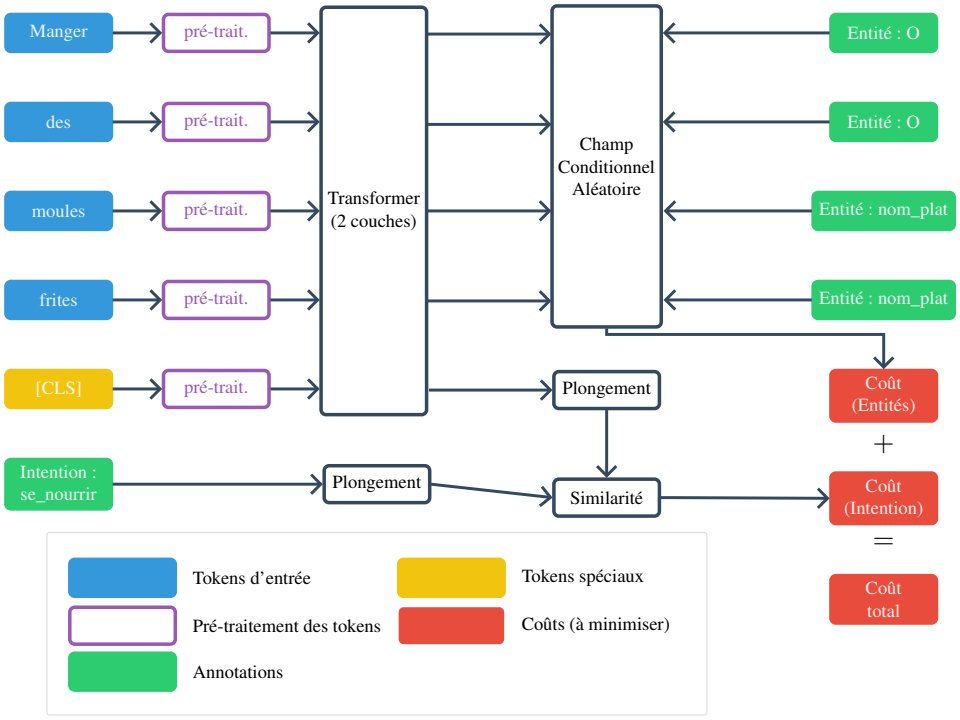


FIGURE 3 – Schéma de l'architecture du classifieur DIET d'après Bunk *et al.* (2020). La version présentée ne comporte pas le masquage des mots, inactive par défaut dans la librairie RASA.

Par la suite, et comme représenté en Figure 3, ces représentations sont présentées à deux couches successives d'un *Transformer* (Vaswani *et al.*, 2017) avec un système d'attention aux positions relatives (Shaw *et al.*, 2018). Afin que les dimensions des représentations soient adaptées à celle du *Transformer*, un passage par une couche entièrement connectée est réalisé au préalable. La prédiction des concepts de la phrase est réalisée à partir des sorties de la seconde couche *Transformer*, passées à une couche d'étiquetage à champ aléatoire conditionnel (Lafferty *et al.*, 2001). Un coût lié aux entités est calculé par log-vraisemblance négative entre la séquence prédite et celle attendue (Lample *et al.*, 2016). La sortie de la seconde couche *Transformer* pour le token **[CLS]** et les étiquettes d'intentions sont intégrées à un unique espace vectoriel sémantique. Un coût par produit scalaire de similarité est calculé (Wu *et al.*, 2018; Henderson *et al.*, 2019; Valsov *et al.*, 2019). Le calcul de ce coût vise à maximiser les similarités entre les prédictions issues de l'espace vectoriel sémantique pour la totalité des tokens **[CLS]** et leur étiquette positive (cible), ainsi qu'à minimiser celles avec les étiquettes négatives (non-cibles). Pour l'inférence, un calcul de similarité par produit scalaire permet d'ordonner les différentes intentions possibles. Lors de l'entraînement d'un classifieur DIET, c'est le coût total qui est minimisé. Ce coût est la somme de ceux calculés pour l'identification des concepts et pour la détection d'intention. Afin de réduire l'effet de déséquilibre des classes (Japkowicz & Stephen, 2002), les lots d'entraînement de données sont arrangés selon une stratégie d'équilibrage (Valsov *et al.*, 2019). La taille de ces lots augmente avec les itérations de l'entraînement dans une optique de régularisation (Smith *et al.*, 2018).

Lors de la conception du prototype, le choix s'est porté sur ce classifieur en raison de son intégration à la librairie RASA. Cette librairie avait été choisie en fonction de contraintes techniques. L'aspect modulaire de RASA a permis d'y intégrer le modèle FrALBERT$_{base}$ (Cattan *et al.*, 2021), non proposé initialement par la librairie. FrALBERT est utilisé dans ce classifieur pour l'obtention des plongements lexicaux et en tant que couches *Transformer*. Seules ces représentations par plongements lexicaux sont utilisées dans notre configuration du classifieur DIET, sans concaténation avec des représentations éparses. Le modèle résultant sera désigné comme DIET FrALBERT.

Concernant FrALBERT, il s'agit d'un modèle compact de langage basé sur l'architecture *Transformer*, pré-entraîné sur un corpus français. Comme sa version anglaise, ALBERT (Lan *et al.*, 2020), FrALBERT utilise des méthodes de partage et de réduction des paramètres permettant de réduire sa complexité et d'accélérer ses phases d'entraînement comme d'inférence. FrALBERT$_{base}$ a été entraîné sur la version française du corpus de l'encyclopédie Wikipédia comprenant 4 gigabytes de texte (abrégé *wiki-4GB*).

RASA offre la possibilité d'utiliser un autre modèle pré-entraîné sur un corpus français et basé sur une architecture *Transformer* : le modèle CamemBERT (Martin *et al.*, 2020) dont l'architecture se rapproche de RoBERTa (Liu *et al.*, 2020). Le modèle proposé par défaut dans la librairie est celui pré-entraîné sur le corpus français OSCAR (pour *Open Super-large Crawled ALMAnaCH coRpus*) (Ortiz Suárez *et al.*, 2020). La mise en place d'un classifieur DIET utilisant $\mathrm{CamemBERT}_{\mathrm{base,wiki-4GB}}$ avait été réalisée pour comparer les performances de DIET FrALBERT à celle d'un modèle de plus grande taille et entraîné sur le même corpus. Malheureusement, le format **.h5** attendu par la librairie RASA étant indisponible pour cette version de CamemBERT, les résultats du modèle DIET $\mathrm{CamemBERT}_{\mathrm{base,wiki-4GB}}$ obtenus étaient aberrants. Plus précisement, les résultats étaient inférieurs de plus d'une vingtaine de points à ceux obtenus avec FrALBERT sur les différentes métriques suivies et dans les mêmes conditions d'entraînement.

Le Tableau 1 résume les principales caractéristiques de taille des modèles FrALBERT$_{base}$ et CamemBERT$_{base}$. Rappelons que le choix de l'utilisation de FrALBERT est motivé par des raisons écologiques et énergétiques. Son coût de calcul (en temps, énergie et CO_2 produit) étant moins élevé que celui de CamemBERT de par sa taille réduite et ses méthodes d'optimisation (Cattan *et al.*, 2022).

Modèle	Nombre de paramètres	Taille du modèle
FrALBERT$_{\mathrm{base}}$	12 millions	50 MB
CamemBERT$_{\mathrm{base}}$	110 millions	445 MB

TABLE 1 – Caractéristiques techniques relatives à la taille des modèles utilisés. Les tailles sont exprimées en mégabytes (MB).

Concernant l'usage des données, l'outil RASA ne nécessite qu'un lot d'entraînement, qu'il sépare en lot d'entraînement et de validation, pour entraîner le modèle. Deux lots, d'entraînement et d'évaluation, avaient donc été constitués depuis les données issues de la plateforme UBIB. La librairie RASA étant avant tout destinée à des fins de production, des « gardes-fou » sont présents et limitent son utilisation. À titre d'exemple, il est impossible de réaliser de l'apprentissage par transfert depuis un modèle préalablement entraîné avec RASA. Ces limites nous ont poussées à chercher un nouveau système permettant toujours cette détection jointe de l'intention et des concepts dans un contexte plus expérimental. Le modèle DIET FrALBERT servira de *baseline* à nos expériences.

3.2 Mise en place d'un modèle à architecture jointe BERT

L'architecture jointe BERT pour la classification d'intention et la détection des concepts (désigné BERT joint) (Chen *et al.*, 2019) est une version modifiée du modèle BERT (Devlin *et al.*, 2019).

Le modèle BERT est un encodeur de type *Transformer* (Vaswani *et al.*, 2017), multicouche et bidirectionnel. Il existe plusieurs variations du modèle BERT selon le corpus de pré-entraînement, le nombre de couches de blocs d'encodeur de modèle *Transformer*, les dimensions des sorties entre ces couches ou encore le nombre de modules d'attention. Les données utilisées pour son pré-entraînement sont d'abord représentées sous forme de plongements lexicaux fondés sur WordPiece (Wu *et al.*,

2016), après ajout d'un token spécial de classification **[CLS]** en début de phrase. Si plusieurs phrases sont assemblées en paire dans une même donnée, des tokens de séparation **[SEP]** sont insérés. Des plongements positionnels, représentant la position des mots dans une phrase, et segmentaires, représentant la position d'une phrase dans une paire de phrases, sont aussi utilisés. Ces différentes représentations sont ensuite concaténées avant d'être présentée au premier bloc d'encodeur de modèle *Transformer*. BERT est pré-entraîné sur deux tâches : Le masquage de mots (*Masked Language Model*) et la prédiction de la prochaine phrase (*Next Sentence Prediction*). Il peut ensuite être affiné par *fine-tuning* sur une variété d'autres tâches dont la détection d'intention ou la détection des concepts.

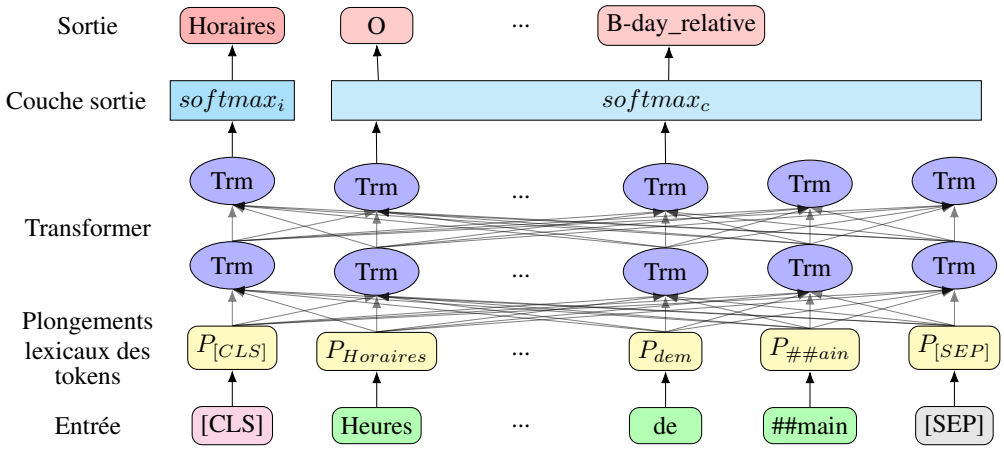


FIGURE 4 – Schéma simplifié de l'architecture du modèle BERT joint d'après Chen *et al.* (2019). La requête en entrée est "Heures d'ouverture de la bibliothèque demain".

Le modèle BERT joint (Chen *et al.*, 2019) propose d'utiliser les états cachés finaux du modèle BERT pour prédire l'intention et les concepts de la phrase passée en entrée. Son architecture est représentée en Figure 4. Dans ce modèle BERT joint, l'état caché final du token **[CLS]** permet de déduire l'intention à l'aide d'une couche de sortie $\mathrm{softmax_i}$. Sa sortie est désignée par la prédiction y^i. Pour les concepts, c'est l'état caché final du premier sous-token qui correspond au début de chacun des n mots de la phrase qui est présenté à une couche de sortie $\mathrm{softmax_c}$. Sa sortie est désignée par la prédiction y^c. L'objectif joint de ce modèle consiste alors au produit des probabilités conditionnelles des concepts et de l'intention :

$$p(y^i, y^c|x) = p(y^i|x) \prod_{n=1}^{N} p(y_n^c|x). \quad (1)$$

Le but est ensuite de maximiser cette probabilité conditionnelle par minimisation d'une fonction de coût par entropie croisée. L'ajout d'une couche à champ aléatoire conditionnel (Lafferty *et al.*, 2001), remplaçant la couche de sortie *softmax*, pour la détection des concepts a été testé par les auteurs Chen *et al.* (2019) mais n'apporte pas toujours une amélioration. L'hypothèse avancée étant que le système d'attention des blocs *Transformer* suffit pour modéliser la structure des différents concepts.

L'architecture du modèle BERT joint peut facilement être ré-implémentée afin de remplacer BERT par $\mathrm{FrALBERT_{base,wiki-4GB}}$ et $\mathrm{CamemBERT_{base,wiki-4GB}}$. Les modèles correspondants seront respectivement désignés par FrALBERT joint et CamemBERT joint.

Précisons que dans l'article sur le classifieur DIET (Bunk *et al.*, 2020), une comparaison avec BERT joint avait été réalisé par les auteurs sur les sets de données ATIS (Hemphill *et al.*, 1990)

et SNIPS (Coucke *et al.*, 2018). Les performances du classifieur DIET étant en deçà de celles du BERT joint de 1 à 2 points pour l'exactitude des intentions et la F-mesure des concepts en utilisant uniquement des représentations éparses. Les auteurs précisaient que ces performances pouvaient être liées à leur méthode d'étiquetage utilisée pour les paramètres, BILOU (Ramshaw & Marcus, 1995), plus complexe que la notation BIO classiquement utilisée. Il est donc probable que FrALBERT joint obtienne de meilleures performances que DIET FrALBERT. Nous utiliserons pour nos modèles à architecture BERT joint la méthode d'étiquetage BIO utilisée dans l'article de Chen *et al.* (2019).

4 Protocoles expérimentaux et résultats

Le but de notre travail est de remplacer le classifieur DIET avec modèle $\mathrm{FrALBERT}_{\mathrm{base,wiki-4GB}}$ préalablement utilisé. Pour cela, nous comparerons ses performances à celles d'un modèle FrALBERT joint. Par ailleurs, les performances de notre modèle compact FrALBERT joint seront comparées à celles d'un modèle CamemBERT joint de plus grande taille. Ceci permettra d'observer les conséquences de la taille du modèle *Transformer* utilisé dans une architecture BERT joint.

4.1 Jeux de données

Comme cela avait été exposé en section 2, nous disposons de deux jeux de données pour lesquels des étiquettes d'intentions et de concepts avaient été décidés préalablement à la création du prototype. Ces deux jeux de données ont été corrigés à l'issue de la campagne d'annotation par un unique annotateur.

— Le corpus UBIB. Il s'agit de questions d'utilisateurs de certaines bibliothèques universitaires métropolitaines françaises destinées à des bibliothécaires, annotées manuellement. Les questions portent généralement sur des informations pratiques, des recherches documentaires ou des problèmes d'accès. Elles ont préalablement été augmentées par génération de données selon la méthode de remplissage de patrons.
— Le corpus DIBISO. Il s'agit des questions obtenues suite à la campagne d'annotation, annotées par le prototype DIET FrALBERT avant correction manuelle. Elles ont été posées par des annotateurs sélectionnés par la DIBISO sans consignes sur les mots clés à utiliser. Ce sont des questions sur des renseignements généraux à propos des BIBLIOTHÈQUES UNIVERSITAIRES PARIS-SACLAY, la plupart ciblées sur les intentions prises en compte par le SQR.

Corpus	DIBISO	$\mathrm{UBIB}_{\mathrm{train}}$	$\mathrm{UBIB}_{\mathrm{test}}$
Nombre de questions	471	12402	997
Taille du vocabulaire	3071	15710	15601
Nombre d'intentions	7	7	7
Nombre de concepts	4	9	9

TABLE 2 – Caractéristiques des différents corpus utilisés. Les lots d'entraînement (*train*) et d'évaluation (*test*) du corpus UBIB sont présentés séparément.

Les caractéristiques de ces jeux de données sont résumées dans la Tableau 2.

Les corpus ont été corrigés avec les 7 intentions utilisées lors de l'élaboration du prototype, présentées dans le Tableau 3. La majorité de ces intentions concernent des catégories de renseignements généraux sur les BIBLIOTHÈQUES UNIVERSITAIRES PARIS-SACLAY : heures d'ouverture, adresse, domaines scientifiques des livres d'une bibliothèque ou bibliothèques comportant des livres d'un domaine scientifique particulier. D'autres sont plus spécifiques au système de SQR : salutations, mise en

question de l'identité du SQR. Une dernière catégorie d'intention comprend toutes les questions n'appartenant pas aux précédentes. Le Tableau 3 révèle aussi la présence d'un déséquilibre entre les proportions des différentes intentions dans le corpus DIBISO comme dans le corpus UBIB pour ses lots d'entraînement (*train*) et d'évaluation (*test*). Concernant les concepts, ils sont étiquetés avec 9 catégories différentes. Ces étiquettes correspondent aux noms des bibliothèques, aux domaines scientifiques, à une date (jour, semaine, mois, période de l'année, etc.) ou des termes relatifs à une date. Les différentes étiquettes utilisées et leurs nombres sont présentés en Tableau 4. Comme les intentions, un fort déséquilibre de répartition des concepts est présent entre les différents corpus.

Intention	DIBISO	$\text{UBIB}_{\text{train}}$	$\text{UBIB}_{\text{test}}$
bot challenge	26	62	9
get timetable of library	7	2812	162
out of scope	346	1433	356
search library fields	14	4925	141
greet	26	11	3
search library address	29	2637	256
search libraries from field	23	522	70

TABLE 3 – Répartition des intentions dans les différents corpus. Les lots d'entraînement (*train*) et d'évaluation (*test*) du corpus UBIB sont présentés séparément.

Concepts	DIBISO	$\text{UBIB}_{\text{train}}$	$\text{UBIB}_{\text{test}}$
library	136	11920	609
field	66	8353	515
month	0	67	30
period_relative	1	545	76
day_of_month	0	1079	9
day_relative	3	287	54
week_relative	0	61	9
day_of_the_week	0	576	43
date_relative	0	43	13

TABLE 4 – Répartition des intentions dans les différents corpus. Les lots d'entraînement (*train*) et d'évaluation (*test*) du corpus UBIB sont présentés séparément.

4.2 Métriques d'évaluation

Nous avons d'abord évalué chacune des tâches de manière indépendante. Pour la détection d'intention, seule l'exactitude sera utilisée. Pour la détection des concepts, la précision, le rappel et la F-mesure seront utilisés. Une autre métrique permet de calculer le résultat conjoint des deux tâches : L'exactitude du cadre sémantique à l'échelle des phrases. Le cadre sémantique d'une phrase est considéré comme exact lorsque son intention et ses concepts ont tous été parfaitement prédits.

4.3 Protocole d'entraînement

Deux séries d'expériences sont réalisées sur les modèles FrALBERT joint et CamemBERT joint afin d'étudier la qualité du corpus DIBISO. La question sous-jacente étant : est-ce que ces données DIBISO vont être suffisantes pour obtenir un modèle avec de bonnes capacités de prédiction ? Pour cela, des modèles joints seront entraînés et testés exclusivement sur le corpus DIBISO. D'autres modèles joints

seront entraînés sur le lot d'entraînement du corpus UBIB associé à une partie du corpus DIBISO. Dans ce second cas, le lot d'évaluation du corpus UBIB servira alors de lot de validation. Les performances des modèles seront exclusivement évaluées sur une sous-partie des données DIBISO.

En raison de la petite quantité de données du corpus DIBISO, une validation croisée à *k*-blocs (Kohavi, 1995) est réalisée : Le corpus est mélangé, découpé en cinq blocs de tailles proches, puis quatre de ces blocs sont assemblés pour former le lot d'entraînement tandis que le dernier bloc constitue celui d'évaluation. Cinq différentes répartitions de lots d'entraînements et d'évaluations sont ainsi obtenues. La moyenne et l'écart-type sur ces différentes répartitions seront calculés.

Des modèles FrALBERT joint et CamemBERT joint sont mis en place et entraînés sur chacune des différentes répartitions. Concernant le classifieur DIET FrALBERT, celui-ci est entraîné pendant 300 itérations totales du corpus d'entraînement (*epoch*) UBIB ré-annoté, pour reproduire des conditions d'entraînement similaires à celles du prototype déployé lors de la campagne d'annotation. Si ce protocole expérimental ne permet pas une véritable comparaison entre classifieur DIET et modèle joint, le but est avant tout de s'assurer que le nouveau modèle mis en place soit plus performant que la *baseline* représentée par le précèdent.

Concernant les optimisations, l'optimiseur AdamW (Loshchilov & Hutter, 2019) est utilisé pour les modèles joints. Des optimisations intégrées à la librairie RASA, difficiles à identifier en raison d'un effet « boîte noire », sont appliquées au classifieur DIET. Avec RASA, les résultats d'un entraînement à un autre sont peu variables. Un seul *run* sera donc réalisé pour le classifieur DIET.

Modèle joint	**Exactitude**$_i$	**Précision**$_c$	**Rappel**$_c$	**F-mesure**$_c$	**Exactitude**$_{cs}$
FrALBERT$_{base,wiki-4GB}$	92,49±0,91	75,87±2,12	78,27±1,97	77,05±1,98	49,11±3,36
CamemBERT$_{base,wiki-4GB}$	**94,67**±0,78	**81,51**±0,62	**85,19**±0,27	**83,31**±0,39	**60,52**±0,49

TABLE 5 – Performances des modèles joints après entraînement de 10 *epoch* sur le corpus ATIS-FR pour les intentions (i), les concepts (c) et le cadre sémantique au niveau des phrases (cs). Les résultats présentés sont les moyennes et écarts-types sur 10 *runs* sur le lot d'évaluation d'ATIS-FR.

Puisque le corpus DIBISO est de faible taille et afin de spécialiser le modèle *Transformer* utilisé sur le système de détection jointe, un pré-entraînement de 10 *epoch* est réalisé sur le corpus français ATIS-FR issu du corpus MultiATIS+++ (Xu *et al.*, 2020) pour les modèles joints. Ce pré-entraînement n'est pas possible pour le classifieur DIET inclus dans la librairie RASA car la librairie ne permet pas l'entraînement continu et l'apprentissage par transfert. À l'issue de ce pré-entraînement simple, dont les résultats sont présentés en Tableau 5, on constate une différence non-négligeable entre les deux modèles joints. Ces différences sont en faveur du modèle CamemBERT joint avec des écarts de 2,18 points pour l'exactitude des intentions, et de 6,26 points pour la F-mesure des concepts. Les auteurs Cattan *et al.* (2022) ont pourtant démontré que les modèles monolingues FrALBERT$_{base,wiki-4GB}$ et CamemBERT$_{base,wiki-4GB}$ ont des performances similaires sur ce corpus ATIS-FR. Dans leurs travaux, les résultats pour la F-mesure des concepts étaient de 92,8 pour FrALBERT$_{base,wiki-4GB}$ et de 92,5 pour CamemBERT$_{base,wiki-4GB}$. L'écart que nous observons dans le Tableau 5 est plus important que celui de 0,3 point qu'ils avaient constaté.

Plusieurs hypothèses peuvent expliquer ces différences, notamment le fait que l'objectif à minimiser dans un modèle joint diffère de celui d'un problème de classification en classes multiples plus classique. Par ailleurs, nous utilisons des hyperparamètres fixes lors de ce pré-entraînement alors qu'une optimisation par *population based training* (Jaderberg *et al.*, 2017) était utilisée dans les

travaux de Cattan *et al.* (2022). Nous choisissons donc d'utiliser le *population based training*, illustré en Figure 5, dans nos expérimentations sur les modèles joints. Nous explorerons un taux d'apprentissage entre 1 et 5, une taille de *batch* d'entraînement entre 8 et 32 et un nombre d'*epoch* d'apprentissage entre 8 et 14. Le nombre de versions parallèles entraînées (épreuves) sera fixé à 8 et la meilleure épreuve pour chaque *run* sera sélectionnée en fonction du plus petit résultat de coût obtenu sur le lot de validation. Le *population based training* ne sera pas utilisé lors du pré-entraînement, notre but n'étant pas de spécialiser nos modèles sur le corpus ATIS-FR. Dans le cas où seul le corpus DIBISO est utilisé, un nouveau découpage de 10% du lot d'entraînement est réalisé pour constituer un lot de validation nécessaire au *population based training*. L'utilisation du même lot de validation que celui des modèles entraînés en partie sur le corpus UBIB n'est pas envisageable : ce lot est de trop grande taille et contient des concepts que ne présente pas le corpus DIBISO. Ce nouveau découpage, ainsi que les différences de distribution des lots de validation (en quantité et origine des données), peuvent constituer des biais statistiques à prendre en compte lors de l'analyse de nos résultats.

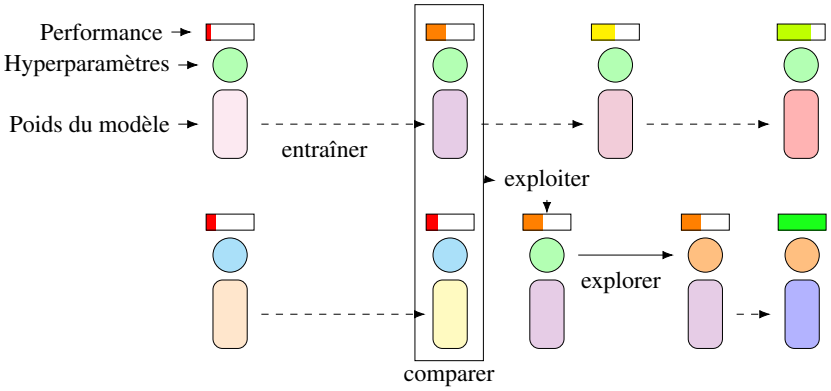


FIGURE 5 – Représentation schématique du *population based training* d'après Jaderberg *et al.* (2017). Plusieurs versions (épreuves) d'un même modèle sont initialisées aléatoirement au niveau de leurs hyperparamètres et de leurs poids. Elle sont entraînées et évaluées parallèlement. Les différentes configurations de poids et d'hyperparamètres sont représentées ici par des couleurs différentes. À intervalles réguliers et après comparaison statistique des différentes versions, les poids et hyperparamètres d'une version peu performante sont remplacés par ceux d'une meilleure version. De nouveaux hyperparamètres sont alors explorés sur cette copie en appliquant un facteur de perturbation ou en les réinitialisant de manière aléatoire.

Les résultats des modèles joints pré-entraînés sur ATIS-FR seront confrontés à ceux du classifieur DIET FrALBERT entraîné sur le corpus UBIB uniquement.

4.4 Résultats et analyse

Les performances de nos modèles sont présentées en Tableau 6 pour les différentes métriques.
Une analyse des erreurs des meilleurs modèles FrALBERT joint et CamemBERT joint est aussi réalisée : Pour chaque expérience, les modèles avec les plus petites valeurs de coûts sur leurs lots d'évaluation sont sélectionnés parmi les différents *runs*. Pour permettre cette comparaison, ces modèles sont sélectionnés sur la même répartition de la validation croisée à k-blocs. Ensuite, des prédictions sur l'intégralité du corpus DIBISO sont réalisées. Les résultats de cette analyse sont représentés par des matrices de confusion en Figures 6 et 7. Des tests de McNemar sont réalisés avec une valeur seuil de 0,05 entre ces meilleurs modèles pour déterminer si leurs différences sont significatives sur les tâches de détection d'intention et d'identification des concepts.

Modèle	Données	Exactitude
DIET FrALBERT$_{\text{base,wiki-4GB}}$	U	72,59$\pm$3, 45
FrALBERT$_{\text{base,wiki-4GB}}$ joint	D	87,15$\pm$5, 43
CamemBERT$_{\text{base,wiki-4GB}}$ joint	D	86,62$\pm$4, 80
FrALBERT$_{\text{base,wiki-4GB}}$ joint	U + D	**87,55**$\pm$4, 76
CamemBERT$_{\text{base,wiki-4GB}}$ joint	U + D	87,52$\pm$4, 97

(a) Performances pour les intentions

Modèle	Données	Précision	Rappel	F-mesure
DIET FrALBERT$_{\text{base,wiki-4GB}}$	U	85,29$\pm$7, 49	88,79$\pm$2, 08	**92,87**$\pm$4, 21
FrALBERT$_{\text{base,wiki-4GB}}$ joint	D	81,50$\pm$10, 60	75,71$\pm$11, 65	78,14$\pm$9, 92
CamemBERT$_{\text{base,wiki-4GB}}$ joint	D	81,33$\pm$7, 52	81,30$\pm$8, 15	81,00$\pm$6, 23
FrALBERT$_{\text{base,wiki-4GB}}$ joint	U + D	**90,97**$\pm$4, 46	**94,04**$\pm$4, 03	92,38$\pm$3, 03
CamemBERT$_{\text{base,wiki-4GB}}$ joint	U + D	85,00$\pm$6, 13	90,95$\pm$5, 37	87,73$\pm$4, 63

(b) Performances pour les concepts

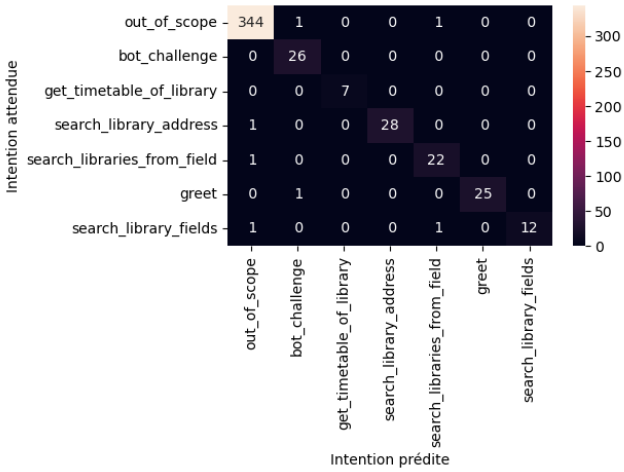
Modèle	Données	Exactitude
DIET FrALBERT$_{\text{base,wiki-4GB}}$	U	67,44$\pm$4, 99
FrALBERT$_{\text{base,wiki-4GB}}$ joint	D	78,38$\pm$6, 09
CamemBERT$_{\text{base,wiki-4GB}}$ joint	D	77,28$\pm$5, 03
FrALBERT$_{\text{base,wiki-4GB}}$ joint	U + D	**83,51**$\pm$5, 31
CamemBERT$_{\text{base,wiki-4GB}}$ joint	U + D	81,42$\pm$5, 45

(c) Performances pour le cadre sémantique au niveau des phrases

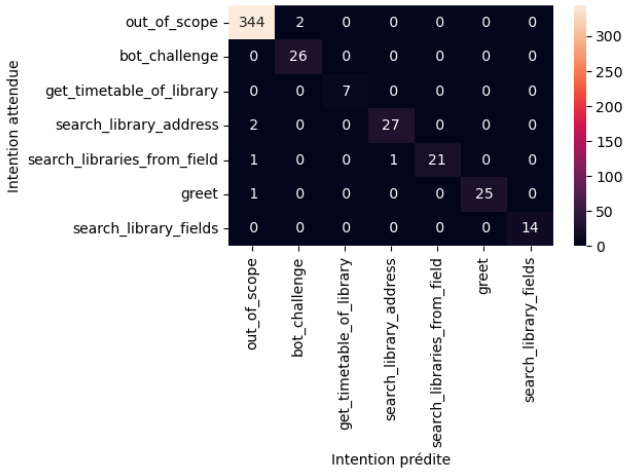
TABLE 6 – Performances des différents modèles. Les modèles sont entraînés sur les données d'entraînement UBIB (U) et/ou DIBISO (D), comme indiqué en colonne "Données", et évalués sur les différents lots d'évaluation issus du corpus DIBISO pour chaque découpage de la validation croisée. Les résultats présentés pour les modèles joints sont les moyennes et écarts-types sur 6 *runs*, après calcul des valeurs sur les 5 blocs de validation croisée pour chaque *run*. Les résultats présentés pour le modèle DIET sont issus d'un unique *run* pour lequel la moyenne et l'écart-type sur les 5 blocs d'évaluation de la validation croisée des modèles joints ont été calculés.

4.4.1 Comparaison des modèles à architecture jointe sur la tâche de détection d'intention

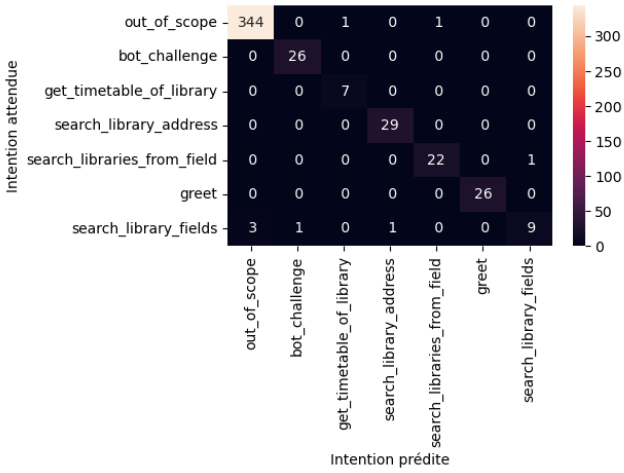
Pour la tâche de détection d'intention dont les résultats sont présentés en Tableau 6a, les modèles FrALBERT joint et CamemBERT joint entraînés uniquement sur le corpus DIBISO ont obtenu respectivement 87,15 et 86,62 d'exactitude, soit seulement 0,53 point d'écart en faveur de FrALBERT joint. Les modèles joints entraînés sur les deux corpus ont des performances très similaires, avec 87,55 pour FrALBERT joint et 87,52 pour CamemBERT joint, soit 0,03 point d'écart. Ces faibles écarts de moyennes entre modèles entraînés sur les mêmes ensembles de données, associés à de forts écarts-types (entre 4,76 et 5,43), montrent que dans nos conditions d'expérimentations FrALBERT joint parvient à égaler CamemBERT joint pour la tâche de détection d'intention, malgré son plus faible nombre de paramètres. Par ailleurs, les performances similaires entre modèles entraînés uniquement sur le corpus DIBISO ou associé au corpus UBIB indiquent que malgré sa faible quantité de données, le contenu du corpus DIBISO est suffisamment complet pour permettre aux modèles de distinguer les différentes intentions.



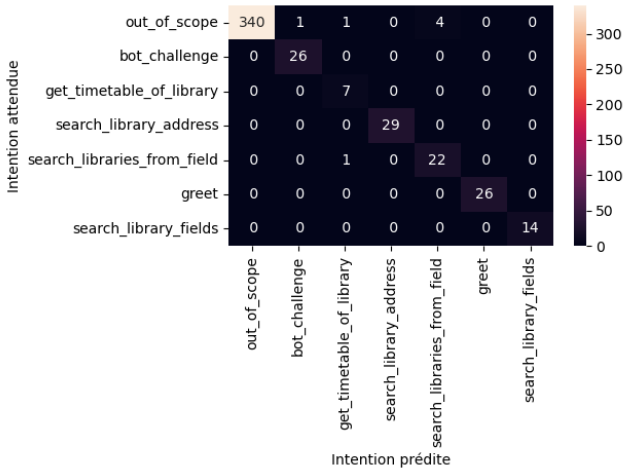
(a) FrALBERT joint
entraîné sur données DIBISO



(b) CamemBERT joint
entraîné sur données DIBISO



(c) FrALBERT joint
entraîné sur données DIBISO et UBIBs



(d) CamemBERT joint
entraîné sur données DIBISO et UBIBs

FIGURE 6 – Matrices de confusion des intentions sur le corpus DIBISO.

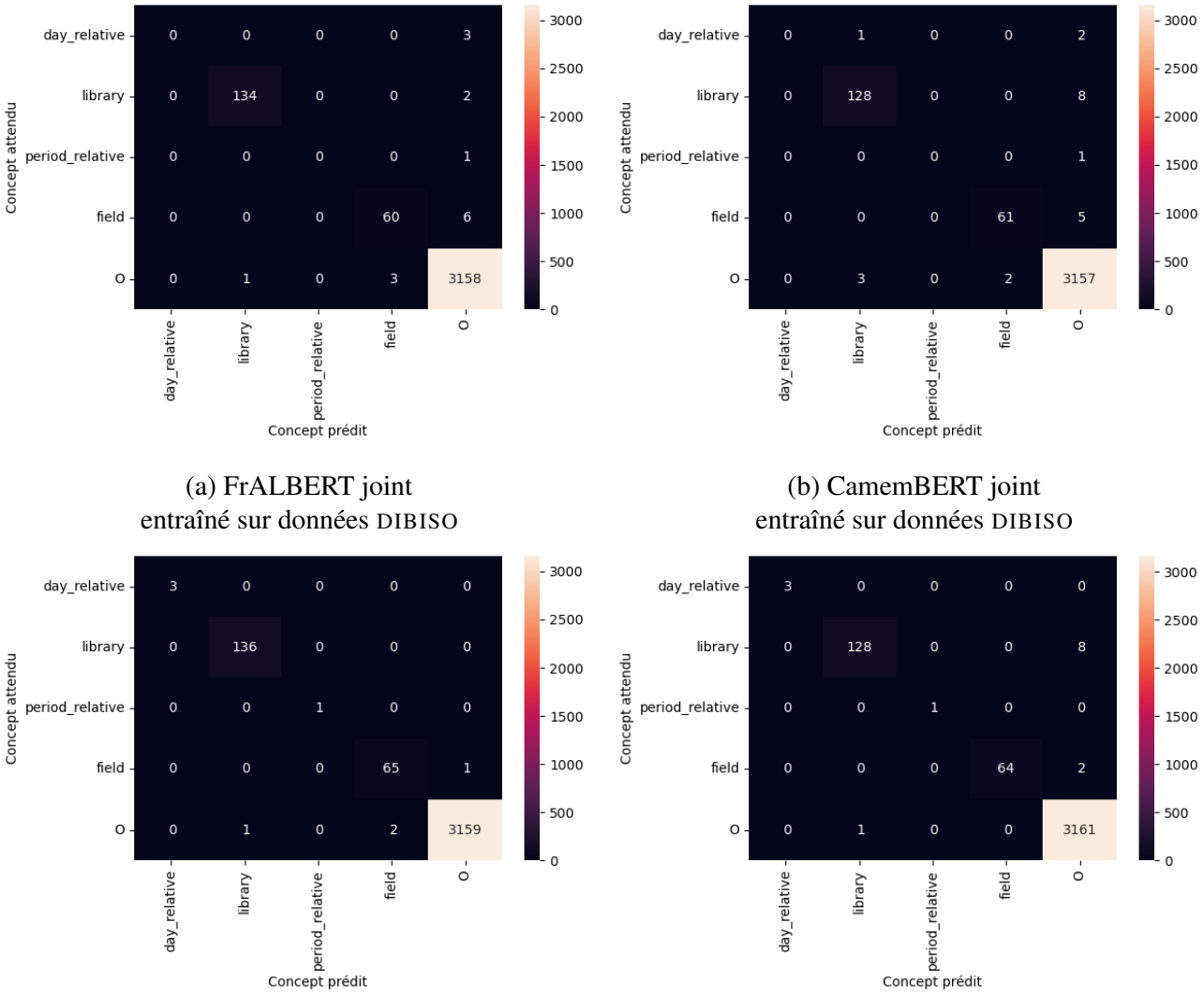
Une analyse des erreurs des prédictions de ces intentions pour les modèles joints, représentée par la Figure 6, confirme cette bonne correspondance des intentions attendues et prédites. De plus, les valeurs-p ne permettent pas d'établir de différences significatives pour la tâche de détection d'intention entre les meilleurs modèles utilisant le même *Transformer* (0,7600 pour les modèles FrALBERT, 1 pour les modèles CamemBERT) ou entre les meilleurs modèles entraînés sur les mêmes corpus (1 pour les modèles entraînés sur le corpus DIBISO uniquement, 0.7600 pour ceux entraînés sur les deux corpus). Pour près de la moitié des erreurs commises par les modèles entraînés uniquement sur le corpus DIBISO, c'est l'intention majoritaire (*out_of_scope*) du corpus qui a été prédite. Dans le cas des modèles entraînés sur les deux corpus, le meilleur modèle FrALBERT joint rencontre plus de difficulté à identifier correctement l'intention *search_library_fields* avec un tiers d'erreurs pour les phrases y correspondant. Il s'agit pourtant de l'intention la plus présente du corpus UBIB utilisé pour l'entraînement. Le meilleur modèle CamemBERT joint identifie certaines questions comme étant étiquetées *search_libraries_from_field* parmi les *out_of_scope* (4 des 7 erreurs commises par ce modèle). En regardant les questions concernées par ces erreurs, certaines formulations peuvent

prêter à confusion entre les deux intentions. Par exemple, la question *"je cherche des livres de droit"* est considérée comme *out_of_scope* car il s'agit d'une recherche de livre, mais est identifiée comme *search_libraries_from_field* par le modèle. La question équivalente *search_libraries_from_field* serait plutôt *"dans quelle bibliothèque trouver des livres de droit"* ou *"où trouver des livres de droit"*.

4.4.2 Comparaison des modèles à architecture jointe sur la tâche d'identification des concepts

Concernant l'identification des concepts en Tableau 6b, on observe un écart important entre les modèles selon les données d'entraînement utilisées : Le modèle FrALBERT joint gagne 14,24 points de F-mesure s'il est entraîné sur les deux corpus (92,38) plutôt que sur le corpus DIBISO seul (78,14). L'écart entre les modèles CamemBERT est de 6,73 points pour la F-mesure avec 81,00 s'il est entraîné sur le corpus DIBISO uniquement contre 87,73 s'il est entraîné sur les deux corpus. Les différences sont significatives pour cette tâche entre les meilleurs modèles FrALBERT entraînés sur corpus DIBISO seul par rapport à celui entraîné sur les deux corpus (valeur-p de 0,0028). Ces différences sont aussi significatives entre les meilleurs modèles CamemBERT selon les données d'entraînement (valeur-p de 0,0201). Ainsi, il semblerait que les modèles joints ne parviennent pas à extraire suffisamment d'informations uniquement du corpus DIBISO pour cette tâche. La faible présence de certains concepts dans ce corpus (*day_relative* et *period_relative*) peut empêcher leur identification correcte et expliquer en partie ces scores des modèles entraînés uniquement sur le corpus DIBISO. Par ailleurs, de par les découpages en lots d'évaluation et de validation du corpus DIBISO, certains de ces modèles peuvent ne pas avoir été entraînés sur des questions présentant ces concepts. Les matrices de confusion correspondant aux meilleurs modèles entraînés uniquement sur le corpus DIBISO, en Figure 7a et 7b, confirment que les étiquettes présentes en fortes quantité dans ce corpus (*library* et *field*) sont plutôt bien identifiées, avec la présence de quelques faux-négatifs et faux-positifs. Les deux autres étiquettes ne sont pas du tout détectées.

Au sujet des différences entre *Transformer* utilisés, elles sont non négligeables au niveau des moyennes. Lors de l'entraînement sur le corpus DIBISO uniquement, le score de rappel est en faveur de CamemBERT joint avec une valeur de 81,20 contre 5,49 points de moins pour FrALBERT joint. L'écart entre les F-mesures est de 2,86 points en faveur de CamemBERT joint et l'écart entre les scores de précision de 0,17 point en faveur de FrALBERT joint (81,50 contre 81,33 pour CamemBERT joint). Cependant, pour ces modèles, FrALBERT présente de plus importants écart-types (entre 9,92 et 11,65) que CamemBERT (entre 6,23 et 8,15). Lors de l'entraînement sur les deux corpus, les résultats sont en faveur de FrALBERT joint uniquement : Son score de précision est de 90,97 soit 5,97 points de plus que CamemBERT joint, son score de rappel avec une valeur de 94,04 est 3,09 points plus élevé que CamemBERT joint et enfin, sa F-mesure est en avance de 4,65 points. Le peu d'erreurs présentes dans la matrice de confusion du meilleur modèle FrALBERT joint entraîné sur les deux corpus, en Figure 7c, corrobore ces résultats. Ces différences de score en faveur de FrALBERT joint sont surprenantes en considérant que ce modèle est de plus petite taille que CamemBERT. Une hypothèse pouvant être formulée est que les conditions de l'optimisation par *population based training* telle que la fourchette du nombre d'*epoch* explorée, ou le petit nombre de *runs* réalisé ont grandement influencés les résultats obtenus. Si les scores de précision et de rappel de CamemBERT joint entraîné sur les deux corpus montrent une faible tendance à la détection de faux-positifs, les prédictions du meilleur modèle en Figure 7d montrent plutôt quelques faux-négatifs. La valeur-p entre les meilleurs modèles confirment que pour les modèles entraînés sur le corpus DIBISO seul, les différences sont significatives avec 0,027. Par contre, la valeur-p de 0,24 entre les meilleurs modèles FrALBERT et CamemBERT entraînés sur l'ensemble des corpus ne permet pas d'établir de différence significative sur cette tâche avec le test de McNemar.



(a) FrALBERT joint entraîné sur données DIBISO

(b) CamemBERT joint entraîné sur données DIBISO

(c) FrALBERT joint entraîné sur données DIBISO et UBIBs

(d) CamemBERT joint entraîné sur données DIBISO et UBIBs

FIGURE 7 – Matrices de confusion des concepts sur le corpus DIBISO.

4.4.3 Comparaison des modèles à architecture jointe sur le cadre sémantique

Les résultats de l'exactitude du cadre sémantique présentés en Tableau 6c suivent ceux observés pour les concepts : Le meilleur modèle sur cette métrique est FrALBERT joint entraîné sur l'ensemble des deux corpus avec 83,51 d'exactitude, contre 81,42 pour le modèle CamemBERT joint équivalent. Les modèles joints entraînés sur association des deux corpus obtiennent de meilleures performances que ceux entraînés sur le corpus DIBISO seul avec un écart de 5,13 points pour les modèles FrALBERT joint et de 4,14 points pour les modèles CamemBERT joint.

4.4.4 Comparaison du classifieur DIET aux modèles à architecture jointe

DIET FrALBERT présente des résultats inférieurs à l'ensemble des différents modèles joints pour la tâche de détection d'intention avec 72,59, et en conséquence pour l'exactitude du cadre sémantique (67,44). Concernant la tâche d'identification des concepts, ses résultats sont meilleurs que les modèles joints entraînés sur le corpus DIBISO uniquement notamment avec une F-mesure de 92,87. Pour les

modèles entraînés sur plus de données, FrALBERT joint obtient globalement de meilleurs résultats que DIET FrALBERT. Nos conditions d'expérimentations ne nous permettent pas d'estimer si ces différences de performances entre classifieur DIET et modèles joints sont liées aux architectures, aux optimisations, au pré-entraînement des modèles joints sur ATIS-FR ou à l'influence des données du corpus DIBISO.

4.4.5 Réflexion sur les écarts-types

On observe sur l'ensemble des résultats la présence d'un fort écart-type pour chacune des métriques. Il s'explique avant tout par l'utilisation d'une validation croisée engendrant des lots potentiellement très déséquilibrés. Une autre explication, plus difficile à prouver, est l'effet en « dent de scie » entre les deux tâches qui peut être observé en fin d'entraînement chez les modèles réalisant une détection jointe (Hui *et al.*, 2021) : l'augmentation de l'exactitude d'une tâche peut déclencher la diminution de l'exactitude de l'autre tâche. Ainsi, pour des calculs de coûts équivalents, deux modèles peuvent avoir des performances relativement différentes pour les tâches de détection d'intention et d'identification des concepts.

5 Conclusion

Dans cet article, nous avons détaillé la mise en place d'un modèle réalisant une détection jointe de l'intention et des concepts, dont l'architecture repose sur un modèle *Transformer* compact. Notre contexte applicatif, un système interactif de questions-réponses (SQR), ne disposait initialement d'aucun corpus répondant précisément à notre besoin. Un premier prototype utilisant un corpus relativement similaire (corpus UBIB), et reposant sur une architecture de classifieur DIET (Bunk *et al.*, 2020) a permis de collecter un corpus de 471 questions. À partir de ces nouvelles données, des modèles reposant sur une architecture BERT joint (Chen *et al.*, 2019) ont été entraînés. Le modèle compact de *Transformer* FrALBERT$_{base,wiki-4GB}$ a été utilisé pour l'élaboration du classifieur DIET et des modèles joints. Dans le cas des modèles joints, FrALBERT a été comparé avec le *Transformer* CamemBERT$_{base,wiki-4GB}$. Une de nos finalités étant d'évaluer si le modèle compact FrALBERT a des performances comparables à un modèle de plus grande taille pour cette détection jointe et dans notre contexte applicatif.

Nos résultats montrent que les nouvelles données obtenues, bien que correspondant mieux à notre contexte, sont insuffisantes à elles seules pour permettre de meilleures performances que le prototype sur la tâche d'identification des concepts. Elles permettent par contre aux modèles joints d'obtenir une meilleure exactitude pour la tâche de détection d'intention. Un entraînement à la fois sur les données UBIB et les données récoltées permet d'obtenir des modèles joints plus performants que le prototype DIET FrALBERT sur les deux tâches. Bien que notre étude ne permette pas une véritable comparaison entre les architectures classifieur DIET et BERT joint, nous pouvons donc envisager remplacer le classificateur DIET actuel de notre SQR par un modèle joint entraîné selon notre protocole expérimental. Par ailleurs, le modèle joint FrALBERT entraîné sur les deux corpus présente des performances similaires à son équivalent CamemBERT joint. L'usage de FrALBERT dans notre SQR permettra donc l'obtention de bons résultats en réduisant notre consommation des ressources. Enfin, nous prévoyons d'enrichir notre corpus DIBISO, notamment pour lui ajouter les concepts qui lui manquent. Pour cela, l'utilisation de méthodes de génération ou d'augmentation de données, comme les méthodes de patron (Boulanger *et al.*, 2022) ou *Tri-training* (Boulanger, 2020), seront envisagées.

Remerciements

Nous tenons à remercier la Direction des Bibliothèques, de l'Information et de la Science Ouverte de l'Université Paris-Saclay pour la proposition de ce projet, ainsi que leur collaboration et leur aide lors de sa mise en place. Nous adressons nos sincères remerciements à la plateforme Ubib et son intermédiaire, Natacha LECLERC, pour nous avoir fourni des conversations issues de leur service. Nous remercions Mathilde VERON pour son travail et son rapport sur l'étude de faisabilité du projet. Tous nos remerciements vont également aux différents acteurs du projet Humane.AI pour leur accueil et échanges stimulants lors de nos réunions hebdomadaires. Enfin, nous tenons à exprimer toute notre gratitude à notre équipe encadrante du LISN pour leur bienveillance, leurs conseils avisés et l'opportunité de rédiger cet article : Sophie ROSSET, Christophe SERVAN, Laure SOULIER et Sahar GHANNAY.

Références

BENDER E. M., GEBRU T., MCMILLAN-MAJOR A. & SHMITCHELL S. (2021). On the dangers of stochastic parrots : Can language models be too big ? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, p. 610–623, New York, NY, USA : Association for Computing Machinery.

BOCKLISCH T., FAULKER J., PAWLOWSKI N. & NICHOL A. (2017). Rasa : Open source language understanding and dialogue management. In *NIPS 2017 Conversational AI workshop*, p. 1–9.

BOULANGER H. (2020). Évaluation systématique d'une méthode commune de génération. In *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 3 : Rencontre des Étudiants Chercheurs en Informatique pour le TAL*, p. 43–56, Nancy, France : ATALA.

BOULANGER H., LAVERGNE T. & ROSSET S. (2022). Generating unlabelled data for a tri-training approach in a low resourced NER task. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, p. 30–37, Hybrid : Association for Computational Linguistics.

BUNK T., VARSHNEYA D., VLASOV V. & NICHOL A. (2020). Diet : Lightweight language understanding for dialogue systems. *arXiv : 2004.09936 [cs]*.

CASTELLUCCI G., BELLOMARIA V., FAVALLI A. & ROMAGNOLI R. (2019). Multi-lingual intent detection and slot filling in a joint bert-based model. *arXiv : 1907.02884 [cs]*.

CATTAN O., GHANNAY S., SERVAN C. & ROSSET S. (2022). Benchmarking Transformers-based models on French Spoken Language Understanding tasks. In *Proc. Interspeech 2022*, p. 1238–1242.

CATTAN O., SERVAN C. & ROSSET S. (2021). On the usability of transformers-based models for a French question-answering task. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, p. 244–255, Held Online : INCOMA Ltd.

CHEN Q., ZHUO Z. & WANG W. (2019). Bert for joint intent classification and slot filling. *arXiv : 1902.10909 [cs]*.

CHEN Y.-N., HAKANNI-TÜR D., TUR G., CELIKYILMAZ A., GUO J. & DENG L. (2016). Syntax or semantics ? knowledge-guided joint semantic frame parsing. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, p. 348–355.

COUCKE A., SAADE A., BALL A., BLUCHE T., CAULIER A., LEROY D., DOUMOURO C., GISSELBRECHT T., CALTAGIRONE F., LAVRIL T., PRIMET M. & DUREAU J. (2018). Snips voice platform : an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv : 1805.10190 [cs]*.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics.

GOO C.-W., GAO G., HSU Y.-K., HUO C.-L., CHEN T.-C., HSU K.-W. & CHEN Y.-N. (2018). Slot-gated modeling for joint slot filling and intent prediction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 2 (Short Papers)*, p. 753–757, New Orleans, Louisiana : Association for Computational Linguistics.

GUO D., TUR G., YIH W.-T. & ZWEIG G. (2014). Joint semantic utterance classification and slot filling with recursive neural networks. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, p. 554–559.

HAKKANI-TÜR D., TUR G., CELIKYILMAZ A., CHEN Y.-N., GAO J., DENG L. & WANG Y.-Y. (2016). Multi-Domain Joint Semantic Frame Parsing Using Bi-Directional RNN-LSTM. In *Proc. Interspeech 2016*, p. 715–719.

HEMPHILL C. T., GODFREY J. J. & DODDINGTON G. R. (1990). The ATIS spoken language systems pilot corpus. In *Speech and Natural Language : Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27,1990*.

HENDERSON M., VULIĆ I., GERZ D., CASANUEVA I., BUDZIANOWSKI P., COOPE S., SPITHOURAKIS G., WEN T.-H., MRKŠIĆ N. & SU P.-H. (2019). Training neural response selection for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 5392–5404, Florence, Italy : Association for Computational Linguistics.

HUI Y., WANG J., CHENG N., YU F., WU T. & XIAO J. (2021). Joint intent detection and slot filling based on continual learning model. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 7643–7647.

JADERBERG M., DALIBARD V., OSINDERO S., CZARNECKI W. M., DONAHUE J., RAZAVI A., VINYALS O., GREEN T., DUNNING I., SIMONYAN K., FERNANDO C. & KAVUKCUOGLU K. (2017). Population based training of neural networks. *arXiv : 1711.09846 [cs]*.

JAPKOWICZ N. & STEPHEN S. (2002). The class imbalance problem : A systematic study. *Intell. Data Anal.*, **6**(5), 429–449.

JEONG M. & LEE G. G. (2008). Triangular-chain conditional random fields. *IEEE Transactions on Audio, Speech, and Language Processing*, **16**(7), 1287–1302.

KOHAVI R. (1995). A study of cross-validation and Bootstrap for accuracy estimation and model selection. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, p. 1137–1143 : Morgan Kaufmann Publishers Inc.

LAFFERTY J. D., MCCALLUM A. & PEREIRA F. C. N. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, p. 282–289, San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

LAMPLE G., BALLESTEROS M., SUBRAMANIAN S., KAWAKAMI K. & DYER C. (2016). Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 260–270, San Diego, California : Association for Computational Linguistics.

LAN Z., CHEN M., GOODMAN S., GIMPEL K., SHARMA P. & SORICUT R. (2020). Albert : A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

LIU B. & LANE I. (2016). Attention-Based Recurrent Neural Network Models for Joint Intent Detection and Slot Filling. In *Proc. Interspeech 2016*, p. 685–689.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2020). Roberta : A robustly optimized bert pretraining approach.

LOSHCHILOV I. & HUTTER F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations*.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., DE LA CLERGERIE É., SEDDAH D. & SAGOT B. (2020). CamemBERT : a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7203–7219, Online : Association for Computational Linguistics.

MOOSAVI N. S., FAN A., SHWARTZ V., GLAVAŠ G., JOTY S., WANG A. & WOLF T., Éds. (2020). *Proceedings of SustaiNLP : Workshop on Simple and Efficient Natural Language Processing*, Online. Association for Computational Linguistics.

ORTIZ SUÁREZ P. J., ROMARY L. & SAGOT B. (2020). A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 1703–1714, Online : Association for Computational Linguistics.

RAMSHAW L. & MARCUS M. (1995). Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

SHAW P., USZKOREIT J. & VASWANI A. (2018). Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 2 (Short Papers)*, p. 464–468, New Orleans, Louisiana : Association for Computational Linguistics.

SMITH S. L., KINDERMANS P.-J. & LE Q. V. (2018). Don't decay the learning rate, increase the batch size. In *International Conference on Learning Representations*.

STRUBELL E., GANESH A. & MCCALLUM A. (2019). Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 3645–3650, Florence, Italy : Association for Computational Linguistics.

VALSOV V., MOSIG J. E. M. & NICHOL A. (2019). Dialogue transformers. *arXiv : 1910.00486 [cs]*.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. U. & POLOSUKHIN I. (2017). Attention is all you need. In I. GUYON, U. V. LUXBURG, S. BENGIO, H. WALLACH, R. FERGUS, S. VISHWANATHAN & R. GARNETT, Éds., *Advances in Neural Information Processing Systems*, volume 30 : Curran Associates, Inc.

WELD H., HUANG X., LONG S., POON J. & HAN S. C. (2022). A survey of joint intent detection and slot filling models in natural language understanding. *ACM Comput. Surv.*, **55**(8).

WU L., FISCH A., CHOPRA S., ADAMS K., BORDES A. & WESTON J. (2018). Starspace : Embed all the things ! *Proceedings of the AAAI Conference on Artificial Intelligence*, **32**(1).

WU Y., SCHUSTER M., CHEN Z., LE Q. V., NOROUZI M., MACHEREY W., KRIKUN M., CAO Y., GAO Q., MACHEREY K., KLINGNER J., SHAH A., JOHNSON M., LIU X., KAISER L., GOUWS S., KATO Y., KUDO T., KAZAWA H., STEVENS K., KURIAN G., PATIL N., WANG W., YOUNG C., SMITH J., RIESA J., RUDNICK A., VINYALS O., CORRADO G., HUGHES M. & DEAN J. (2016). Google's neural machine translation system : Bridging the gap between human and machine translation. *arXiv : 1609.08144 [cs]*.

XU P. & SARIKAYA R. (2013). Convolutional neural network based triangular crf for joint intent detection and slot filling. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, p. 78–83.

XU W., HAIDER B. & MANSOUR S. (2020). End-to-end slot alignment and recognition for cross-lingual NLU. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 5052–5063, Online : Association for Computational Linguistics.

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Utiliser les syntagmes nominaux complexes anglais pour évaluer la robustesse des systèmes de traduction anglais-français en langue de spécialité

Maud Bénard
Laboratoire CLILLAC-ARP, université Paris Cité, 8 place Paul Ricœur, 75013 Paris, France
maud.benard@u-paris.fr

RÉSUMÉ

Nous défendons l'idée que l'analyse des erreurs faites lors de la traduction des syntagmes nominaux complexes présente un intérêt pour évaluer la robustesse des systèmes de traduction automatique anglais-français en langue de spécialité. Ces constructions syntaxiques impliquent des questions de syntaxe et de lexique qui constituent un obstacle important à leur compréhension et leur production pour les locuteurs d'anglais non natifs. Nous soutenons que ces analyses contribueraient à garantir que les systèmes de TA répondent aux exigences linguistiques des utilisateurs finaux auxquels ils sont destinés.

ABSTRACT

Using English complex noun phrases to evaluate the robustness of specialized English to French machine translation systems

We argue that the analysis of errors made in the translation of complex noun phrases might be pertinent to evaluate the robustness of specialized English to French machine translation systems. These syntactic constructions involve syntactic and lexical issues that hinder their use and their understanding by non-native English speakers. We argue that such analyses would help ensure that MT systems meet the requirements of their intended end-users.

MOTS-CLÉS : traduction automatique ; évaluation linguistique ; syntagmes nominaux complexes ; langue de spécialité ; discours scientifique.

KEYWORDS: machine translation ; linguistic evaluation ; complex noun phrases ; English for specific purposes ; academic discourse.

1 Introduction

Nous défendons l'idée que l'analyse des erreurs faites lors de la traduction des syntagmes nominaux complexes présente un intérêt pour évaluer la robustesse des systèmes de traduction automatique (TA) en langue de spécialité. Dans cet article, nous justifions tout d'abord l'intérêt d'une analyse linguistique des systèmes de TA (partie 2), puis nous présentons les syntagmes nominaux complexes et les difficultés qu'ils soulèvent (partie 3). Par l'analyse de l'état de l'art, nous montrons que cette construction syntaxique ne semble pourtant pas assez prise en compte aujourd'hui dans l'évaluation des systèmes de TA (partie 4). Puis, nous présentons une typologie d'erreurs adaptées à l'analyse de ce phénomène (partie 5). Nous illustrons ensuite l'intérêt de cette analyse par quelques exemples tirés

Actes des 16e Rencontres Jeunes Chercheurs en RI (RJCRI) et 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL), pages 57–71.

de nos travaux de thèse (partie 6).

2 Pourquoi une analyse linguistique des erreurs ?

Les progrès récents de la traduction automatique neuronale s'accompagnent d'attentes importantes de la part des usagers potentiels (de Sinety, 2019). Les progrès en matière de fluidité font que cet outil est de plus en plus perçu comme une aide potentielle à la rédaction en anglais (Goulet *et al.*, 2017), et comme un outil pour accéder à des connaissances scientifiques dans des langues non maîtrisées (Yvon, 2019; Cros & Kübler, 2019). En parallèle, l'usage de la postédition se développe en milieu professionnel. Dès 2017, Moorkens (Moorkens, 2017) supposait que son utilisation devrait certainement s'accroître en raison de la pression sur les coûts et la productivité. L'étude annuelle European Language Industry Survey menée par un groupement d'associations et d'organisations relevant des services linguistiques semble prouver cette tendance : la part des entreprises et des indépendants déclarant utiliser la traduction automatique est ainsi passée de plus de 50 % en 2018 à 58 % des entreprises et plus de 70 % des indépendants en 2022 (ELIS survey 2022 [1]).

Or, si la traduction neuronale s'est accompagnée d'un bond qualitatif indiscutable (meilleure fluidité, moins d'erreurs lexicales et morphologiques, etc.) et d'une réduction de certains efforts de postédition (Bentivogli *et al.*, 2016), elle n'est pas exempte d'erreurs (Castilho *et al.*, 2017; Esperança-Rodier & Becker, 2018). Les métriques automatiques (BLEU, METEOR...) classiquement utilisées pour évaluer les traductions automatiques ne suffisent plus à comprendre la capacité des systèmes neuronaux au regard de la fluidité des sorties (Burlot & Yvon, 2018).

Dans ce cadre, une analyse de phénomènes linguistiques précis, connus pour poser des difficultés aux utilisateurs finaux, prend tout son sens. En effet, si le système produit des erreurs similaires à celles des humains, il est probable que leur identification soit plus difficile dans le cas d'une postédition professionnelle (tâche par laquelle un être humain vient corriger un texte traduit automatiquement pour le rendre conforme aux attentes du client final) ou d'une utilisation de ces systèmes comme aide à la rédaction.

Depuis 2017, cette approche linguistique est de plus en plus utilisée pour évaluer les systèmes de TA. Macketanz et al. (Macketanz *et al.*, 2017) ont montré l'intérêt d'une telle approche pour identifier les faiblesses des systèmes neuronaux en utilisant un jeu de test contenant des phénomènes linguistiques connus pour poser des difficultés de traduction en anglais-allemand. Isabelle et al. (Isabelle *et al.*, 2017) ont effectué un travail similaire pour la paire anglais-français. Le traitement des noms composés du type < *Noun Noun* > (*juice filter*, *butter knife*...) a par exemple été étudié et les traductions se sont avérées décevantes. Loock (Loock, 2018) a mené une évaluation similaire sur une série de phénomènes linguistiques ayant la réputation de poser problème aux traducteurs anglais-français du fait d'une différence de fréquence d'utilisation entre les deux langues originales. Il a mis en évidence une surreprésentation de ces phénomènes lexicaux et grammaticaux en français traduit automatiquement par rapport au français original. Burlot et Yvon se sont intéressés à la compétence morphologique des systèmes : leurs travaux suggèrent que certains phénomènes grammaticaux sont moins bien modélisés par la traduction neuronale (Burlot & Yvon, 2018).

1. https ://elis-survey.org

3 Les syntagmes nominaux complexes : une structure caractéristique du discours scientifique anglais

Si le syntagme nominal « simple » est constitué d'un nom tête et éventuellement d'un déterminant, le syntagme nominal complexe résulte alors de l'ajout d'unités grammaticales à ces structures simples (syntagme prépositionnel, prémodifieurs adjectivaux, nom adjectival...).

Le discours scientifique en anglais se caractérise par le recours important et croissant aux syntagmes nominaux prémodifiés. Ils permettent l'identification très précise des référents (Biber & Conrad, 2009; Biber & Gray, 2016). Ces groupes sont notamment employés pour compacter ou condenser l'information, et leur usage, comparativement, s'inscrit également dans des choix communicationnels spécifiques à la construction du discours scientifique, comme, par exemple, l'alternance de la pré-et postmodification dans le domaine médical (Gledhill & Pecman, 2018).

Au sein de ces syntagmes nominaux prémodifiés, les syntagmes comprenant une prémodification nominale sont particulièrement complexes. Ces groupes impliquent des questions de syntaxe et de lexique qui entraînent fréquemment des difficultés de découpage syntaxique. Leur complexité peut être accrue par la multiplication des éléments de complexification sur un même groupe considéré ((Rouleau, 2006; Berlage, 2014).

Considérons les exemples suivants tirés d'articles du domaine de la Traduction Automatique des Langues (TAL).

TABLE 1: Exemples de SNC (en gras)

(1)	An attention mechanism implemented by a feedforward neural network is then used to attend specific parts of the input and to generate an alignment between **input and output sequence**.
(2)	While both publications report results of an extensive analysis and comparison of NMT and PBMT approaches, neither of publications deals with language related issues based on **the source and the target language properties** and their differences.
(3)	This scenario is relevant for **the so-called computer-assisted translation (CAT) framework**, which now represents the standard operating environment in the translation industry.
(4)	**A PBSMT and an NMT system** were compared across four translation directions (i.e. from English (EN) into German (DE), Greek (EL), Portuguese (PT), and Russian (RU) in a series of extensive assessment tasks.

Les exemples ci-dessus illustrent la manière dont la complexité n'est pas seulement lexicale (présence d'un terme du domaine, c'est-à-dire d'un mot ou groupe de mots renvoyant à une notion définie dans un domaine de spécialité), mais peut également résulter de la longueur du syntagme en raison du phénomène de récursivité dans la construction des SNC (exemple (3)) ou de l'enchâssement des modifieurs avec par exemple la présence d'une coordination nominale (exemples (1) et (2)).

L'exemple (4) combine deux éléments de complexification avec une difficulté lexicale (présence de deux termes du domaine du TAL : *PBSMT* et *NMT*, et d'une coordination des prémodifieurs de la tête, *system*).

L'exemple (3) illustre la difficulté de découpage syntaxique qui résulte de la longueur potentielle des syntagmes : *so-called* doit-il s'interpréter comme un modifieur du seul *computer-assisted translation (CAT)* ou de *computer-assisted translation (CAT) framework* ?

Ces particularités d'usage et de construction constituent un obstacle important à leur compréhension, leur production et leur traduction pour les locuteurs d'anglais non natifs (Chuquet & Paillard, 1987; Biber & Conrad, 2009; Kübler *et al.*, 2022).

Dans ce contexte, une analyse de la capacité des systèmes de traduction automatique (TA) à traiter de ces syntagmes nominaux complexes comprenant une prémodification nominale (appelés "SNC" dans la suite de cet article) tirés de textes spécialisés authentiques prend tout son sens. Il s'agit de garantir que les textes scientifiques traduits automatiquement et postédités répondent aux exigences linguistiques de la communauté de discours à laquelle ils sont destinés.

4 Un phénomène peu abordé dans les évaluations des systèmes de traduction

Malgré l'importance des SNC pour une part des utilisateurs finaux, l'efficacité des systèmes de traduction automatique sur ce point n'a été que peu abordée (Nakov 2013). De nombreuses études n'abordent la question des SNC qu'à travers soit les mots ou termes composés (*noun compound*) (Lauer, 1995, Baldwin and Tanaka, 2004, Cao and Li, 2002, Moldovan et al., 2004, Nakov, 2013, Balyan and Chatterjee, 2015), et plus particulièrement les seuls syntagmes binominaux (structure du type < Nom Nom >), soit la question de l'identification des mots complexes (*complex word*) lexicalement dans le cadre des recherches sur la simplification des textes. Pour une étude de SNC plus complexes syntaxiquement, il faut se tourner vers les études des unités phraséologiques, collocations ou lexies composées (*multi-word units*) portant sur les n-grammes. La question des SNC y restent cependant très centrée sur les binominaux et les études se focalisent surtout sur la traduction des collocations et des idiomes.

Cette approche des SNC est réductrice, car en se concentrant sur la seule complexité lexicale et plus particulièrement celle portant sur les seuls syntagmes binominaux, une partie des difficultés de traduction est occultée.

5 Une typologie des erreurs permettant de confronter les difficultés des systèmes de traduction et les besoins des utilisateurs finaux

La définition d'une typologie d'erreurs est une tâche complexe : il faut déterminer les classes d'erreurs et leurs niveaux de détails adaptés aux phénomènes que l'on souhaite étudier (Popović, 2018; Chatzikoumi, 2020). Aucune typologie existante dans le domaine de la recherche en traduction

automatique ne semble s'être particulièrement concentrée sur l'identification des erreurs liées aux syntagmes nominaux complexes (SNC).

Les études répertoriées, se concentrant en majorité sur les syntagmes binominaux et l'ambiguïté lexicale qui les caractérisent, n'ont développé des typologies que pour décrire les relations sémantiques entre ces termes. Les autres typologies d'erreurs classiquement utilisées en TAL s'avèrent souvent trop générales pour prendre en compte les spécificités des erreurs des SNC, ou au contraire trop spécialisées sur l'étude d'un phénomène linguistique précis ou une paire de langues données (Vilar *et al.*, 2006; Stymne & Ahrenberg, 2012; Kirchhoff *et al.*, 2012; Costa *et al.*, 2015). Ainsi, une catégorie classique comme "Ordre des mots" pourrait correspondre à des phénomènes très différents pour l'analyse d'un SNC comme une mauvaise attribution d'un modifieur ou une mauvaise identification de la tête. Or, les anglophones langue seconde éprouvent généralement plus de difficultés dans l'attribution des modifieurs aux bons constituants du SNC plutôt qu'à l'identification de la tête. Une erreur du premier type sera donc plus difficilement identifiée et corrigée par un utilisateur final de TA.

Dans le domaine de la traduction professionnelle (agences de traduction, indépendants...), si les approches linguistiques prédominent pour analyser les erreurs produites, elles font appel, ne serait-ce qu'implicitement, à la finalité des traductions pour évaluer la gravité des erreurs commises. Les typologies d'analyses des erreurs, comme la typologie Dynamic Quality Framework (DQF), la Multitidimensional Quality Metrics (MQM) ou la LISA QA, combinent ainsi la prise en compte des erreurs langagières, les attentes du client (guide de rédaction, terminologie propre…), les nécessités de distinction entre différentes variantes linguistiques (par exemple, entre anglais britannique ou américain, ou entre français de France ou français canadien) et les contraintes du processus de production (temps passé, délai de livraison…) ; chacun de ces éléments est pondéré en fonction de sa gravité. Ces typologies professionnelles s'avèrent également inadaptées à l'analyse des erreurs linguistiques sur les SNC en raison du fort l'accent mis sur les erreurs de localisation pour répondre aux besoins d'un client spécifique, ce qui rend difficile leur utilisation sur une grande diversité de genres textuels. Concernant l'identification des erreurs, elles présentent les mêmes limites que les typologies utilisées dans le domaine de la TA.

Pour trouver une typologie adaptée, il faut se tourner vers les typologies utilisées dans l'enseignement de la traduction anglais-français ou de l'anglais comme langue seconde. Mais, si la question des termes complexes en anglais a fait l'objet de très nombreuses études, peu d'entre elles se sont intéressées à la question des syntagmes nominaux complexes dans les langues de spécialité et la grande majorité s'est concentrée sur les syntagmes binominaux et l'ambiguïté lexicale qui les caractérisent (Tournier, 1985). Maniez (Maniez, 2008, 2017, 2020) s'est ainsi intéressé aux stratégies de traduction des traducteurs français et des systèmes de TA face aux SNC anglais dans le domaine médical sans toutefois développer une typologie détaillée des erreurs.

La typologie d'erreurs développée par Kübler, Mestivier et Pecman (Kübler *et al.*, 2022) s'avère particulièrement intéressante : elle résulte d'un inventaire sur plusieurs années portant sur les erreurs faites par les apprenants en traduction lors de la postédition d'articles de recherche du domaine des sciences de la Terre. Cette typologie nous parait parfaitement adaptée aux besoins présentés dans cet article, car elle permet de comparer les difficultés rencontrées par des utilisateurs finaux (non-identification ou sous-identification d'un type d'erreur, difficultés d'identification des composants d'un syntagme...) avec les difficultés éventuelles des systèmes de TA.

Afin de tenir compte des spécificités d'erreurs produites par les systèmes de TA, Bénard (Bénard, 2019) a adapté cette typologie d'erreurs à l'évaluation de textes produits par des systèmes neuronaux. Des erreurs spécifiques à la TA ont ainsi été ajoutées comme l'instabilité des choix de traduction d'un

mot au sein d'un même texte ou l'ajout d'unités lexicales lors de la traduction tendant à modifier le sens du syntagme nominal source.

La typologie finale obtenue est présentée dans la Table 2 page 6. C'est cette typologie que nous proposons d'utiliser pour l'analyse des erreurs.

Niveau 1	Niveau 2
Erreurs terminologiques, en genre ou en nombre	1. Incohérence de la traduction intra- et extratextuelle 2. Mauvaise traduction d'une unité lexicale qui n'est pas un terme du domaine ou du genre 3. Traduction d'un terme par un non-terme 4. Unité lexicale non traduite 5. Traduction d'un nom propre ou d'un nom commun qui ne doit pas être traduit 6. Problème de détermination, d'erreur en nombre ou en genre
Erreurs d'analyse syntaxique du texte source	7. Identification erronée de la tête 8. Attribution du modifieur au mauvais constituant 9. Factorisation non détectée (ou partiellement) d'un constituant dans la coordination 10. Ajout d'une factorisation n'existant pas en langue source 11. Mauvaise interprétation des liens sémantiques entre les constituants, hors factorisation
Adaptation erronée à la syntaxe de la langue cible	12. Absence d'une explicitation pourtant nécessaire 13. Calque erroné sur la langue source ou influence de la langue source 14. Inadaptation des poids des constituants
Erreurs de transferts	15. Ajout injustifié d'une unité lexicale ou d'une ponctuation
	16. Omission d'un constituant, voire de tout le syntagme lors de la traduction

TABLE 2: Typologie des erreurs de Bénard (Bénard, 2019) sur la base de Kübler et al. (Kübler *et al.*, 2022)

6 Résultats préliminaires

6.1 Méthode

6.1.1 Choix du corpus

Nous avons appliqué cette approche linguistique dans le cadre de notre thèse en analysant un corpus d'articles de recherche, traduits simultanément par plusieurs systèmes. Afin d'illustrer l'intérêt d'une telle approche, nous présentons, dans cette partie, deux exemples tirés d'une analyse diachronique à l'aide de deux systèmes commerciaux, non spécialisés et accessibles en ligne : DeepL et Systran. Les textes du corpus de test ont été traduits une première fois en décembre 2019 et une seconde fois en août 2022.

L'ensemble du corpus de test de cette analyse diachronique est constitué de 4 articles de recherche rédigés en anglais, tirés de ACL Anthology [2] : 2 articles de synthèse (*review papers*) et 2 articles de recherche expérimentale (*research papers*).

Les textes retenus ont été écrits ou co-écrits par des chercheurs spécialistes du domaine de la traduction automatique : la majorité des auteurs ont un indice de Hirsch (indice h) supérieur à 10 – six d'entre eux ont un indice supérieur à 20 – et leur indice i10 de Google Scholar (nombre d'articles qui ont été cités au moins 10 fois) est généralement élevé. Les articles eux-mêmes ont été publiés dans des revues à comité de lecture. Ils ont également été cités de très nombreuses fois depuis leur publication en 2017, ce qui pourrait témoigner de la capacité des chercheurs du domaine à s'approprier ces textes. Nous pouvons donc supposer que ces articles suivent les conventions d'écriture, la phraséologie et la terminologie attendues par le public du domaine du TAL.

Les textes ont été modifiés avant d'être soumis à la traduction. Les images, les références, les en-têtes, les pieds de page et les éléments spécifiques de mise en page (italiques, saut de page...) ont été supprimés. Seuls l'abstract et le corps de l'article ont été conservés. Les phrases éventuellement coupées par une image ou un tableau ont été reconstituées.
En revanche, les textes n'ont pas été corrigés : les erreurs d'orthographe, de grammaire ou de typographie ont été conservées. En effet, ce sont des éléments que les systèmes de TA devront être capables de traiter lors de leur utilisation par des utilisateurs finaux (postéditeur, rédacteur...). De plus, les textes n'ont pas été segmentés avant d'être soumis aux systèmes de traduction.

6.1.2 Identification des syntagmes

Les SNC ont été extraits de manière semi-automatique. Les textes traduits ont été étiquetés à l'aide du logiciel Le Trameur [3] qui intègre TreeTagger, un système d'étiquetage automatique des catégories grammaticales des mots avec lemmatisation sur la base du jeu d'étiquettes The Penn Treebank. Notre choix s'est porté sur cette combinaison d'outils en raison de de leur bon équilibre entre l'efficience de l'étiquetage (ratio rapidité/fiabilité) en recourant au système intégré, la souplesse offerte par les formats en sortie et l'affichage de corpus parallèle multilingue, et les possibilités d'analyse statistique et d'annotation intégrées.

2. Voir les références complètes en fin d'article
3. Version 12.176. http ://www.tal.univ-paris3.fr/trameur/

Sur la base de cet étiquetage, les syntagmes candidats sont extraits par repérage de patrons syntaxiques, c'est-à-dire sur la base de l'ordre d'enchaînement des catégories grammaticales formant un SNC. La table ci-dessous présente un extrait des patrons considérés (Table 3).

TABLE 3: Exemples de patrons syntaxiques identifiés lors de l'analyse

Patrons syntaxiques	**Exemples de syntagmes identifiés**
NN NN	MT systems
et par récursivité :	
DT NP NP NN NN	the "a posteriori" adaptation strategy
DT NN NN	a product listing
DT JJ NN NN	high quality submissions
JJ NN NN NNS	a large output vocabulary
etc.	
JJ NN CC NN	additional nouns and adjectives
et par récursivité :	
JJ NN NNS CC NNS	chemical patent titles and abstracts
etc.	

Toutefois, des erreurs sont présentes en raison des limitations techniques du système d'étiquetage utilisé et de l'ambiguïté inhérente à certains termes anglais. Afin de réaliser une étude qualitative, il est nécessaire de corriger ces erreurs manuellement comme l'illustre l'exemple ci-dessous.

ID_partition-phrase	GNC_ID_2022	list_GNC_POS_ID	list_GNC_Lemme	list_GNC_Forme	list_GNC_POS	Texte source
146_T2	1907.0	['13829', '13830', '13831', '13832', '13833', '13834', '13835']	leverage past gradient information	leverage past gradient information	NN JJ JJ NN	For this reason, optimizers that can leverage past gradient information are usually more reliable.

FIGURE 1 – Exemple de SNC identifiés automatiquement à corriger

L'extrait de notre base (Figure 1) montre l'un des syntagmes identifiés automatiquement (*leverage past gradient information*) et les informations données par l'étiqueteur Tree-Tagger. Nous observons que l'étiqueteur a identifié de manière erronée *leverage* comme un nom, comme le prouve la mention *NN* dans la colonne *list_GNC_POS*. Il s'agit en réalité de la forme verbale et le mot n'aurait jamais dû être identifié comme un composant du syntagme nominal à étudier. Il apparaît donc nécessaire de supprimer l'intégralité des éléments marqués en rouge dans les cellules afin d'obtenir le syntagme correct.

Ces syntagmes candidats ont donc été corrigés manuellement par un linguiste au fur et à mesure de l'identification des erreurs. Ces bases de données, créées automatiquement, contiennent de nombreuses informations indispensables à l'analyse des syntagmes (liste des catégories grammaticales, liste des lemmes, liste des formes, etc. constituant les syntagmes). Afin de limiter les biais dans l'analyse, un

second annotateur a également annoté une partie du corpus et un accord inter-annotateur a été calculé. Au total, les traductions de 1072 SNC ont été étudiées, soit 4288 syntagmes traduits.

Les sous-section suivantes, 6.2 et 6.3, présentent deux exemples tirés de ce corpus qui illustrent l'intéret d'analyser les SNC pour l'évaluation des systèmes de traduction spécialisés.

6.2 Exemple 1 : Des différences de performances significatives

En 2019, Systran se démarquait par une difficulté à gérer la présence de coordination dans les prémodifieurs nominaux avec 23% des groupes de ce type faisant l'objet d'une erreur contre une seule occurence pour DeepL. Considérons les exemples ci-dessous (Table 4).

TABLE 4: Exemples de traduction erronées produites en 2019 par le système commercial non spécialisé Systran (le syntagme analysé est en gras)

(1.a)	An attention mechanism implemented by a feedforward neural network is then used to attend specific parts of the input and to generate an alignment between **input and output sequence**.
(1.b)	Un mécanisme d'attention mis en place par un réseau neuronal de flux est ensuite utilisé pour assister à des parties spécifiques de l'entrée et générer un alignement entre **l'entrée et la séquence de sortie.**
(2.a)	While both publications report results of an extensive analysis and comparison of NMT and PBMT approaches, neither of publications deals with language related issues based on **the source and the target language properties** and their differences.
(2.b)	Bien que les deux publications fassent état des résultats d'une analyse et d'une comparaison approfondies des approches NMT et PBMT, aucune de ces publications ne traite des questions liées à la langue en fonction de **la source et des propriétés linguistiques cibles** et de leurs différences.

Le exemples (1.a) et (2.a) correspondent aux phrases d'origine soumises aux systèmes et les exemples (1.b) et (2.b) correspondent aux traductions.

Les erreurs portent essentiellement sur l'identification de la première partie du constituant : le moteur n'identifie généralement pas que le premier nom du syntagme coordonné (respectivement *input* et *the source*) est également un modifieur de la tête (respectivement *sequence* et *language properties*). Il traite chacun des constituants de la coordination comme un syntagme nominal différent.

Cette différence marquée entre deux systèmes commerciaux peut s'avérer intéressante pour un utilisateur final. En effet, il peut être difficile pour un traducteur ou un utilisateur final de déterminer l'intention du locuteur et donc d'identifier correctement les limites des syntagmes nominaux. Cela pourrait amener un utilisateur final à privilégier un système de TA plus efficace sur ce point, lorsque ce type de construction syntaxique coordonnée s'avère particulièrement utilisée dans son domaine de spécialité, et ce, afin de limiter le temps à passer en postédition.

6.3 Exemple 2 : Une pérennité des performances qui n'est pas garantie

Les exemples présents dans la Table 5 ci-dessous montrent la continuité des difficultés de traduction dans le temps.

TABLE 5: Exemples de traduction erronées produites par les systèmes commerciaux non spécialisés DeepL et Systran en 2019 et 2022 (les syntagmes analysés sont en gras)

(3.a)	Texte original	The goal of this evaluation was to compare the performance between **the mature Chinese to English patent MT engines** used in production at Iconic with **novel NMT engines** developed at the ADAPT Centre on an 'apples to apples' basis, trained on the same available data.
(3.b)	DeepL 2019	Le but de cette évaluation était de comparer les performances **des moteurs de TA brevetés chinois à anglais** utilisés dans la production à Iconic avec les nouveaux moteurs de TA développés au Centre ADAPT sur une base de " pommes à pommes ", formés sur la base des mêmes données disponibles.
(3.c)	DeepL 2022	L'objectif de cette évaluation était de comparer les performances **des moteurs de TA du chinois à l'anglais** utilisés en production chez Iconic avec les nouveaux moteurs de NMT développés au Centre ADAPT, sur une base de comparaison, entraînés sur les mêmes données disponibles.
(3.d)	Systran 2019	L'objectif de cette évaluation était de comparer la performance entre **les moteurs chinois matures et anglais brevetés MT** utilisés en production à Iconic avec les nouveaux moteurs NMT mis au point au Centre ADAPT sur une base de 'pommes à pommes', formés sur les mêmes données disponibles.
(3.e)	Systran 2022	Le but de cette évaluation était de comparer les performances **des moteurs MT de brevets chinois à anglais** utilisés dans la production chez Iconic avec les nouveaux moteurs NMT développés au Centre ADAPT sur une base 'pomme à pomme', formés sur la base des mêmes données disponibles.

L'exemple (3.a) correspond à la phrases d'origine soumise aux systèmes de traduction, les exemples (3.b) et (3.d) correspondent aux traductions produites respectivement par DeepL et Systran en 2019 et les exemples (3.c) et (3.e) correspondent aux traductions produites respectivement par DeepL et Systran en 2022.

De manière générale, nous constatons une difficulté à interpréter les liens sémantiques dans les SNC de grande taille. Si les solutions ou les types d'erreurs observées peuvent évoluer ou les erreurs ne pas concerner exactement les mêmes composants, on constate un certain maintien des erreurs observées

ou leur apparition sur d'autres SNC. En sens inverse, des SNC ou des éléments de SNC qui pouvaient être bien traités en 2019 s'avèrent mal traduit en 2022.

Considérons le syntagme *the mature Chinese to English patent MT engines* dans l'exemple (3a) dont la traduction attendue est du type *les moteurs aboutis pour la traduction automatique de brevets du chinois vers l'anglais*. il faut noter que ce syntagme combine plusieurs niveaux de complexité, notamment une complexité lexicale (présence de terme du domaine du Traitement Automatique des Langues et de non termes du domaine) et une complexité syntaxique (8 éléments).

Nous constatons que le système DeepL conserve la difficulté à traduire *patent* : même si la solution adoptée est différente entre 2019 (erreur d'analyse syntaxique avec une mauvaise interprétation du lien sémantique modifieur/modifié) et 2022 (non traduction du terme, c'est-à-dire une erreur de transfert), la traduction reste erronée. De manière générale, le système éprouve des difficultés à interpréter les liens sémantiques entre les composants du SNC, ce qui entraîne une mauvaise attribution des modifieurs et une traduction fausse. Le système Systran continue également à faire une erreur terminologique en ne traduisant pas *MT* par *TA*, alors même que cette traduction est bien faite dans d'autres segments traduits du même texte.

Si nous considérons le système Systran, nous constatons une amélioration de l'identification des liens sémantiques entre 2019 et 2022 : le mot *patent* est bien considéré en 2022 comme limitant le sens de *engines* en indiquant le domaine de spécialité du moteur considéré (traduction par *de brevets*), alors qu'il avait été analysé comme qualifiant le statut du moteur (traduction par *brevetés*). Cette amélioration s'accompagne cependant d'une dégradation importante avec l'omission de la traduction du mot *mature* ce qui entraîne une distorsion du sens en français par rapport à l'original.

Cet exemple illustre la pérennité des difficultés dans le temps pour traduire les SNC et la difficulté à maintenir les performances observées à un moment donné. Or, dans le cadre des systèmes disponibles en ligne, les utilisateurs finaux n'ont aucune information sur les évolutions de version qui leur permettrait de juger des évolutions sur la qualité des traduction. Ils peuvent également manquer des compétences linguistiques ou informatiques et des ressources financières ou matérielles pour mener une telle évaluation. Il apparaît donc important de prendre en compte ce point lors de l'évaluation de la qualité des systèmes de TA dès leur développement.

7 Conclusion et perspectives

Nous défendons l'idée que l'analyse des erreurs faites lors de la traduction des syntagmes nominaux complexes présente un intérêt pour évaluer la robustesse des systèmes de traduction automatique (TA) en langue de spécialité. Plus particulièrement, l'analyse des syntagmes nominaux présentant une prémodification nominale pour le premier modifieur (noté "SNC" dans cet article) est particulièrement adapté à la prise en compte des besoins des utilisateurs finaux, qu'il soit un professionnel de la langue (traducteur-postéditeur, rédacteur...) ou un spécialiste du domaine recherchant une aide à la rédaction vers l'anglais ou une compréhension rapide du contenu d'un texte.

Cette construction syntaxique ne semble pourtant pas assez prise en compte aujourd'hui dans l'évaluation des systèmes de TA. Au mieux, ces questions sont partiellement intégrées à l'étude des composés nominaux, des unités phraséologiques ou aux questions sur la complexité lexicale des textes en vue de leur simplification. Les analyses se limitent souvent à la question des syntagmes binominaux ou à l'aspect terminologique sans prendre en compte les autres facettes possibles de la complexité comme

la longueur ou l'enchâssement.

Nous avons mené une première analyse diachronique basée sur une typologie permettant de mettre en parallèle les difficultés des systèmes avec les besoins des utilisateurs finaux. Dans cet article, nous nous sommes concentrés sur la présentation de notre méthodologie. Nous présentons deux exemples illustrant qu'en appliquant cette méthodologie à des textes du domaine du Traitement Automatique des Langues, il est possible de montrer que les systèmes de TA produisent des erreurs similaires en partie à celles produites par un traducteur humain, ce qui peut rendre leur détection difficile par un utilisateur final. Considérant que la construction syntaxique de ces syntagmes varient d'une langue de spécialité à l'autre, nous pensons donc que l'efficacité des systèmes de TA en manière de traduction des SNC pourrait constituer un facteur dans le choix d'un utilisateur final d'utiliser l'un ou l'autre des systèmes à sa disposition.

Nos travaux actuels et futurs visent à présenter de manière approfondie les résultats de cette analyse et à étendre celle-ci à d'autres domaines de spécialité (pour prendre en compte la diversité de construction des SNC) et à des systèmes de TA spécialisés (pour voir si la spécialisation dans un domaine permet d'améliorer la traduction des SNC). Au-delà du développement logiciel, une analyse fine des erreurs produites par les systèmes neuronaux est essentielle dans une visée pédagogique pour la formation des postéditeurs professionnels ou occasionnels (comme, par exemple, les spécialistes d'autres disciplines souhaitant une aide à la rédaction pour l'écriture d'un article scientifique dans une langue autre que sa langue maternelle).

Les corpus annotés seront mis librement à disposition de la communauté à l'automne 2023.

Remerciements

Ma directrice de thèse, Natalie Kübler, professeur des universités, laboratoire CLILLAC-ARP, université Paris Cité (France).

Références des textes inclus dans le corpus de test

CASTILHO S., MOORKENS J., GASPARI F., CALIXTO I., TINSLEY J. & WAY A. (2017). Is Neural Machine Translation the New State of the Art ? *The Prague Bulletin of Mathematical Linguistics*, 108(1), 109–120. DOI : 10.1515/pralin-2017-0013.

COSTA-JUSSÀ M. R., ALLAUZEN A., BARRAULT L., CHO K. & SCHWENK H. (2017). Introduction to the special issue on deep learning approaches for machine translation. *Computer Speech I& Language*, 46(Journal Article), 367–373. DOI : 10.1016/j.csl.2017.03.001.

POPOVIC M. (2017). Comparing Language Related Issues for NMT and PBMT between German and English. *The Prague Bulletin of Mathematical Linguistics*, 108(1), 209–220. DOI : 10.1515/pralin-2017-0021.

TURCHI M., NEGRI M., FARAJIAN M. A. & FEDERICO M. (2017). Continuous Learning from Human Post-Edits for Neural Machine Translation. *The Prague Bulletin of Mathematical Linguistics*, 108(1), 233–244. DOI : 10.1515/pralin-2017-0023.

Références

BENTIVOGLI L., BISAZZA A., CETTOLO M. & FEDERICO M. (2016). Neural versus Phrase-Based Machine Translation Quality : a Case Study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, p. 257–267, Austin, Texas : Association for Computational Linguistics. DOI : 10.18653/v1/D16-1025.

BERLAGE E. (2014). *Noun Phrase Complexity in English*. Studies in English Language. United Kingdom : Cambridge University Press.

BIBER D. & CONRAD S. (2009). *Register, Genre, and Style*. Cambridge Texbooks in Linguistics. Cambridge University Press.

BIBER D. & GRAY B. (2016). *Grammatical Complexity in Academic English : Linguistic Change in Writing*. Studies in English Language. United Kingdom : Cambridge University Press.

BURLOT F. & YVON F. (2018). Évaluation morphologique pour la traduction automatique : adaptation au français (Morphological Evaluation for Machine Translation : Adaptation to French). In *Actes de la Conférence TALN. Volume 1 - Articles longs, articles courts de TALN*, p. 61–74, Rennes, France : ATALA.

BÉNARD M. (2019). Mémoire de Master 2 Recherche LSCT : Évaluation de la qualité des systèmes neuronaux en matière de traduction des groupes nominaux complexes à prémodification nominale.

CASTILHO S., MOORKENS J., GASPARI F., SENNRICH R., SOSONI V., GEORGAKOPOULOU P., LOHAR P., WAY A., VALERIO MICELI BARONE A. & GIALAMA M. (2017). A Comparative Quality Evaluation of PBSMT and NMT using Professional Translators. In *Proceedings of MT Summit XVI*, volume Vol. 1 : Research Track, p. 116–131, Nagoya, Aichi, Japan. https ://www.researchgate.net/profile/Vilelmini-Sosoni/publication/320016264_A_Comparative_Quality_Evaluation_of_PBSMT_and_NMT_using_Prc Comparative-Quality-Evaluation-of-PBSMT-and-NMT-using-Professional-Translators.pdf.

CHATZIKOUMI E. (2020). How to evaluate machine translation : A review of automated and human metrics. *Natural Language Engineering*, **26**(2), 137–161. Publisher : Cambridge University Press, DOI : 10.1017/S1351324919000469.

CHUQUET H. & PAILLARD M. (1987). *Approche linguistique des problèmes de traduction*. Paris, France : Éditions Ophrys. Réimpression 2015 d'un ouvrage de 1987.

COSTA A., LING W., LUÍS T., CORREIA R. & COHEUR L. (2015). A linguistically motivated taxonomy for Machine Translation error analysis. *Machine Translation*, **29**, 127–161. DOI : 10.1007/s10590-015-9169-0.

CROS I. & KÜBLER N. (2019). Du FOU au FLA à l'Université française ou de la pédagogie universitaire à l'écrit scientifique spécialisé. *FIU Francophonie et innovation à l'université*.

DE SINETY P. (2019). "Pour un plurilinguisme de la pensée scientifique". Library Catalog : www.culture.gouv.fr.

ESPERANÇA-RODIER E. & BECKER N. (2018). Comparaison de systèmes de traduction automatique, probabiliste et neuronal, par analyse d'erreurs. Nancy, France.

GLEDHILL C. & PECMAN M. (2018). On alternating pre-modified and post-modified nominals such as aspirin synthesis vs. synthesis of aspirin : Rhetorical and cognitive packing in English science writing. *Fachsprache : Internationale Zeitschrift für Fachsprachenforschung- didaktik und Terminologie*, **40**(1), 24–46.

GOULET M.-J., SIMARD M., PARRA ESCARTÍN C. & O'BRIEN S. (2017). La traduction automatique comme outil d'aide à la rédaction scientifique en anglais langue seconde : résultats d'une étude exploratoire sur la qualité linguistique. *ASp. la revue du GERAS*, (72), 5–28. Number : 72 Publisher : Groupe d'étude et de recherche en anglais de spécialité, DOI : 10.4000/asp.5045.

ISABELLE P., CHERRY C. & FOSTER G. (2017). A Challenge Set Approach to Evaluating Machine Translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, p. 2476–2486, Copenhagen, Denmark. arXiv : 1704.07431.

KIRCHHOFF K., CAPURRO D. & TURNER A. (2012). Evaluating User Preferences in Machine Translation Using ConjointAnalysis. In *Proceedings of the 16th EAMT Conference*, p. 119–126, Trento, Italy : European Association for Machine Translation.

KÜBLER N., MESTIVIER A. & PECMAN M. (2022). Using Comparable Corpora for Translating and Post-Editing Complex Noun Phrases in Specialised Texts : Insights from English-to-French Specialised Translation. In S. GRANGER & M.-A. LEFER, Éds., *Extending the Scope of Corpus-Based Translation Studies*, Bloomsbury Advances in Translation. Bloomsbury Academic. https ://www.bloomsbury.com/us/extending-the-scope-of-corpusbased-translation-studies-9781350143258 24/02/2022.

LOOCK R. (2018). Traduction automatique et usage linguistique : une analyse de traductions anglais-français réunies en corpus. *Meta : journal des traducteurs / Meta : Translators' Journal*, **63**(3), 786–806. DOI : https ://doi.org/10.7202/1060173ar.

MACKETANZ V., AVRAMIDIS E., BURCHARDT A., HELCL J. & SRIVASTAVA A. (2017). Machine Translation : Phrase-Based, Rule-Based and Neural Approaches with Linguistic Evaluation. *Cybernetics and Information Technologies*, **17**(2), 28–43. Publisher : Sciendo Section : Cybernetics and Information Technologies, DOI : 10.1515/cait-2017-0014.

MANIEZ F. (2008). Traduction automatique et ambiguïté syntaxique : le cas de la coordination dans les groupes nominaux complexes en anglais médical. p. 765–776, Lyon. DOI : 10.4000/asp.500.

MANIEZ F. (2017). An appraisal of recent breakthroughs in machine translation : the case of past participle-based compound adjectives in ESP. *ASp*, **72**, 29–48. DOI : 10.4000/asp.5059.

MANIEZ F. (2020). Chapter 18 : The identification of potentially ambiguous noun phrases in scientific english : a crucial aspect of translator. In *Strategies and Analyses of Language and Communication in Multilingual and International Contexts*, p. 187–195. Cambridge Scholars Publishing.

MOORKENS J. (2017). Under pressure : translation in times of austerity. *Perspectives*, **25**, 464–477. DOI : 10.1080/0907676X.2017.1285331.

POPOVIĆ M. (2018). Error Classification and Analysis for Machine Translation Quality Assessment. In *Translation Quality Assessment : From Principles to Practice.*, volume 1 de Machine Translation : Technologies and Applications, p. 129–158. Springer International Publishing, 1sd. ed édition. https ://doi.org/10.1007/978-3-319-91241-7_7.

ROULEAU M. (2006). Complexité de la phrase en langue de spécialité : mythe ou réalité ? Le cas de la langue médicale. *Panace@*, **VII**(24), 298–306.

STYMNE S. & AHRENBERG L. (2012). On the practice of error analysis for machine translation evaluation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, p. 1785–1790, Istanbul, Turkey : European Language Resources Association (ELRA).

TOURNIER J. (1985). *Introduction à la lexicogénétique de l'anglais contemporain*. Genève : Édition Slatkine. Réimpression de l'édition de 1985.

VILAR D., XU J., D'HARO L. & NEY H. (2006). Error Analysis of Machine Translation Output. p. 697–702, Genoa, Italy.

YVON F. (2019). Les deux voies de la traduction automatique. *Hermes, La Revue*, **n° 85**(3), 62–68. Publisher : C.N.R.S. Editions.

Timothy Baldwin and Takaaki Tanaka. 2004. Translation by machine of complex nominals. In Proceedings of the Workshop on Multiword Expressions Integrating Processing - MWE '04, pages 24–31. Association for Computational Linguistics.

Renu Balyan and Niladri Chatterjee. 2015. Translating noun compounds using semantic relations. Comput Speech Lang., 32(1) :91–108.

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Vers une implémentation de la théorie sens-texte avec les grammaires catégorielles abstraites

Marie Cousin[1]
(1) Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
marie.cousin@loria.fr

RÉSUMÉ

La théorie sens-texte est une théorie linguistique visant à décrire la correspondance entre le sens et le texte d'un énoncé à l'aide d'un outil formel qui simule l'activité langagière d'un locuteur natif. Nous avons mis en place une premiere implémentation de cette théorie à l'aide des grammaires catégorielles abstraites, qui sont un formalisme grammatical basé sur le λ-calcul. Cette implémentation représente les trois niveaux de représentation sémantique, syntaxique profonde et syntaxique de surface de la théorie sens-texte. Elle montre que la transition de l'un à l'autre de ces niveaux (en particulier la génération d'une représentation de syntaxe de surface à partir d'une représentation sémantique d'un même énoncé) peut être implémentée en utilisant les propriétés avantageuses des grammaires catégorielles abstraites, dont la transduction.

ABSTRACT

Towards an implementation of meaning-text theory with abstract categorial grammars

Meaning-text theory is a linguistic theory aimed at describing the relation between meaning and text in natural language, with a formal device that simulates the linguistic activity of a native speaker. We established a first implementation of this theory with abstract categorial grammars, which are a grammatical formalism based on λ-calculus. This implementation represents the following three representation levels of meaning-text theory : semantic, deep syntactic and surface syntactic. It shows that transitions between these levels (especially the generation of a surface syntactic representation starting from a semantic representation of the same utterance) can be implemented using advantageous properties of abstract categorial grammars, like transduction.

MOTS-CLÉS : Grammaires Catégorielles Abstraites, Théorie Sens-Texte, Paraphrase.

KEYWORDS: Abstract Categorial Grammars, Meaning-Text Theory, Paraphrase.

1 Introduction

Cet article s'intéresse à l'implémentation de la théorie sens-texte (TST) à l'aide de grammaires catégorielles abstraites (ACG), afin de générer du texte. La TST est une théorie linguistique ayant déjà été utilisée en génération, tandis que les ACG sont un formalisme grammatical s'appuyant sur un système de types. Nous utiliserons donc des méthodes formelles. Ce choix est motivé par notre volonté d'avoir un texte très contrôlé, afin d'être sûrs que le texte généré est bien celui voulu et que le message transmis est bien le message que l'on voulait transmettre. Ces mêmes motivations se retrouvent aussi dans Grammatical Framework (Ranta, 2004), qui s'appuie, comme les ACG, sur un système de types. L'encodage et les liens vers d'autres formalismes dans les ACG a beaucoup été étudié (cf. tableau 1).

Actes des 16e Rencontres Jeunes Chercheurs en RI (RJCRI) et 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL), pages 72–86.

Étant donné que nous voulons avoir un très fort contrôle sur les différentes structures linguistiques aux différents niveaux, pour avoir les structures de la TST, nous avons privilégié une approche avec des méthodes formelles et non pas des méthodes numériques à base de réseaux de neurones.

Les ACG (de Groote, 2001) sont un formalisme, un cadre logique, utilisant des λ-termes et les propriétés du λ-calcul telles que la β-réduction. Ce cadre logique que sont les ACG permet de représenter d'autres formalismes grammaticaux. Les ACG ont de nombreux avantages, dont les suivants : elles sont réversibles, on peut donc les utiliser en génération ou analyse (Kanazawa, 2007). Leurs capacités à encoder d'autres formalismes, comme les TAG (Pogodalla, 2017a), et à être utilisées en génération ont été mises en oeuvre pour généraliser le modèle G-TAG (Danlos *et al.*, 2014). Elles sont également utilisées dans un cadre industriel chez Yseop. Le tableau ci-dessous (Pogodalla, 2017b) illustre le pouvoir expressif des ACG. Une hiérarchie des ACG est utilisée dans ce tableau, se basant sur deux notions (l'ordre et la complexité d'une ACG), qui sont définis plus bas en section 3.

ACG	langage généré
$\text{ACG}_{(1,n)}$	langages finis
$\text{ACG}_{(2,1)}$	langages réguliers
$\text{ACG}_{(2,2)}$	langages hors-contexte
$\text{ACG}_{(2,3)}$	grammaires hors-contexte multiples bien équilibrées
$\text{ACG}_{(2,4)}$	grammaires faiblement contextuelles
$\text{ACG}_{(2,4+n)}$	$\text{ACG}_{(2,4)}$
$\text{ACG}_{(3,n)}$	décidabilité de MELL

TABLE 1 – Pouvoir expressif des ACGs (Pogodalla, 2017b)

Dans cet article, nous cherchons à appliquer ces propriétés à une autre théorie linguistique ayant déjà démontré son intérêt pour les systèmes de génération : la TST.

La TST est en effet une théorie linguistique qui se donne pour objectif de décrire la correspondance entre le sens d'un énoncé et sa représentation sous forme textuelle par un outil formel, le modèle sens-texte, qui simule l'activité langagière d'un locuteur natif. Elle utilise entre autres une syntaxe de dépendances et les concepts clés de paraphrase et de fonctions lexicales (FL) (cf. sections suivantes), ces derniers permettant de rendre plus naturel le texte en question. En particulier, les FL syntagmatiques (cf. section 2) encodent les collocations ou encore les verbes supports. Elles sont donc très importantes, surtout lors des réalisations de surface. La direction privilégiée par cette théorie est la direction du sens vers le texte, soit la génération. Certaines formalisations et implémentations de ce modèle ont déjà eu lieu, telles que les grammaires polarisées et GUST (Kahane, 2005; Kahane & Lareau, 2005; Lareau, 2007) et MARQUIS (Wanner *et al.*, 2010; Lambrey & Lareau, 2015; Lareau *et al.*, 2018).

Nous cherchons ici à implémenter la TST avec les ACG et en particulier les structures linguistiques utilisées par la TST, même si pour chacun de ses niveaux de représentation il existe des formalismes relativement proches. Par exemple, au niveau sémantique la TST utilise en effet des graphes, les notions de prédicats et d'arguments, et se rapproche des AMR (Banarescu *et al.*, 2013). Au niveau de représentation syntaxique, elle utilise des arbres de dépendances, tout en ayant des types d'étiquettes différents d'autres formalismes en dépendance. Nous nous intéressons ici à l'articulation et l'unité au sein de la TST de ces structures linguistiques.

Une question se pose alors : les ACG sont-elles adaptées à implémenter une théorie linguistique basée sur une syntaxe de dépendance dans un but de génération ? Cet article présente la faisabilité d'une telle implémentation en nous appuyant sur un nombre restreint d'exemples qui mettent en œuvre plusieurs spécificités de la TST.

2 Théorie sens-texte et les fonctions lexicales

La TST (Mel'čuk *et al.*, 2012; Milićević, 2006) utilise le modèle sens-texte pour faire le lien entre le sens et le texte d'un énoncé. Le sens est le contenu linguistique à communiquer (Milićević, 2006) ; il n'est pas directement observable. Le texte est un fragment de discours (Milićević, 2006), et est immédiatement perceptible.

2.1 Théorie sens-texte

Le modèle sens-texte est composé de 7 niveaux de représentation : les niveaux de représentation sémantique (SemR), syntaxique profond (DSyntR), syntaxique de surface (SSyntR), morphologique profond (DMorphR), morphologique de surface (SMorphR), phonétique profond (DPhonR) et phonétique de surface (SPhonR) (cf. figure 1). Il comprend également 6 modules de transition entre ces niveaux, avec entre autres le module sémantique ↔ syntaxe profonde. Le modèle proposé par la TST permet donc d'analyser ou de générer un énoncé, selon le sens de transition choisi entre les niveaux.

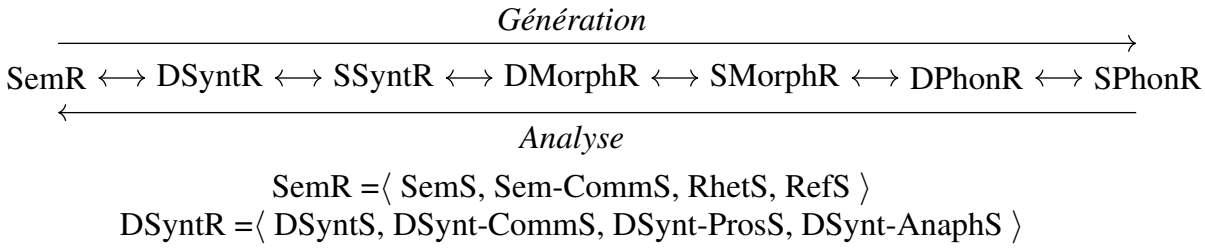


FIGURE 1 – Schéma du modèle sens-texte et détail des structures composant les représentations sémantiques et syntaxiques profondes (Mel'čuk *et al.*, 2012; Milićević, 2006)

Chacun de ces niveaux de représentation est composé de différentes structures : la structure principale ainsi que une ou plusieurs structures complémentaires (cf. figure 1 détaillant les structures dont sont composés les niveaux SemR et DSyntR). Par exemple, SemR est composé des structures sémantique (SemS, cf. figure 2 où cette structure est représentée pour deux expressions), sémantique communicative (Sem-CommS), sémantique rhétorique (RhetS) et sémantique de référence (RefS). La structure principale de SemR est SemS et est représentée sous forme de graphe orienté (cf. figure 2). Elle représente les liens (sémantiques) entre les sémantèmes (ou unités sémantiques) composant le sens exprimé. Mel'čuk *et al.* (2012) donnent plus de détail quant aux règles de bonne formation de cette structure. Sem-CommS indique (sur SemS, comme une sorte d'annotation) où vont se situer le thème et le rhème (cf. Milićević (2006), page 10, où un exemple est détaillé), RhetS va indiquer le style de l'expression (neutre, ironique, joyeux, etc.), et RefS est un ensemble de pointeurs partant des representations sémantiques vers les entités leur correspondant du monde réel. Cependant, ces trois dernières structures ne seront pas détaillées ou utilisées ici, seule l'est SemS, car c'est sur

celle-ci que se concentre cet article et notre implémentation. Les structures principales de DSyntR et SSyntR, respectivement DSyntS et SSyntS, sont représentées sous la forme d'arbres de dépendance (cf. figure 5). Ici aussi, pour les mêmes raisons que précédemment, nous ne détaillerons pas les autres structures de DSyntR et SSyntR. Les structures principales des autres niveaux de représentation de la TST sont représentées par des chaînes de caractères.

Dans ces modules de transition, ainsi qu'au sein de certains niveaux de représentation, la TST utilise les concepts clés de paraphrase et de fonctions lexicales (FL).

2.2 Fonctions lexicales

Les FL ont pour but de décrire les phénomènes linguistiques existant au sein des langages. Ce sont en effet des fonctions qui décrivent les relations entre les unités lexicales. Une FL va associer à une unité lexicale un ensemble constitué de toutes les autres unités lexicales alternatives correspondant à la relation qu'elle décrit (voir ci-dessous où des exemples sont donnés). Elles sont utilisées dans la TST dans le dictionnaire explicatif combinatoire (nous ne décrirons pas cet outil, voir Mel'čuk *et al.* (2012, 2013) pour plus de détails) ainsi que pour effectuer l'une des étapes de paraphrase ayant lieu lors de la génération.

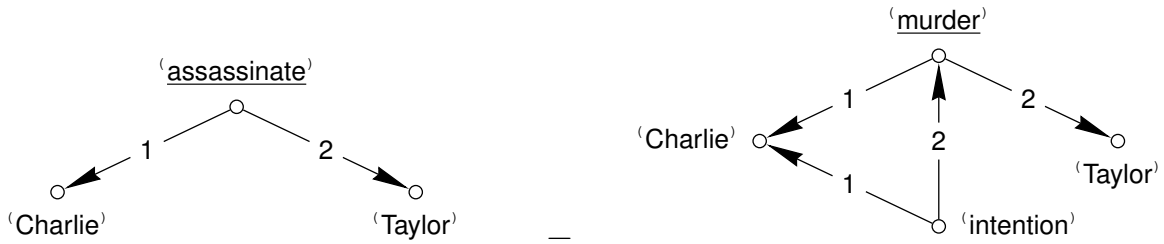
Il existe deux grands types de FL : les FL paradigmatiques et les FL syntagmatiques (Mel'Čuk & Polguère, 2021). Les FL paradigmatiques sont relatives aux propriétés sémantiques des unités lexicales, alors que les FL syntagmatiques expriment les propriétés combinatoires des unités lexicales. Par exemple, `anti` est une FL paradigmatique qui à un lexème associe son contraire : `anti`(CALM) = {UPSET, RESTLESS} (à partir de Mel'čuk *et al.* (2013)), et `causFunc` une FL syntagmatique qui à un lexème L associe un verbe support signifiant *faire en sorte que* L *existe* : `causFunc`(ATTENTION) = DRAW (pour l'expression « *to draw attention* ») (Milićević, 2006). Les FL permettent ainsi d'encoder des phénomènes lexicaux tels que les collocations ou les verbes supports. Plus de détails sur les FL sont donnés dans Mel'Čuk & Polguère (2021). Nous ne détaillerons pas plus cette notion ici.

2.3 Paraphrase

Lors de la génération d'un énoncé à partir d'une représentation sémantique, la TST utilise plusieurs étapes de paraphrase (Iordanskaja *et al.*, 1991) : la paraphrase sémantique (entre SemR et DSyntR), la paraphrase syntaxique profonde (au niveau de DSyntR), la réalisation des FL (entre DSyntR et SSyntR), et au moment de choisir la réalisation de syntaxe de surface (au niveau de SSyntR). Chacune de ces étapes de paraphrase est effectuée par un module de transitions. Nous nous intéressons dans cet article à la paraphrase sémantique.

La figure 2 illustre un exemple de paraphrase au niveau de représentation sémantique de la TST. La paraphrase sémantique est basée sur la composition des sens. En effet, chaque sémantème est défini dans la TST par des sémantèmes plus simples. Ainsi, les sens sont "dépliables" (en une sorte de définition). Un sens est dit composé s'il peut être défini par des sens plus simples. Sinon, il est dit primitif (Mel'čuk *et al.*, 2012). La figure 2 illustre ce phénomène en "dépliant" ⟨assassinate⟩.

Nous nous plaçons ici dans un contexte de génération. Nous nous occuperons principalement de SemR et DSyntR, et nous donnerons en section 5 un exemple détaillé.



(a) « *Charlie assassinates Taylor.* » (b) « *Charlie murders Taylor intentionally.* »

FIGURE 2 – Représentation de deux structures sémantiques illustrant l'équivalence de deux graphes pour la paraphrase sémantique de « *Charlie assassinates Taylor* » (à partir de Mel'čuk *et al.* (2013)).

3 Grammaires catégorielles abstraites (ACG)

Les ACG (de Groote (2001), dont nous reprenons ici les définitions) sont un formalisme grammatical basé sur le λ-calcul. Une ACG est composée de deux langages, liés par un lexique. Le premier langage est appelé *langage abstrait* et correspond à l'ensemble des structures grammaticales abstraites, comme des arbres d'analyse. Le second langage, le *langage objet*, correspond à l'ensemble des réalisations concrètes générées par le langage abstrait, tel que des chaînes de caractère ou des représentations logiques sous forme de graphe. Chacun de ces deux langages est un ensemble de λ-termes obtenu par induction sur une signature.

Définition 1 *Soit A l'ensemble des types atomiques. On note $\mathcal{T}(A)$ l'ensemble des* ***types implicatifs linéaires*** *obtenu par induction sur A :*

— *si $a \in A$ alors $a \in \mathcal{T}(A)$*
— *si $\alpha, \beta \in \mathcal{T}(A)$ alors $(\alpha \to \beta) \in \mathcal{T}(A)$*

Définition 2 *Soit Σ une* ***signature d'ordre supérieur****. Σ est de la forme $\Sigma = \langle A, C, \tau \rangle$, où :*

— *A est un ensemble de types atomiques,*
— *C un ensemble de constantes,*
— *$\tau : C \longrightarrow \mathcal{T}(A)$ une fonction.*

Pour indiquer qu'un λ-terme t est de type s dans la signature Σ, on note $\vdash_\Sigma t : s$, ou encore $t : s$ s'il n'y a pas d'ambiguïté.

On note $\Lambda(\Sigma)$ l'ensemble des λ-termes obtenus en utilisant les constantes de C, les variables, les abstractions et les applications.

Définition 3 *Étant donné deux signatures Σ_1 et Σ_2, un* ***lexique*** *$\mathcal{L}_{12}$ de Σ_1 dans Σ_2 est un couple de morphismes $\langle F, G \rangle$ tel que $F : \tau(A_1) \longrightarrow \tau(A_2)$ et $G : \Lambda(\Sigma_1) \longrightarrow \Lambda(\Sigma_2)$.*

On note alors $\mathcal{L}_{12}(t) = \gamma$ l'interprétation de t par $\mathcal{L}_{12}$ (qui est alors égale à γ), ou encore $t := \gamma$ s'il n'y a pas d'ambiguïté sur le lexique utilisé.

Les signatures et ACG détaillées ici utilisent des λ-termes presque linéaires. Nous ne détaillerons cependant pas cette notion car elle n'a que peu d'intérêt par rapport à ce que nous exposons ici, les variables utilisées n'étant ni effacées ni dupliquées dans les lexiques. Nous notons toutefois λ^o

et λ pour les abstractions respectivement linéaires et non linéaires, et $\rightarrow$ et $\Rightarrow$ pour les types non atomiques respectivement linéaires et non linéaires.

Nous définissons ainsi $\Sigma_{semantic}$ qui correspond au niveau SemR (et permet de représenter les SemS illustrées en figure 2), et qui illustre la définition 2, de la manière suivante :

— $A_{semantic} = \{n, t\}$
— $C_{semantic} = \{true,\ c^s_{intention},\ c^s_{Charlie},\ c^s_{Taylor},\ c^s_{murder},\ R_1,\ R_2\}$
— $\tau_{semantic}$ est donné par la table 2

Constante		Type
$c^s_{intention}$	:	$n \Rightarrow t$
$c^s_{Charlie}$	:	$n \Rightarrow t$
c^s_{Taylor}	:	$n \Rightarrow t$
c^s_{murder}	:	$n \Rightarrow t$
R_1	:	$n \rightarrow t \rightarrow t$
R_2	:	$n \rightarrow t \rightarrow t$
$true$	:	t

TABLE 2 – $\tau_{semantic}$

Pour des raisons de concision, nous ne définirons plus par la suite les ensembles A et C ; ils s'induisent de la table représentant τ. Nous pouvons maintenant définir formellement une ACG et les notions de langages abstrait et objet :

Définition 4 *Une **grammaire catégorielle abstraite** est un quadruplet $\mathcal{G} = \langle \Sigma_1, \Sigma_2, \mathcal{L}, s \rangle$ où :*
— *$\Sigma_1 = \langle A_1, C_1, \tau_1 \rangle$ et $\Sigma_2 = \langle A_2, C_2, \tau_2 \rangle$ sont deux signatures d'ordre supérieur,*
— *$\mathcal{L} = \Sigma_1 \longrightarrow \Sigma_2$ est le lexique,*
— *$s \in \mathcal{T}(A_1)$ est le type distingué de la grammaire.*

Définition 5 *Le **langage abstrait** $\mathcal{A}$ et le **langage objet** $\mathcal{O}$ d'une ACG $\mathcal{G} = \langle \Sigma_1, \Sigma_2, \mathcal{L}, s \rangle$ sont :*
— *$\mathcal{A} = \{t \in \Lambda(\Sigma_1) | \vdash_{\Sigma_1} t : s$ est dérivable$\}$*
— *$\mathcal{O} = \{t \in \Lambda(\Sigma_2) | \exists\, u \in \mathcal{A}(\mathcal{G})$ tel que $t = \mathcal{L}(u)\}$*
Nous utilisons dans cet article la $\beta\eta$-equivalence comme égalité entre les λ-termes

Afin d'illustrer ces concepts, nous introduisons la signature abstraite $\Sigma_{deep\text{-}syntactic}$ (représentée table 3). Nous pouvons maintenant définir le lexique $\mathcal{L}_{sem}$ (représenté table 4), qui formera avec les signatures $\Sigma_{semantic}$ et $\Sigma_{deep\text{-}syntactic}$ une ACG. Cette ACG permettra la première partie de la transition de SemR ($\Sigma_{semantic}$) vers DSyntR ($\Sigma_{dsynt\text{-}tree}$) puisque $\Sigma_{deep\text{-}syntactic}$ correspond à une représentation syntaxique profonde abstraite.

Constante		Type
$c^{ds}_{assassinate}$	:	$G' \rightarrow G \rightarrow G$
$c^{ds}_{intentionally}$	:	$(MOD \rightarrow G' \rightarrow G) \rightarrow G' \rightarrow G$
$c^{ds}_{Charlie}$	:	G
c^{ds}_{Taylor}	:	G
c^{ds}_{murder}	:	$MOD \rightarrow G' \rightarrow G \rightarrow G$
EXP	:	$G \rightarrow G'$

TABLE 3 – $\tau_{deep-syntactic}$

Nous définissons également les notions d'ordre et de complexité d'une ACG. Ces notions sont utilisées dans la table 1 qui décrit le pouvoir expressif des ACG.

$\Sigma_{\textit{deep-syntactic}}$		$\Sigma_{\textit{semantic}}$
G	:=	$n \Rightarrow t$
G'	:=	$n \Rightarrow t$
MOD	:=	t
$c^{ds}_{assassinate}$	:=	$\lambda^{o}\ x\ y.\ \lambda\ e_0.\ (\exists e_x.\ \exists\ e_y.\ (c^{s}_{murder}\ e_0) \wedge (R_1\ e_0\ e_x) \wedge (x\ e_x) \wedge (y\ e_y)$ $\wedge(\exists e_1.\ (c^{s}_{intention}\ e_1) \wedge (R_1\ e_1\ e_x) \wedge (R_2\ e_1\ e_0)))$
$c^{ds}_{intentionally}$	:=	$\lambda^{o}\ pred.\ \lambda^{o}\ x.\ \lambda\ e0.\ \exists\ e_1.\ (c^{s}_{intention}\ e_1) \wedge (R_2\ e_1\ e_0)$ $\wedge(pred\ true\ (\lambda\ e_2.\ (R_1\ e_1\ e_2) \wedge (e_x\ e_2))\ e_0)$
$c^{ds}_{Charlie}$	:=	$\lambda\ e_0.\ (c^{s}_{Charlie}\ e_0)$
c^{ds}_{Taylor}	:=	$\lambda\ e_0.\ (c^{s}_{Taylor}\ e_0)$
c^{ds}_{murder}	:=	$\lambda^{o}\ A.\ \lambda^{o}\ x\ y.\ \lambda\ e_0.\ (\exists\ e_x.\ \exists\ e_y.\ (c^{s}_{murder}\ e0) \wedge (R_1\ e_0\ e_x)$ $\wedge(R_2\ e_0\ e_y) \wedge A \wedge (x\ e_x) \wedge (y\ e_y))$

TABLE 4 – $\mathcal{L}_{sem}$

Définition 6 (Pogodalla, 2017b) *L'**ordre** d'une ACG est le maximum de l'ordre de ses constantes abstraites. L'ordre d'une constante abstraite est l'ordre de son type τ. L'ordre d'un type $\tau \in \mathcal{T}(A)$ est défini par induction :*

- $ordre(\tau) = 1$ *si* $\tau \in A$,
- $ordre(\alpha \to \beta) = max(1 + ordre(\alpha), ordre(\beta))$ *sinon.*

*La **complexité** d'une ACG est le maximum des ordres des réalisations de ses types atomiques. Une ACG d'ordre γ et de complexité η est notée $ACG_{(\gamma,\eta)}$.*

Nous notons dans cet article c^{X}_{L} pour la constante de Σ_X représentant le lexème L . Nous utilisons aussi la notation γ^{X}_{i} pour le λ-terme complexe de $\Lambda(\Sigma_X)$ indexé par i. Ces λ-termes complexes encodent en fait une représentation (dans $\Lambda(\Sigma_X)$) d'une expression, expression que l'on indique grâce à i. Ainsi, nous utiliserons l'indice a pour « *Charlie assassinates Taylor* » et l'indice m pour « *Charlie murders Taylor intentionally* ». Nous notons respectivement s, ds et dt au lieu de $semantic$, $deep - syntactic$ et $dsynt - tree$ dans cette notation.

On définit alors les λ-termes complexes grâce aux tables 2, 3 et 4 :

$$\begin{aligned}\gamma^{s}_{m} &= \lambda^{o}\ A.\ \lambda^{o}\ x\ y.\ \lambda\ e_0.\ (\exists\ e_x.\ \exists\ e_y.\ (c^{s}_{murder}\ e0) \wedge (R_1\ e_0\ e_x) \\ &\qquad \wedge (R_2\ e_0\ e_y) \wedge A \wedge (c^{s}_{Charlie}\ e_x) \wedge (c^{s}_{Taylor}\ e_y)) \qquad (1)\\ \gamma^{ds}_{m} &= (c^{ds}_{intentionally}(\lambda^{o}\ A\ x.\ c^{ds}_{murder}\ A\ x\ c^{ds}_{Taylor}))(EXP\ c^{ds}_{Charlie}) \qquad (2)\\ \gamma^{ds}_{a} &= c^{ds}_{assassinate}\ (EXP\ c^{ds}_{Charlie})\ c^{ds}_{Taylor} \qquad (3)\end{aligned}$$

Une autre propriété des ACG est la transduction : étant donné deux ACG partageant la même signature abstraite, la transduction (cf. figure 3) consiste en la composition de l'analyse (c'est-à-dire une inversion de morphisme, comme $\mathcal{L}_{sem}{}^{-1}$ dans la figure 3) et de l'application (c'est-à-dire une application de morphisme, comme $\mathcal{L}_{dsyntRel}$ dans la figure 3) à l'aide de chacune des ACG. La transduction permet donc de mettre en relation les termes de deux languages objets.

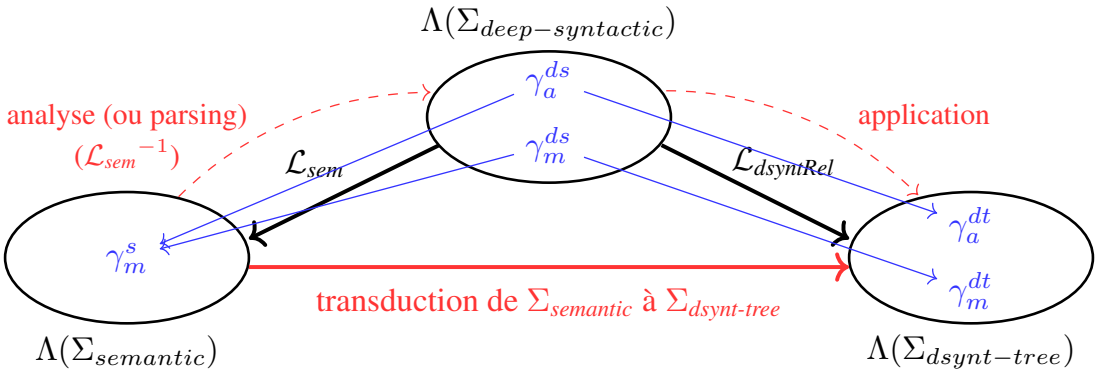


FIGURE 3 – Illustration de la transduction dans le cas de la transition de SemR à DSyntR par inversion de $\mathcal{L}_{sem}$ et sa composition avec $\mathcal{L}_{dsyntRel}$. Le lexique $\mathcal{L}_{sem}$ est tel que $\mathcal{L}_{sem}(\gamma_a^{ds}) = \mathcal{L}_{sem}(\gamma_m^{ds})$

4 Implémentation et résultats obtenus

Nous avons implémenté différentes signatures et lexiques afin de modéliser le fonctionnement du modèle sens-texte de SemR à SSyntR. Cependant, pour des soucis de simplification et cette implémentation se voulant exploratoire, nous n'avons représenté que les principaux niveaux de représentation de la TST, et non pas les structures communicatives ou rhétoriques par exemple (cf. figure 1).

Cette implémentation exploite la transduction (cf. figure 3), c'est-à-dire la composition entre une inversion de morphismes et l'application de morphismes dans les ACG. C'est en effet un moyen de calcul qui va permettre la transition entre les modules (ou les signatures objet dans notre implémentation) pour la transformation des structures. Dans notre implémentation, cette propriété de transduction est cruciale et exploitée afin de réaliser toutes les transformations. En effet, étant donné un λ-terme initial, on obtient par transduction un second λ-terme, sans avoir modifié le premier terme. Nous allons donc utiliser la transduction, qui nous semble particulièrement judicieuse, car la TST est comme une cascade de transformations de structures si l'on regarde comment elle est construite (cf. figure 1). Cette cascade se devine également sur la figure 4 (qui représente l'architecture des ACG dans notre implémentation) dans les zones 1, 2, 4 et 5.

Cependant, la zone 3 (cf. figure 4) forme comme une excroissance. En effet, lors de la génération, bien que la TST ne modifie pas les structures de représentation d'un niveau à l'autre (similairement à la transduction), nous voulons garder les premières structures. Dans les étapes de paraphrase par exemple, il se peut que plusieurs paraphrases différentes soient possibles (cf. table 5 illustrant le nombre de structures syntaxiques profondes obtenues après la paraphrase syntaxique profonde). Nous voulons donc garder une trace de toutes les applications de règles de paraphrases (initiales, intermédiaires et finales), et ce en particulier lors de l'étape de paraphrase syntaxique profonde, ce qui n'est pas propre à la transduction. Autrement dit, la TST fait de la réécriture tout en restant dans un même niveau de représentation, et sans supprimer les structures antérieures du même niveau : elles restent toutes dans l'ensemble qui nous intéresse, et doivent toutes passer à l'étape suivante. Cette étape est particulièrement compliquée à implémenter avec les ACG, car cela forme une boucle. La zone 3 représente cette boucle de paraphrase syntaxique profonde. Il faut, pour chacune des structures syntaxiques profondes, effectuer la transduction de $\Sigma_{dsynt\text{-}tree}$ à $\Sigma_{dsynt\text{-}rule}$, puis de $\Sigma_{dsynt\text{-}rule}$ à $\Sigma_{dsynt\text{-}tree}$, et recommencer jusqu'à ce qu'aucune nouvelle structure ne soit renvoyée. Nous verrons par la suite

que cette boucle est problématique.

La figure 4 illustre les liens entre les différentes ACG implémentées. On retrouve les différents niveaux de représentation de la TST : $\Sigma_{semantic}$ correspond au niveau SemR, $\Sigma_{dsynt\text{-}tree}$ correspond au niveau DSyntR, $\Sigma_{dsynt\text{-}0\text{-}fl}$ est une étape intermédiaire correspondant au niveau DSyntR mais où les FL auront été réalisées, et $\Sigma_{ssynt\text{-}tree}$ correspond au niveau SSyntR. La transduction (cf. figure 3) est exploitée ici, et a lieu :

— entre les signatures $\Sigma_{semantic}$ et $\Sigma_{dsynt\text{-}tree}$: pour faire la transition du niveau sémantique à celui de syntaxe profonde (il s'agit des zones 1 et 2, qui seront détaillées en section 5),
— entre les signatures $\Sigma_{dsynt\text{-}tree}$ et $\Sigma_{dsynt\text{-}rule}$: pour effectuer la paraphrase syntaxique profonde (zone 3),
— entre les signatures $\Sigma_{dsynt\text{-}tree}$ et $\Sigma_{dsynt\text{-}0\text{-}fl}$: pour réaliser les FL (zone 4),
— entre les signatures $\Sigma_{dsynt\text{-}0\text{-}fl}$ et $\Sigma_{ssynt\text{-}tree}$: pour faire la transition entre la syntaxe profonde et la syntaxe de surface (zone 5).

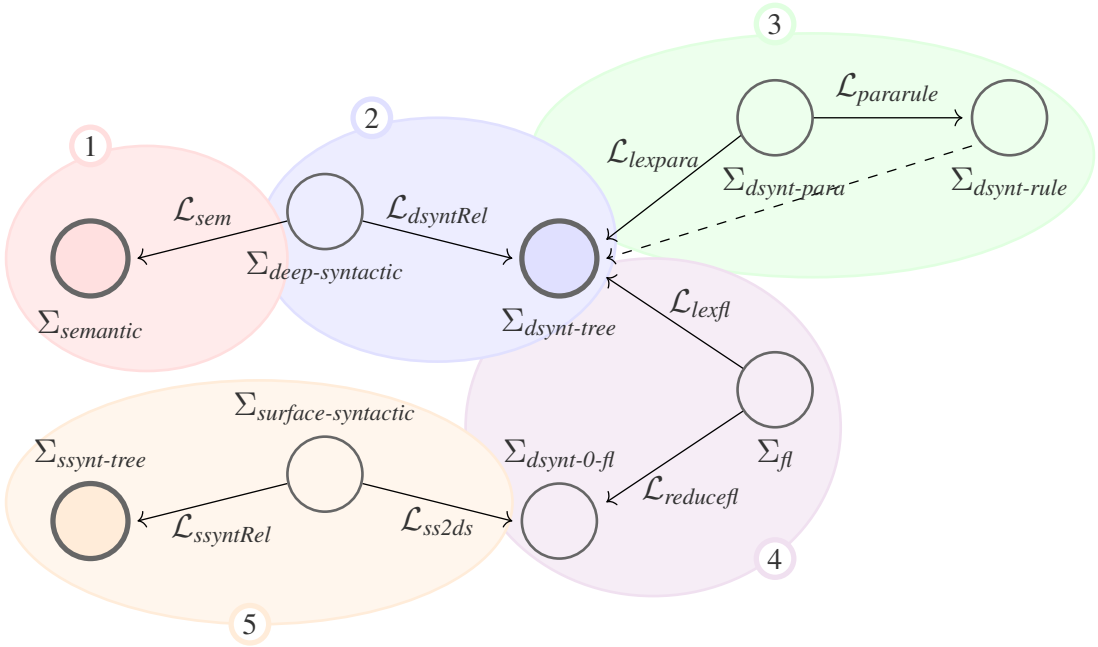


FIGURE 4 – Aperçu de l'architecture des ACG. La zone 1 correspond à l'étape de paraphrase sémantique, la zone 2 à la transition entre la sémantique et la syntaxe profonde, la zone 3 à l'étape de paraphrase syntaxique profonde, la zone 4 à l'étape de réalisation des FL et la zone 5 à la transition entre une représentation de syntaxe profonde où les FL seraient réalisées et la syntaxe de surface.

Cette implémentation a été réalisée avec le logiciel ACGtk (Pogodalla, 2016), et testée sur un ensemble de phrases exemples. Cet ensemble d'exemples est certes restreint, mais couvre plusieurs phénomènes linguistiques, tels que les collocations, le calcul et la manipulation des FL, les équivalences sémantiques ou syntaxiques entre deux énoncés, ou les arguments optionnellement exprimables par exemple. La table 5 illustre le nombre de structures obtenues par étape de génération pour deux graphes sémantiques pris comme exemples, correspondant aux expressions « *Charlie murders Taylor intentionnally* » et « *Alain is calm* ». La première, détaillée dans cet article, illustre la paraphrase sémantique et le traitement des groupes adverbiaux. La seconde, non détaillée ici par manque de place, illustre la paraphrase syntaxique profonde ainsi que le traitement des FL. En outre, les structures d'un niveau de représentation n'étant pas destinées à être réalisées dans le niveau de représentation suivant

dans la TST (parce qu'elles sont incorrectes par exemple) ne posséderont aucun antécédent dans la signature correspondant à ce niveau suivant (cf. table 5 montrant ce phénomène). Si une structure ne doit mener à rien par rapport au formalisme de la TST, alors notre implémentation des ACG est telle que, par transduction, aucun antécédent ne sera trouvé.

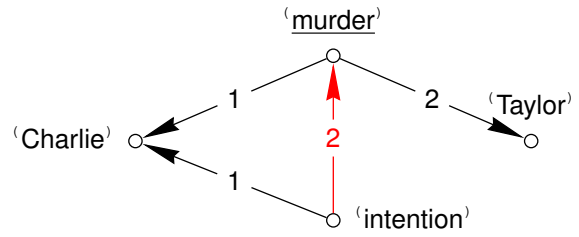
Le code complet contenant les ACG et les exemples est disponible à cette adresse. En effet, les exemples d'ACG que vous pouvez trouver dans cet article ont été simplifiés pour des soucis de clarté et de place.

Expression associée au graphe sémantique	« *Charlie murders Taylor intentionally* »	« *Alain is calm* »
SemR ($\Sigma_{semantic}$)	1	1
DSyntR avant paraphrase syntaxique profonde ($\Sigma_{dsynt\text{-}tree}$)	2	1
dont seront acceptées à l'étape suivante	2	1
DSyntR après paraphrase syntaxique profonde ($\Sigma_{dsynt\text{-}tree}$)	2	6 [1]
dont seront acceptées à l'étape suivante	2	3
DSyntR après réalisation des FL ($\Sigma_{dsynt\text{-}0\text{-}fl}$)	2	5
dont seront acceptées à l'étape suivante	2	3
SSyntR ($\Sigma_{ssynt\text{-}tree}$)	2	3

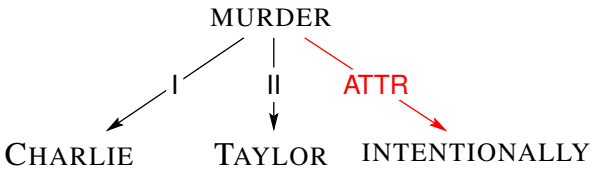
TABLE 5 – Nombre de structures obtenues par étape de la génération pour deux graphes sémantiques initiaux, correspondant aux expressions « *Charlie murders Taylor intentionally* » et « *Alain is calm* ».

De plus, cette implémentation permet, outre les transitions d'un niveau à un autre, le traitement des groupes adverbiaux, qui ont un fonctionnement particulier inspiré du travail sur les TAG de Pogodalla 2017a. Leur traitement n'est en effet pas le même selon que l'on se trouve dans $\Lambda(\Sigma_{semantic})$ ou $\Lambda(\Sigma_{dsynt\text{-}tree})$. Si l'on compare la direction de l'arc reliant le groupe adverbial au reste du graphe sémantique dans SemR et celle de la branche reliant le groupe adverbial au reste de l'arbre de syntaxe profonde dans DSyntR (cf. figure 5), alors on remarque qu'elles sont inversées. En effet, le groupe adverbial est un modificateur, son comportement change alors du comportement "standard" des autres unités lexicales : il pointe sur le sémantème qu'il modifie dans SemS, et est pointé par le lexème qu'il modifie en DSyntS (Mel'čuk *et al.*, 2013, 2015). Nous avons utilisé la même approche que les TAG pour les arbres de dérivation et les arbres dérivés lors des inversions de dépendances (Candito & Kahane, 1998). Notre implémentation permet aussi de ne pas exprimer un argument obligatoire mais optionnellement exprimable d'une SemR (Mel'čuk *et al.*, 2015). Ce dernier point est effectué à l'aide de constantes (EXP, IMP et I_MOD) de $\Sigma_{deep\text{-}syntactic}$, et a été inspiré de Blom *et al.* (2011) mais ne sera pas détaillé ici. Cependant, le traitement des arguments optionnellement exprimables est détaillé dans Cousin (2022), et la section 5 détaille l'exemple de paraphrase sémantique de la figure 2. Cet exemple illustre bien les différentes équivalences pouvant avoir lieu entre deux ACG en implémentant l'équivalence sémantique vue en figure 3, ainsi que le traitement des groupes adverbiaux avec « *intentionally* ».

1. En réalité, plus de structures sont obtenues, mais elles sont incorrectes par construction, elles ne sont donc pas considérées ici (et n'ont dans tous les cas pas d'antécédent dans $\Sigma_{surface\text{-}syntactic}$).



(a) Illustration de la SemS de « *Charlie murders Taylor intentionally.* », encodée par γ_m^s



(b) Illustration de la DSyntS de « *Charlie murders Taylor intentionally* », encodée par γ_m^{dt}

FIGURE 5 – Illustration de l'inversion de dépendance (en rouge) pour le cas de l'adverbe « *intentionally* » dans « *Charlie murders Taylor intentionally* »

5 Un exemple détaillé

Nous donnons dans cette section un exemple détaillé de transduction entre deux signatures et des équivalences pouvant intervenir. Nous nous plaçons dans le cadre de la transition de SemR à DSyntR, soit entre les signatures $\Sigma_{semantic}$ et $\Sigma_{dsynt\text{-}tree}$ (cf. figure 3). Nous n'allons nous intéresser qu'à l'équivalence sémantique dans cet exemple, c'est-à-dire deux termes ayant une DSyntR différente mais qui sont associées au même graphe sémantique, soit à la même SemR.

Prenons l'exemple de la paraphrase sémantique entre les deux expressions suivantes (inspiré de l'exemple de (Mel'čuk *et al.*, 2013), page 207) dont les graphes sémantiques donnés par la TST sont illustrés en figure 2, et dont la paraphrase sémantique est donnée en figure 3 :

(4) a. « *Charlie assassinates Taylor* »

b. « *Charlie murders Taylor intentionally* »

Nous voulons, en notant $\mathcal{T}$ la relation de transduction, avoir les équations suivantes :

$$\mathcal{L}_{sem}(\gamma_m^{ds}) = \gamma_m^s = \mathcal{L}_{sem}(\gamma_a^{ds}) \tag{5}$$

$$\mathcal{L}_{dsyntRel}(\gamma_m^{ds}) = \gamma_m^{dt} \tag{6}$$

$$\mathcal{L}_{dsyntRel}(\gamma_a^{ds}) = \gamma_a^{dt} \tag{7}$$

$$\mathcal{T}(\gamma_m^s,\ \gamma_m^{dt}) \text{ et } \mathcal{T}(\gamma_m^s,\ \gamma_a^{dt}) \tag{8}$$

$$\gamma_m^{dt} \equiv \gamma_a^{dt} \tag{9}$$

Les tables 2, 3, 4, 6 et 7 indiquent les termes des signatures et lexiques que nous allons utiliser dans cette section. Ces trois signatures et deux lexiques permettent la transition de SemR à DSyntR.

Nous avons choisi de modéliser la paraphrase sémantique en exploitant la transduction et les propriétés de β-réduction (entre autres) du λ-calcul : en effet, ces deux expressions ((4a) et (4b)) vont partager la même représentation sémantique (cf. section 2 et figure 2) au niveau de $\Sigma_{semantic}$ mais auront deux arbres de syntaxe profonde différents au niveau de $\Sigma_{dsynt\text{-}tree}$. Dans la modélisation de ce lien de paraphrase, nous utilisions des équivalences sur plusieurs niveaux différents. Cet exemple est ainsi bien adapté pour illustrer les différents niveaux d'équivalence sur lesquels nous pouvons travailler.

Constante		Type
lex_0	:	$l \to T$
lex_2	:	$l \to rel \to T \to rel \to T \to T$
lex_3	:	$l \to rel \to T \to rel \to T \to rel \to T \to T$
A_1	:	rel
A_2	:	rel
$ATTR$	:	rel
$c^{dt}_{assassinate}$	:	l
$c^{dt}_{intentionally}$	:	l
$c^{dt}_{Charlie}$	:	l
c^{dt}_{Taylor}	:	l
c^{dt}_{murder}	:	l

TABLE 6 – $\tau_{dsynt-tree}$

$\Sigma_{deep\text{-}syntactic}$		$\Sigma_{dsynt\text{-}tree}$
G	:=	T
G'	:=	T
MOD	:=	T
EXP	:=	λ^{o} LEX. LEX
$c^{ds}_{assassinate}$	:=	λ^{o} X Y. $lex_2\ c^{dt}_{assassinate}\ A_1$ X A_2 Y
$c^{ds}_{intentionally}$	:=	λ^{o} LEX. λ^{o} X. LEX $(lex_0\ c^{dt}_{intentionally})$X
$c^{ds}_{Charlie}$	:=	$lex_0\ c^{dt}_{Charlie}$
c^{ds}_{Taylor}	:=	$lex_0\ c^{dt}_{Taylor}$
c^{ds}_{murder}	:=	λ^{o} A. λ^{o} X Y. $lex_3\ c^{dt}_{murder}\ A_1$ X A_2 Y $ATTR$ A

TABLE 7 – $\mathcal{L}_{dsyntRel}$

Les termes γ^{dt}_a et γ^{dt}_m de la figure 3 sont construits grâce aux constantes de $\Sigma_{dsynt\text{-}tree}$ (illustrée table 6), et sont respectivement égaux à :

$$\gamma^{dt}_a = lex_2\ c^{dt}_{assassinate}\ A_1\ (lex_0\ c^{dt}_{Charlie})\ A_2\ (lex_0\ c^{dt}_{Taylor}) \tag{10}$$

$$\gamma^{dt}_m = lex_3\ c^{dt}_{murder}\ A_1\ (lex_0\ c^{dt}_{Charlie})\ A_2\ (lex_0\ c^{dt}_{Taylor})\ ATTR\ (lex_0\ c^{dt}_{intentionally}) \tag{11}$$

Ils correspondent aux DSyntR des expressions (4a) et (4b) respectivement.

Ces niveaux d'équivalence en question sont les suivants (cf. figure 3) :

— au sein de $\Sigma_{semantic}$ (cf. table 2) et au niveau du parsing par $\mathcal{L}_{sem}$ (cf. table 4) : les deux phrases (4a) et (4b) ont deux représentations différentes (γ^{ds}_m et γ^{ds}_a) dans $\Sigma_{deep\text{-}syntactic}$ (cf. table 3) mais le même graphe sémantique, qui est lui représenté au niveau de $\Sigma_{semantic}$ (par γ^{s}_m). On obtient donc, par β-réduction, l'égalité (12) modulo β. En effet, nous utilisons la β-équivalence au niveau d'une signature (ici $\Sigma_{semantic}$), et nous utilisons cette équivalence lorsqu'on dit que deux termes ont la même interprétation, car nous raisonnons en utilisant cette β-équivalence. (Les deux termes $\mathcal{L}_{sem}(\gamma^{ds}_m)$ et $\mathcal{L}_{sem}(\gamma^{ds}_a)$ sont bien équivalents à l'ordre des sous-expressions près : la représentation des termes n'est en effet pas unique. C'est d'ailleurs un problème sur lequel nous devons encore travailler.)

$$\mathcal{L}_{sem}(\gamma^{ds}_m) =_{\beta} \mathcal{L}_{sem}(\gamma^{ds}_a) \tag{12}$$

— au niveau de la transduction : les deux arbres de DSyntR et leurs représentations γ_a^{dt} et γ_m^{dt} dans $\Sigma_{dsynt\text{-}tree}$ (cf. table 6) sont equivalents. En effet (cf. figure 3), par application des lexiques $\mathcal{L}_{sem}$ (cf. table 4) et $\mathcal{L}_{dsyntRel}$ (cf. table 7), on a bien les égalités (5), (6) et (7) ci-dessus, puis (8), et donc (9) par transduction.

La transduction entre les signatures $\Sigma_{semantic}$ et $\Sigma_{dsynt\text{-}tree}$ permet donc de modéliser la paraphrase ayant lieu au niveau sémantique.

6 Conclusion

Nous avons montré ici une implémentation possible de la TST avec des ACG. Cette implémentation a eu lieu entre les niveaux SemR et SSyntR du modèle sens-texte. Bien que n'utilisant uniquement les structures principales des trois niveaux de représentation en question, et pas leurs autres structures, comme la structure communicative notamment, cette implémentation a montré, sur un ensemble de phrases exemples, que leurs structures syntaxiques de surface pouvaient être obtenues correctement (cf. table 5). En effet, pour un niveau de représentation donné, les structures incorrectes ne sont pas produites, car elles ne trouvent aucun antécédent par parsing du lexique menant à la signature abstraite associée à ce lexique. Il y a ainsi des moments de la génération où l'on produit des structures et d'autres où l'on filtre les structures obtenues. En outre, si l'on teste la capacité d'analyse de ce modèle, nous obtenons également les graphes sémantiques voulus.

Le modèle réalisé permet d'effectuer la paraphrase sémantique, les transitions entre les niveaux de représentation grâce à la transduction, ainsi que la réalisation des FL, également grâce à la transduction des ACG. De plus, le traitement des arguments optionnellement exprimables et des groupes adverbiaux est aussi permis.

Cependant, ce modèle montre quelques limites. En effet, la paraphrase syntaxique profonde n'est pas optimale. Elle est actuellement possible, et réalisée par l'excroissance dont nous parlions en section 4 (cf. figure 4, zone 3). Mais, elle demande d'être effectuée à la main, en stockant les étapes intermédiaires, et d'effectuer la boucle de transduction jusqu'à ce que plus aucune nouvelle structure ne soit renvoyée. Nous pouvons imaginer un algorithme automatisant ces manipulations, mais ce n'est pas ce que nous recherchons ; cette alternative ne serait pas plus optimale. La transduction montre ici des limites que nous souhaitons dépasser dans nos recherches futures. Aussi, tous les types de paraphrases possibles (Iordanskaja *et al.*, 1991) n'ont pas été exploités et considérés ici : nous voulons continuer dans cette direction par la suite afin de couvrir cela. De plus, nous voulons réussir à ajouter les autres structures de la TST, au moins les structures communicatives des différents niveaux, afin d'avoir une implémentation encore plus conforme à la TST.

Remerciements

Je remercie Sylvain Pogodalla ainsi que les relecteurs pour leurs commentaires et remarques constructives, qui ont permis l'amélioration de cet article. Je tenais aussi à remercier mes deux directeurs de thèse, Philippe de Groote et Sylvain Pogodalla, pour leur encadrement et nos échanges qui m'ont permis d'écrire cet article.

Références

BANARESCU L., BONIAL C., CAI S., GEORGESCU M., GRIFFITT K., HERMJAKOB U., KNIGHT K., KOEHN P., PALMER M. & SCHNEIDER N. (2013). Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, p. 178–186, Sofia, Bulgaria : Association for Computational Linguistics.

BLOM C., DE GROOTE P., WINTER Y. & ZWARTS J. (2011). Implicit Arguments : Event Modification or Option Type Categories ? In M. ALONI, V. KIMMELMAN, F. ROELOFSEN, G. W. SASSOON, K. SCHULZ & M. WESTERA, Éds., *18th Amsterdam Colloquium on Logic, Language and Meaning*, volume 7218 de *Lecture Notes in Computer Science*, p. 240–250, Amsterdam, Netherlands : Springer. DOI : 10.1007/978-3-642-31482-7_25, HAL : hal-00763102.

CANDITO M.-H. & KAHANE S. (1998). Can the TAG derivation tree represent a semantic graph ? an answer in the light of meaning-text theory. In *Proceedings of the Fourth International Workshop on Tree Adjoining Grammars and Related Frameworks (TAG+4)*, p. 21–24, University of Pennsylvania : Institute for Research in Cognitive Science.

COUSIN M. (2022). Génération de texte avec les grammaires catégorielles abstraites et la théorie sens-texte. Mémoire de master, Grenoble INP Ensimag. HAL : hal-03942766.

DANLOS L., MASKHARASHVILI A. & POGODALLA S. (2014). Génération de textes : G-tag revisité avec les grammaires catégorielles abstraites. In B. BIGI, Éd., *Actes de TALN 2014 (Traitement automatique des langues naturelles)*, Marseille : ATALA LPL.

DE GROOTE P. (2001). Towards abstract categorial grammars. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, p. 252–259, Toulouse, France : Association for Computational Linguistics. DOI : 10.3115/1073012.1073045.

IORDANSKAJA L., KITTREDGE R. & POLGUÈRE A. (1991). *Lexical Selection and Paraphrase in a Meaning-Text Generation Model*, In C. L. PARIS, W. R. SWARTOUT & W. C. MANN, Éds., *Natural Language Generation in Artificial Intelligence and Computational Linguistics*, p. 293–312. Springer US : Boston, MA. DOI : 10.1007/978-1-4757-5945-7_11.

JARDINO M., Éd. (2005). *Actes de TALN 2005 (Traitement automatique des langues naturelles)*, Dourdan. ATALA, LIMSI.

KAHANE S. (2005). Structure des représentations logiques, polarisation et sous-spécification. In (Jardino, 2005).

KAHANE S. & LAREAU F. (2005). Grammaire d'unification sens-texte : modularité et polarisation. In (Jardino, 2005).

KANAZAWA M. (2007). Parsing and generation as datalog queries. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, p. 176–183, Prague, Czech Republic : Association for Computational Linguistics.

LAMBREY F. & LAREAU F. (2015). Le traitement des collocations en génération de texte multilingue. In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles. Articles courts*, p. 263–269, Caen, France : ATALA.

LAREAU F. (2007). Vers une formalisation des décompositions sémantiques dans la grammaire d'unification sens-texte. In F. BENAMARA, N. HATOUT, P. MULLER & S. OZDOWSKA, Éds., *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse : ATALA IRIT.

LAREAU F., LAMBREY F., DUBINSKAITE I., GALARRETA-PIQUETTE D. & NEJAT M. (2018). GenDR : A generic deep realizer with complex lexicalization. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan : European Language Resources Association (ELRA).

MEL'ČUK I. & POLGUÈRE A. (2021). Les fonctions lexicales dernier cri. In S. MARENGO, Éd., *La Théorie Sens-Texte. Concepts-clés et applications*, Dixit Grammatica, p. 75–155. L'Harmattan. HAL : hal-03311348.

MEL'ČUK I., MEL'ČUK I., BECK D. & POLGUÈRE A. (2012). *Semantics : From Meaning to Text*, volume 1 de *Semantics : From Meaning to Text*. John Benjamins Publishing Company.

MEL'ČUK I., MEL'ČUK I., BECK D. & POLGUÈRE A. (2013). *Semantics : From Meaning to Text*, volume 2 de *Semantics : From Meaning to Text*. John Benjamins Publishing Company.

MEL'ČUK I., MEL'ČUK I., BECK D. & POLGUÈRE A. (2015). *Semantics : From Meaning to Text*, volume 3 de *Semantics : From Meaning to Text*. John Benjamins Publishing Company.

MILIĆEVIĆ J. (2006). A short guide to the meaning-text linguistic theory. *Journal of Koralex*, **8**, 187–233.

POGODALLA S. (2016). ACGtk : un outil de développement et de test pour les grammaires catégorielles abstraites (ACG TK : a toolkit to develop and test abstract categorial grammars). In *Actes de la conférence conjointe JEP-TALN-RECITAL 2016. volume 5 : Démonstrations*, p. 1–2, Paris, France : AFCP - ATALA.

POGODALLA S. (2017a). A syntax-semantics interface for Tree-Adjoining Grammars through Abstract Categorial Grammars. *Journal of Language Modelling*, **5**(3), 527–605. DOI : 10.15398/jlm.v5i3.193, HAL : hal-01242154.

POGODALLA S. (2017b). Abstract Categorial Grammars as a Model of the Syntax-Semantics Interface for TAG. In *FSMNLP 2017 and TAG+13 conference*, Umeå, Sweden. HAL : hal-01583962.

RANTA A. (2004). Grammatical framework. *J. Funct. Program.*, **14**, 145–189. DOI : 10.1017/S0956796803004738.

WANNER L., BOHNET B., BOUAYAD-AGHA N., LAREAU F. & NICKLASS D. (2010). Marquis : Generation of user-tailored multilingual air quality bulletins. *Applied Artificial Intelligence*, **24**(10), 914–952. DOI : 10.1080/08839514.2010.529258.

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Analyse de la légitimité des start-ups

Asmaa LAGRID[1, 2]
(1) LIS, Aix-Marseille université, 13013 Marseille, France
(2) CERGAM, Aix-Marseille université, 13100 Aix-en-Provence, France
asmaa.lagrid@lis-lab.fr, asmaa.lagrid@univ-amu.fr

RÉSUMÉ

La légitimité est un élément crucial pour la stabilité et la survie des startups en phase de croissance. Ce concept est défini dans la littérature comme étant la perception de l'adéquation d'une organisation à un système social en termes de règles, valeurs, normes et définitions. En d'autres termes, la légitimité des startups repose sur l'alignement des jugements subjectifs avec les jugements objectifs des experts, basés sur les performances des startups. Cette mesure de la subjectivité de la légitimité est très similaire à l'analyse des sentiments financiers réalisée sur les entreprises pour évaluer leur santé financière et prendre des décisions d'investissement. Dans ce travail, nous présentons les travaux sur la légitimité et les avancées de l'analyse des sentiments qui peuvent nous aider à analyser la légitimité. Nous examinons également les similitudes et les différences entre la légitimité et l'analyse des sentiments financiers. Nous présentons une première expérimentation sur les annonces de projets sur une plateforme de crowdfunding, en utilisant le modèle DistilBERT, qui a déjà été largement utilisé pour la classification de texte. En conclusion, nous discutons des perspectives de notre recherche pour mesurer la légitimité des startups.

ABSTRACT

Mesure legitimacy of new ventures

The legitimacy of startups is a crucial factor for their stability and survival during the growth phase. Legitimacy is defined in the literature as "the perception of an organization's adequacy to a social system in terms of rules, values, norms, and definitions." In other words, startup legitimacy is based on the alignment of subjective judgments with objective judgments of experts based on the startup's performance. Measuring legitimacy subjectivity is similar to financial sentiment analysis, which analyzes a company's financial health to make investment decisions. This paper presents the concept of legitimacy and advances in sentiment analysis that can be used to analyze legitimacy. It examines the similarities and differences between legitimacy and financial sentiment analysis. The paper also presents a preliminary experiment on project announcements on a crowdfunding platform, using the DistilBERT model widely used for text classification. The paper concludes by discussing the future prospects of research in measuring the legitimacy of startups.

MOTS-CLÉS : Légitimité, startup, mesure, analyse des sentiments.

KEYWORDS: Legitimacy, startup, measure, sentiment analysis.

Actes des 16e Rencontres Jeunes Chercheurs en RI (RJCRI) et 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL), pages 87–100.

1 Introduction

La légitimité est un concept clé pour les startups innovantes qui cherchent à croître et à devenir plus stables. Selon (Aldrich & Fiol, 1994), l'accès aux ressources et aux marchés dépend du niveau de légitimité de l'organisation ; en particulier les startup qui sont dans leur phase de croissance (Alexiou & Wiggins, 2019). En d'autres termes, une startup qui est perçue comme légitime aura plus de facilité à obtenir du financement, à attirer des clients et à développer son activité.

De nombreuses études ont également montré que la légitimité peut augmenter la probabilité de survie des startups (Deephouse *et al.*, 2017; Tost, 2011; Zimmerman & Zeitz, 2002). Les startups qui sont perçues comme légitimes ont plus de chances de surmonter les défis et les obstacles qui se présentent à elles, et sont mieux préparées à faire face aux incertitudes et aux changements qui peuvent survenir dans leur environnement.

A savoir que la légitimité est définie par (Deephouse *et al.*, 2017) comme "la perception de l'adéquation d'une organisation à un système social en termes de règles, valeurs, normes et définitions." Autrement dit, la légitimité est une question de perception, et elle est évaluée par un ensemble de parties prenantes internes et externes qui émettent des jugements en comparant l'organisation à un certain nombre de critères.

Récemment, une étude a proposé un modèle conceptuel (Schoon, 2022) visant à opérationnaliser le concept de légitimité en identifiant ses critères mesurables à partir de différentes définitions présentes dans la littérature. Selon ce modèle, la légitimité des startups est définie comme une dyade constituée de la startup à évaluer et d'une audience qui formule un jugement sur cette startup. L'audience est représentative d'une population partageant des normes et valeurs sociétales similaires, et les relations entre ces deux éléments sont définies par des attentes. Dans ce cadre, la légitimité est déterminée par l'acceptation de l'audience et la conformité de l'entreprise aux attentes formulées.

De nos jours, les organisations utilisent les médias pour présenter et annoncer leurs activités, tout en permettant aux individus de donner leurs opinions sur ces organisations. Cette relation entre les rapports médiatiques et l'opinion publique (Deephouse *et al.*, 2017) facilite l'analyse de l'opinion publique en utilisant des techniques d'analyse des sentiments. La captation et l'analyse de l'expression de la légitimité dans ces médias possède une certaine proximité avec l'analyse de sentiment. Toutefois, l'analyse de la légimité à certe besoin d'analyser des éléments subjectifs, ce qui rejoint l'analyse de sentiment mais doit aussi tenir compte d'un certain nombre d'éléments factuels objectifs. Nos travaux vont s'appuyer sur les travaux actuels en analyse de sentiment pour caractériser la part subjectif dans l'analyse de la légitmité. Nous utiliserons les techniques de text mining et de NLP, depuis les plus performantes dans l'analyse des sentiments.

Comme nous venons de le mentionner, l'analyse des sentiments dans ce cadre diffère de la méthode traditionnelle où le sentiment représente une opinion sur un produit, un service, une pratique ou une activité d'une startup. Dans notre cas, la polarité de l'opinion reflète le degré d'acceptation du public envers la startup, en se basant sur des **normes**, des **valeurs** et des **logiques institutionnelles**. Bien que le sentiment public soit subjectif, il peut être agrégé et objectivé au niveau collectif en mesurant l'alignement de ces opinions avec les indicateurs de performance de la startup et en analysant l'opinion objective des analystes et experts du domaine.

Ce travail se concentre sur la mesure de la légitimité des startups à partir des médias en répondant aux questions suivantes :

— Comment peut-on mesurer l'acceptation publique d'une startup ?

— Comment peut-on mesurer sa conformité aux attentes ?

Pour répondre à ces questions, cette recherche examine dans un premier temps les travaux existants concernant la légitimité, puis fait un rapide panorama sur les avancées de l'analyse des sentiments pour les adapter à la problématique de l'analyse de la légitimité. Une distinction est faite entre l'analyse de sentiments dans le contexte de la légitimité et l'analyse de sentiments financiers. Enfin, nous présentons les premières expérimentations menées sur l'analyse de la légitimité. Il s'agit de l'analyse de la légitimité d'annonces déposées sur une plateforme de crowdfunding. Pour finir, nous présentons nos perspectives et les prochaines étapes pour répondre à nos questions de recherche.

2 Travaux existants

2.1 La légitimité entrepreneuriale

La légitimité est un concept complexe qui a été initialement proposé dans les travaux de Max Weber (Greenwood & Lawrence, 2005; Díez-Martín *et al.*, 2021) et qui a suscité de nombreux travaux pour clarifier ses dimensions, ses sources d'évaluation, (Deephouse *et al.*, 2008, 2017; Suddaby *et al.*, 2017; Suchman, 1995), les mécanismes et les stratégies qui permettent son acquisition(Fisher, 2020; Suddaby *et al.*, 2017; Suchman, 1995). Cependant, peu d'études empiriques ont traité des mesures de la légitimité (Bitektine *et al.*, 2020; Alexiou & Wiggins, 2019), ce qui a conduit à des critiques sur leur manque de généralisation (Díez-Martín *et al.*, 2021). Schoon dans son article (Schoon, 2022) a distingué les travaux existants selon trois approches pour mesurer la légitimité : (1) une approche basée sur l'évaluation de la perception d'une population envers la startup, (2) une approche basée sur la conformité des activités, services et produits de la startup aux normes, valeurs et croyances sociétales, et (3) une approche basée sur la nécessité de l'existence d'une startup dans l'environnement social et l'absence de questions sur ses activités et services.

Il a été souligné dans l'article (Haack & Sieweke, 2020) qu'il existait deux types de jugements individuels qui n'ont pas été pris en compte lors du développement des instruments de mesure de la légitimité : les jugements du premier ordre, qui représentent des jugements individuels privés sur la légitimité d'une organisation, et qui peuvent être biaisés car ils sont basés sur des expériences subjectives, des préférences et des sentiments. Les jugements du deuxième ordre, quant à eux, représentent des jugements d'une collectivité d'individus et sont considérés comme étant les meilleurs prédicteurs du comportement d'une organisation. Ils sont des jugements qui valident les jugements individuels et sont plus objectifs.

Il en ressort que la légitimité est une **perception collective** généralisée composée de jugements individuels subjectifs (Bitektine, 2011; Tost, 2011) qui sont agrégés et objectivés au niveau collectif (Bitektine & Haack, 2015). Ainsi, le jugement social et la validation collective sont des éléments clés pour mesurer la légitimité d'une entreprise, en particulier les startups. À cet égard, une conceptualisation plus générale a été proposée dans (Schoon, 2022) pour répondre à la question de la façon de mesurer la légitimité des startups. Cela implique de prendre en compte les différentes approches pour mesurer la légitimité, ainsi que les jugements individuels et collectifs, afin d'élaborer des mesures plus généralisables et plus précises de la légitimité des startups. En accord avec le modèle conceptuel de Schoon (Schoon, 2022), notre étude propose une conceptualisation illustrée dans la figure 2, qui se concentre sur l'utilisation des médias traditionnels et sociaux comme source de jugement. Nous mesurons l'opinion publique à travers ces médias pour refléter les attentes des parties prenantes

de la startup envers plusieurs aspects de l'entreprise, tels que l'aspect organisationnel, cognitive, associatif, pragmatique, réglementaire, moral, et responsabilité envers la société. Ces attentes reflètent la santé globale de l'entreprise, son savoir-faire, son expertise, son adéquation aux normes, valeurs et définitions sociétales, ainsi que sa position dans l'environnement social. Cette opinion publique, qui est un jugement subjectif, est validée collectivement par les experts à travers l'analyse d'enquêtes et de rapports, ainsi que par l'analyse des indicateurs de performances de la startup (KPIs) et son capital social, qui représente l'ensemble des ressources mobilisables par la startup en analysant le contenu des bases de données publiques et les rapports annuels.

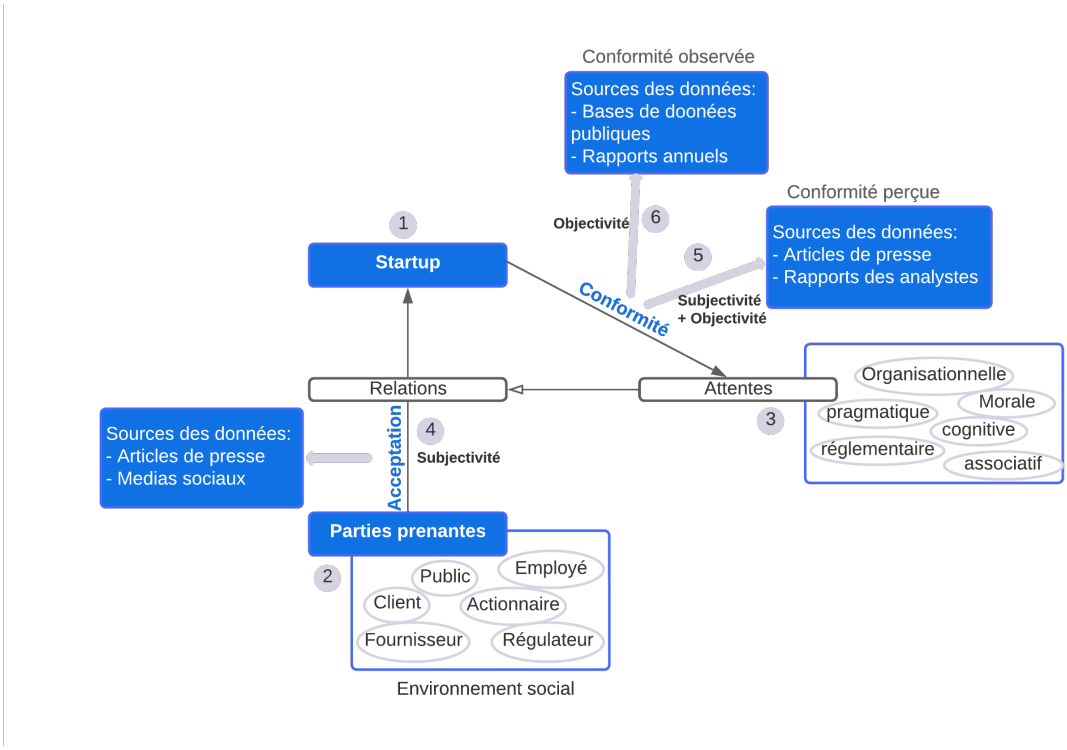


FIGURE 1 – Conceptualisation de la légitimité

2.2 Analyse des sentiments

Comme nous venons de le voir, la légitimité est une perception collective composée de jugements individuels subjectif. Les travaux se rapprochant le plus actuellement de cette définition est l'analyse de sentiment. Le champ de recherche de l'analyse des sentiments (ou opinion mining) s'attache à l'examen des attitudes, des opinions et des émotions exprimées par les individus dans les textes en langage naturel (Liu, 2012), en vue d'extraire des informations subjectives relatives à des entités telles que des produits, des services ou des organisations économiques. Cette analyse peut être menée à différents niveaux (Hu & Liu, 2004), notamment au niveau du document, de la phrase ou d'un sentiment exprimé envers une entité cible.

Dans ce contexte, les tâches d'analyse des sentiments sont classées selon plusieurs catégories, notamment le coarse-grain, le fine-grain, le cross-domain et le cross-lingual. Le coarse-grain consiste à classifier les sentiments en positif, négatif ou neutre au niveau du document et de la phrase. Cette

analyse peut être utilisée pour d'autres tâches telles que l'analyse du sarcasme, l'analyse du langage offensant et l'analyse du discours haineux (Abdullah & Ahmet, 2022). Le fine-grain comprend le targeted-sentiment-analysis (TSA), qui permet d'identifier la polarité d'un ensemble d'entités dans un texte, et l'aspect-based-sentiment-analysis (ABSA), qui permet d'évaluer la polarité de plusieurs aspects (caractéristiques) d'une entité. L'ABSA peut également être utilisé pour l'extraction de termes d'aspect et la catégorisation de la polarité d'aspects. Le cross-domain consiste à classer les sentiments dans des domaines différents de ceux sur lesquels le modèle a été entraîné, tandis que le cross-lingual consiste à classer les sentiments dans des langues différentes de celles sur lesquelles le modèle a été entraîné.

Les dernières avancées en matière d'architecture de deep learning en NLP ont été utilisées pour améliorer les performances de l'analyse des sentiments. Selon l'enquête (Abdullah & Ahmet, 2022) en 2022, les transformers (Vaswani *et al.*, 2017) ont une grande capacité de transfert de connaissances syntaxiques et sémantiques par rapport aux architectures CNN/RNN/Attention. Ils sont également plus simples à ajuster et nécessitent moins d'échantillons pour atteindre les performances de pointe pour toutes les tâches d'analyse des sentiments. De plus, les transformers sont au centre de l'intérêt de la communauté des chercheurs en NLP.

En termes d'applications, les techniques d'analyse des sentiments sont également utilisées dans divers domaines, tels que le marketing, la surveillance de la réputation en ligne, l'analyse des critiques des produits et des films, et l'analyse des données financières. C'est ce dernier domaine qui est le plus proche de l'analyse de la légitimité.

2.3 L'analyse des sentiments financiers

L'analyse des sentiments financiers consiste à déterminer le sentiment et l'opinion dans les textes financiers, elle se diffère de l'analyse des sentiments traditionnelle en raison de plusieurs facteurs spécifiques au domaine financier. Tout d'abord, le concept de sentiment dans les textes financiers représente "les attentes des acteurs du marché" (Brown & Cliff, 2004). Plus précisément, le concept de sentiment représente les opinions des investisseurs ainsi que de leurs croyances (Kearney & Liu, 2014) pessimistes ou optimistes (Baker & Wurgler, 2006) vis-à-vis du marché boursier ou des actions individuelles. Par exemple une polarité positive représente un avis optimiste quant aux perspectives d'avenir d'une entreprise, ce qui peut encourager les investisseurs à acheter des actions, tandis qu'une polarité négative indique que les investisseurs peuvent être plus susceptibles de vendre des actions (Sohangir *et al.*, 2018).

De plus, les textes financiers contiennent des sentiments explicites et implicites (Van de Kauter *et al.*, 2015), et leur structure linguistique est souvent pauvre (Zhang *et al.*, 2018), en raison des termes techniques employés. Il est important de noter également que le même mot dans un texte financier peut représenter plusieurs sentiments selon différentes perspectives (Man *et al.*, 2019).

Dans ce contexte, le sentiment des investisseurs a un impact significatif sur la dynamique du marché (Brown & Cliff, 2004; Baker & Wurgler, 2006; Kearney & Liu, 2014), ce qui rend la mesure des sentiments des investisseurs essentielle pour la prévision sur le marché, car le sentiment du marché reflète le sentiment des investisseurs dans leurs comportements d'investissement (Sohangir *et al.*, 2018). Ce sentiment peut être mesuré par des enquêtes, des textes financiers (Kearney & Liu, 2014) (news financier, communiqué des entreprises, médias sociaux), des indices du marché et les rapports annuels.

2.3.1 Sources des données

L'analyse des sentiments financiers (FSA) est réalisée en utilisant des données de sources diverses, incluant :

- **Les news financiers** : Les news financièrs sont l'une des sources les plus courantes utilisées pour l'analyse des sentiments financiers. Ces articles de presse peuvent être extraites de différents canaux de communication tels que les sites web financiers, les journaux, les magazines spécialisés.
- **Les communications des entreprises** : Les discours et les rapports annuels des entreprises sont également une source importante pour l'analyse des sentiments financiers. Ces communications peuvent donner des indications sur la performance de l'entreprise, ses stratégies, ses projets futurs.
- **Les médias sociaux** : Les plateformes de médias sociaux telles que Twitter et StockTwits sont de plus en plus utilisées pour l'analyse des sentiments financiers. Les postes et micoblogs publiés sur ces plateformes peuvent donner des informations sur l'opinion des investisseurs, les rumeurs, les nouvelles importantes.
- **Les rapports des investisseurs et les enquêtes sur leurs opinions** : Les rapports des investisseurs, tels que les rapports annuels et trimestriels, peuvent fournir des informations sur la performance de l'entreprise, les indicateurs financiers, les projets futurs, a titre d'exemple AAII, UMSC et Sentix, NYSE.

La collecte de données pour l'analyse de sentiments financiers peut être difficile. Les données financières sont souvent privées et confidentielles, ce qui limite l'accès à ces données pour les chercheurs et les scientifiques des données.

L'annotation de textes financiers pour l'analyse de sentiments est également une tâche complexe qui nécessite une expertise dans le domaine financier. Les textes financiers sont souvent très techniques, comportent un langage spécialisé et sont remplis de jargon financier. Les nuances de langage et de tonalité dans ces textes sont également difficiles à saisir, ce qui rend l'annotation encore plus difficile. Pour cette raison plusieurs datasets annotés sont disponibles pour l'analyse des sentiments financiers, incluant :

- **FiQA Task 1** (Maia *et al.*, 2018) : Un dataset qui contient 529 titres de presse financière (436 examples pour l'entrainement et 93 examples pour le test) et 774 microblogs financiers (675 examples pour l'entrainement et 99 examples pour le test) annotés avec plusieurs aspects relatifs à l'entreprise (Réputation, Communication de l'entreprise, Statut, état Financier, Réglementaire, Vente, M&A, Etat légale, Risques, rumeurs, stratégies), la bourse (IPO, analyse fondamentale, analyse technique, Prix des actions), l'économie (Banque central, Trad) et le marché (Liquidité, volatilité, état du marché). L'annotation a été faite selon un echelle de fine-grained entre -1 et 1.
- **PhraseBank** (Malo *et al.*, 2014) : Un dataset qui comprend 4845 phrases extraites des titres des news en anglais sur toutes les entreprises dans la liste de OMX Helsinki, annotées avec des polarités (positive, négative, neutre). Les articles de presse étaient récupéré a partir de la base de données LexisNexis, ensuite annoté par 16 experts en économie et finance.
- **SemEval 2017 Task 5** (Cortis *et al.*, 2017) : Un dataset de 2836 entrées, annotées par des experts en finance, contenant des titres de news comme Yahoo Finance et des microblogs, se concentrant sur les événements du marché boursier et les évaluations des investisseurs et des traders, échangés via la plateforme de microblogging StockTwits, aussi certaines discussions sur le marché boursier ont également lieu sur la plateforme Twitter marqués par des cashtags (des mots clés précédés par des symboles du marché boursier).

— **SSIX Corpora** (Gaillat *et al.*, 2018) : Ce dataset contient 2886 messages avec des opinions ciblés liés au marché boursier extraits de deux plateformes de microblogging financières, StockTwits et Twitter. Le corpus a été annoté par des experts dans différentes langues, notamment l'anglais, l'espagnol et l'allemand, qui ont évalué la polarité des messages sur une échelle continue.
— **FinLin** (Daudert, 2022) : Un dataset récent, constitué de 3 811 entrées. Ces données ont été extraites de tweets StockTwits, d'articles de presse, de rapports d'entreprise et de rapports d'investisseurs, portant sur plusieurs entités issues de l'industrie automobile et couvrant une période de 3 mois. Les entrées de FinLin ont été annotées avec des scores de sentiment dans la plage de -1,0 à 1,0 et avec des scores de pertinence dans la plage de 0,0 à 1,0.

En examinant les sources de données utilisées dans l'analyse des sentiments financiers et dans l'évaluation de la légitimité des entreprises, des similarités ont été observées. Dans notre étude, nous allons également utiliser des sources de données telles que les communications d'entreprise pour extraire des indicateurs de performance et des stratégies, ainsi que des rapports annuels et trimestriels. Nous pouvons également utiliser la partie du corpus FIQA Task 1 qui concerne les entreprises. Nous examinerons dans les section à venir, les points de similarités et les différences entre l'évaluation de la légitimité des start-up et l'évaluation des sentiments financiers

2.4 Similitudes et différences

La mesure de la légitimité et l'analyse des sentiments financiers présentent plusieurs similitudes. Tout d'abord, ces deux méthodes utilisent des sources de données similaires pour évaluer la santé d'une entreprise. Ces sources de données incluent les articles de presse, les médias sociaux, les communications des entreprises et les rapports financiers. En effet, l'analyse des sentiments financiers analyse ces sources pour évaluer la santé financière de l'entreprise d'un point de vue des investisseurs, tandis que la légitimité les analyse pour identifier les pratiques de l'entreprise et ses engagements sociétaux pour aider l'entreprise à concevoir des stratégies efficaces afin d'assurer sa croissance et sa survie.

En outre, la mesure de la légitimité et l'analyse des sentiments financiers partagent des techniques d'extraction d'informations similaires. Par exemple, l'analyse des sentiments financiers vise à extraire des informations qui aident les investisseurs à prendre des décisions d'investissement, telles que le chiffre d'affaires de l'entreprise, le bénéfice par action, le retour sur investissement et la capacité de gestion des dettes. De même, la légitimité utilise ces mêmes indicateurs de performance clés (KPI) pour évaluer la transparence et la crédibilité de l'entreprise sur plusieurs dimensions réglementaires, de savoir-faire, d'expertise et d'identité, de morale, d'impact écologique et sociétal (RSE), ainsi que la santé de l'entreprise d'un point de vue organisationnel.

Cependant, la mesure de la légitimité et l'analyse des sentiments financiers présentent également des différences notables. L'analyse des sentiments financiers se concentre sur les performances financières de l'entreprise, alors que la légitimité se concentre sur des KPI plus larges qui incluent également des pratiques éthiques, la responsabilité sociale, le capital social et la transparence. Ces aspects peuvent avoir un impact important sur la perception de l'entreprise par ses parties prenantes.

En outre, il est important de souligner que les sentiments analysés en Financial Sentiment Analysis (FSA) et dans la légitimité ne sont pas les mêmes et peuvent refléter des attentes différentes. En effet, les acteurs du marché ont souvent des attentes financières et cherchent à maximiser leur rendement financier, tandis que les parties prenantes (clients, fournisseurs, employés, public) ont des attentes

plus larges et diverses telles que la responsabilité sociale, l'éthique et l'impact environnemental. Par exemple, les parties prenantes peuvent attendre que l'entreprise prenne des mesures pour réduire son impact environnemental ou pour améliorer les conditions de travail de ses employés, ce qui peut être différent de ce que les investisseurs attendent pour maximiser leur rendement financier.

Ces attentes différentes peuvent conduire à des interprétations différentes de la polarité des phrases dans les analyses de sentiment. Par exemple, un investisseur pourrait considérer positif le fait qu'une entreprise réduise ses coûts, tandis qu'un client pourrait considérer cela comme un signe de mauvaise qualité ou de manque d'engagement envers la satisfaction du client. Ainsi, il est important de tenir compte de ces attentes différentes lors de l'analyse des sentiments financiers et de la légitimité.

Les entreprises sont de plus en plus conscientes de l'importance des attentes de leurs parties prenantes en matière de légitimité. Cela est dû à l'impact que ces attentes peuvent avoir sur l'image et la survie de l'entreprise à long terme. Les parties prenantes jouent un rôle essentiel dans l'identification des pratiques éthiques et responsables d'une entreprise, qui sont cruciales pour sa crédibilité et sa transparence. L'attention des régulateurs envers la conformité des entreprises aux normes sociales et environnementales ne cesse de croître. Cette tendance renforce davantage l'importance de la légitimité pour les entreprises.

En somme, comme illustre le tableau 1 la mesure de la légitimité et l'analyse des sentiments financiers ont des similitudes importantes dans leur approche et leur utilisation de sources de données et de techniques d'extraction d'informations. Cependant, il est important de reconnaître leurs différences, notamment en ce qui concerne les aspects évalués et les attentes des parties prenantes. Ce qui nécessite le développement des modèles plus spécifique a ce domaine.

2.5 Les avancées de l'analyse des sentiments

Les approches pour l'analyse des sentiments financiers comprennent les méthodes basées sur les lexiques, les méthodes basées sur l'apprentissage automatique (ML), l'apprentissage profond, les approches hybrides combinant les trois et les modèles de langage pré-entraînés.

Les premiers dictionnaires utilisés étaient General Inquirer (GI) (Stone *et al.*, 1966) et Diction, mais ils n'étaient pas suffisamment spécialisés pour capturer les nuances des textes financiers. Des dictionnaires spécifiques au domaine financier, tels que le Loughran-McDonald Financial Sentiment Dictionary (LMFSD) (Loughran & McDonald, 2011), ont été développés pour remédier à ce problème (Mishev *et al.*, 2020; Man *et al.*, 2019).

Les approches basées sur les dictionnaires ne nécessitent pas de données d'entraînement, mais ne peuvent pas capturer toutes les informations critiques dans les textes financiers. Pour améliorer la précision, plusieurs approches d'apprentissage automatique classiques, telles que SVM, Naïve Bayes, les arbres de décision, la régression linéaire, la régression LASSO et Ridge, ont été appliquées sur des données des tweets et des news. Les études ont montré que ces modèles d'apprentissage automatique surpassent les approches basées sur les dictionnaires en termes de performance. Ainsi les modèles hydrique (ML, lexique) et (DL, lexique) ont montré plus de précision en comparent avec tous ces approches appliqué seule (Cortis *et al.*, 2017).

Ces modèles basés sur les n-grammes de bas niveau ne sont pas capables de capturer les caractéristiques complexes des phrases en ne prenant pas en compte l'ordre des mots. Les modèles de deep learning utilisant des techniques d'embedding de mots ont donc été développés pour remédier à cette

	L'analyse des sentiments financiers	La mesure de légitimité
Les sources des données	— Article de presse — Communications des entreprises — Médias sociaux — Rapports des investisseurs — Rapports annuels — Base de données publiques	— Article de presse — Communications des entreprises — Médias sociaux — Rapports annuels — Base de données publiques
Aspects	— Entreprise : réputation, communications, statut, santé financiere, reglementaire, vente, M&A, légale, risques, rumeurs — Bourse : IPO, analyse fondamentale, analyse technique — Économie : banque centrale, trade — Marché : liquidité, volatilité, état du marché	Réglementaire, associatif, RSE, pragmatique, identitaire, organisationnel, cognitive, savoir-faire, expertise, capital social
Sentiment	Attentes des acteurs du marché	Attentes des parties prenantes
Impact	Mouvement et tendance du marché	Croissance et survie de la startup

TABLE 1 – tab : similitudes et différence entre la légitimité et ASF

limitation, parfois combinés avec des approches basées sur les dictionnaires. Les premiers modèles d'embedding de mots tels que Word2Vec et GloVe ont été très influents pour le traitement du langage naturel, mais leurs limites en termes de représentation sémantique des mots et de capacité à capturer les nuances du langage ont conduit au développement de nouveaux modèles plus performants tels que ELMo et GPT, qui peuvent capturer les relations contextuelles entre les mots.

Le modèle BERT a été développé avec l'architecture transformer et le mécanisme d'attention, et diffère des modèles précédents en prenant en compte le contexte autour de chaque mot. En utilisant des modèles de langage masqués, BERT est capable de comprendre comment chaque mot est utilisé dans une phrase et de créer une représentation contextuelle plus précise. BERT existe en deux versions, BERT-base avec 110 millions de paramètres et BERT-large avec 340 millions de paramètres, toutes deux entraînées sur Wikipedia et BookCorpus en anglais.

FinBERT est une version de BERT pré-entraînée sur un corpus de 1,8 million d'articles financiers. Cette version a permis d'améliorer la précision de 15% dans le domaine financier.

En vue de résoudre les limites associées à l'utilisation de grandes représentations de langage naturel pré-entraînées, qui exigent une mémoire importante et des temps d'entraînement plus longs, des modèles ont été proposés tels que ALBERT, RoBERTa, DistilBERT qui optimisent BERT. Ces modèles surpassent BERT dans plusieurs tâches, notamment la classification de texte.

D'après l'enquête de (Mishev *et al.*, 2020), les modèles susmentionnés ont montré de bonnes performances en matière de classification de textes volumineux et peuvent être efficaces pour les applications d'analyse des sentiments financiers, telles que la prédiction des fluctuations du marché boursier, la prévision des risques financiers et la gestion de portefeuilles.

Nous allons maintenant présenter la première expérimentation menée sur les données relatives aux descriptions de projets de crowdfunding, en utilisant le modèle DistilBERT qui a déjà été employé pour l'analyse des sentiments financiers.

3 Expérimentation

3.1 Corpus

Le corpus utilisé dans cette expérimentation est constitué de descriptions de 50 projets de startups en France, extraites de la base de données de crowdfunding Kickstarter [1]. Ces projets ont été sélectionnés en fonction d'un objectif de financement supérieur à 5000 euros, et les 50 projets les plus récents ont été choisis. Les textes des projets ont ensuite été découpés en phrases et annotés manuellement par quatre experts en management, business et entrepreneuriat, selon des mécanismes de légitimité organisationnel, identitaire et associatif. Ces experts ont travaillé indépendamment les uns des autres, ensuite les annotations de chaque expert ont été comparé pour aboutir à une annotation finale. Le corpus se compose donc de 1672 phrases annotées en anglais, dont 1454 sont négatives et 218 positives. Les phrases annotées reflètent la présence ou l'absence de ces mécanismes dans les descriptions de projets.

	Phrases
Négative	1453
Positive	218
Total	1672

TABLE 2 – Un tableau

3.2 Pré-entraînement

Afin de traiter le problème des données non équilibrées dans le corpus, nous avons sélectionné aléatoirement le même nombre de phrases pour les deux classes, ce qui a équilibré les deux classes avec 218 phrases négatives et 218 phrases positives. Ensuite, nous avons effectué un prétraitement de nettoyage des données en supprimant les emojis, les adresses mails et les liens, et nous avons gardé uniquement les mots en caractères alphabétiques. Nous avons également converti tous les mots en minuscules et effectué une lemmatisation. Par la suite, nous avons divisé le dataset en un ensemble d'entraînement (80%) et un ensemble de test (20%).

1. https ://www.kickstarter.com/

3.3 Modèle

Dans le cadre de notre étude, nous avons sélectionné le modèle DistilBert pour entraîner nos données. Ce modèle, une version optimisée de Bert, utilise une architecture de réseau de neurones pré-entraînée pour le traitement de langage naturel. Il présente des avantages significatifs tels qu'une réduction de 40% en termes de mémoire et de temps d'entraînement par rapport à Bert, ainsi qu'une vitesse accrue de 60%. Pour notre classification binaire, nous avons choisi la version distil-bert-uncased. Ce choix s'appuie sur les performances avérées du modèle dans la classification de texte, ainsi que sur son entraînement sur des données provenant de BookCorpus et de Wikipédia, ce qui renforce sa pertinence pour l'analyse de notre corpus constitué des descriptions de projets collectées à partir de la plateforme Kickstarter. Nous avons entraîné le modèle sur quatre époques afin d'améliorer sa précision et sa fiabilité.

3.4 Évaluation

Dans un premier temps, nous avons entraîné le modèle DistilBert sur l'ensemble de données complet, ce qui nous a permis d'obtenir une précision de 97% sur les données d'entraînement et de 88% sur les données de test. Par la suite, nous avons entraîné notre modèle sur un ensemble de données équilibré, ce qui a donné lieu à une précision de 98% sur les données d'entraînement et de 79% sur les données de test. Afin de vérifier la présence de surapprentissage dans notre modèle, nous avons prévu une validation manuelle par des experts. Nous envisageons également de résoudre le problème des données non-équilibrées en appliquant des algorithmes de suréchantillonnage et de collecter davantage de données pour éviter tout risque de surapprentissage.

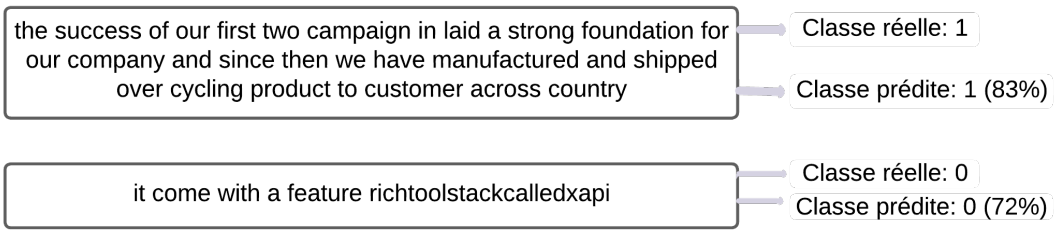


FIGURE 2 – Exemple d'une phrase négative et une autre phrase positive prédite avec le modèle DistilBert sur les données équilibrés

Précision	Dataset non-equilibré	Dataset équilibré
Précision d'entrainement	97%	98%
Précision de test	88%	79%

TABLE 3 – Un tableau

4 Conclusion et perspectives

En somme, cette étude a examiné les travaux antérieurs sur la légitimité et l'analyse des sentiments financiers dans le but d'identifier des critères mesurables pour évaluer la légitimité et opérationnaliser ce concept complexe. Nous avons également étudié les techniques courantes de collecte et d'annotation de données en matière d'analyse des sentiments financiers et nous allons inspirés des travaux de (Garkavenko *et al.*, 2022) pour identifier plusieurs caractéristiques des startups pouvant prédire leur capacité à obtenir des financements à court et à long terme, permettant ainsi leur survie et leur croissance. D'après nos résultats, il ressort que le modèle DistilBert s'avère adéquat pour la classification binaire de notre corpus de données. De plus, nous avons constaté une amélioration de la précision en adoptant un ensemble de données équilibré. Cependant, afin de confirmer la présence ou l'absence de surapprentissage dans notre modèle, une validation manuelle effectuée par des experts est requise. Pour nos perspectives futures, nous prévoyons d'explorer d'autres modèles avancés tels que BERT, GPT et FinBert pour améliorer nos résultats. Nous allons également poursuivre la collecte de données sur la légitimité afin de mieux comprendre les différentes catégories et les différents aspects à analyser. En outre, nous avons l'intention d'étudier l'utilisation de techniques de prétraitement de données telles que l'augmentation de données et l'élargissement des classes de légitimité pour améliorer encore la performance de notre modèle.

Références

ABDULLAH T. & AHMET A. (2022). Deep learning in sentiment analysis : Recent architectures. *ACM Computing Surveys*, **55**(8), 1–37.

ALDRICH H. E. & FIOL C. M. (1994). Fools rush in ? the institutional context of industry creation. *Academy of management review*, **19**(4), 645–670.

ALEXIOU K. & WIGGINS J. (2019). Measuring individual legitimacy perceptions : Scale development and validation. *Strategic Organization*, **17**(4), 470–496.

BAKER M. & WURGLER J. (2006). Investor sentiment and the cross-section of stock returns. *The journal of Finance*, **61**(4), 1645–1680.

BITEKTINE A. (2011). Toward a theory of social judgments of organizations : The case of legitimacy, reputation, and status. *Academy of management review*, **36**(1), 151–179.

BITEKTINE A. & HAACK P. (2015). The "macro" and the "micro" of legitimacy : Toward a multilevel theory of the legitimacy process. *Academy of management review*, **40**(1), 49–75.

BITEKTINE A., HILL K., SONG F. & VANDENBERGHE C. (2020). Organizational legitimacy, reputation, and status : Insights from micro-level measurement. *Academy of Management Discoveries*, **6**(1), 107–136.

BROWN G. W. & CLIFF M. T. (2004). Investor sentiment and the near-term stock market. *Journal of empirical finance*, **11**(1), 1–27.

CORTIS K., FREITAS A., DAUDERT T., HUERLIMANN M., ZARROUK M., HANDSCHUH S. & DAVIS B. (2017). Semeval-2017 task 5 : Fine-grained sentiment analysis on financial microblogs and news. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, p. 519–535.

DAUDERT T. (2022). A multi-source entity-level sentiment corpus for the financial domain : the finlin corpus. *Language Resources and Evaluation*, **56**(1), 333–356.

DEEPHOUSE D. L., BUNDY J., TOST L. P., SUCHMAN M. C. *et al.* (2017). Organizational legitimacy : Six key questions. *The SAGE handbook of organizational institutionalism*, **4**(2), 27–54.

DEEPHOUSE D. L., SUCHMAN M. *et al.* (2008). Legitimacy in organizational institutionalism. *The Sage handbook of organizational institutionalism*, **49**, 77.

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

DÍEZ-MARTÍN F., BLANCO-GONZÁLEZ A. & PRADO-ROMÁN C. (2021). The intellectual structure of organizational legitimacy research : a co-citation analysis in business journals. *Review of Managerial Science*, **15**(4), 1007–1043.

FISHER G. (2020). The complexities of new venture legitimacy. *Organization Theory*, **1**(2), 2631787720913881.

GAILLAT T., ZARROUK M., FREITAS A. & DAVIS B. (2018). The ssix corpora : Three gold standard corpora for sentiment analysis in english, spanish and german financial microblogs. In *LREC : Language Resources and Evaluation Conference*, p. 2671–2675 : European Languages Resources Association (ELRA).

GARKAVENKO M., GAUSSIER E., MIRISAEE H., LAGNIER C. & GUERRAZ A. (2022). Where do you want to invest ? predicting startup funding from freely, publicly available web information. *arXiv preprint arXiv :2204.06479*.

GREENWOOD R. & LAWRENCE T. B. (2005). The iron cage in the information age : The legacy and relevance of max weber for organization studies. editorial.

HAACK P. & SIEWEKE J. (2020). Advancing the measurement of organizational legitimacy, reputation, and status : First-order judgments vs second-order judgments—commentary on "organizational legitimacy, reputation and status : Insights from micro-level management". *Academy of Management Discoveries*, **6**(1), 153–158.

HU M. & LIU B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 168–177.

KEARNEY C. & LIU S. (2014). Textual sentiment in finance : A survey of methods and models. *International Review of Financial Analysis*, **33**, 171–185.

LIU B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, **5**(1), 1–167.

LOUGHRAN T. & MCDONALD B. (2011). When is a liability not a liability ? textual analysis, dictionaries, and 10-ks. *The Journal of finance*, **66**(1), 35–65.

MAIA M., HANDSCHUH S., FREITAS A., DAVIS B., MCDERMOTT R., ZARROUK M. & BALAHUR A. (2018). Www'18 open challenge : financial opinion mining and question answering. In *Companion proceedings of the the web conference 2018*, p. 1941–1942.

MALO P., SINHA A., KORHONEN P., WALLENIUS J. & TAKALA P. (2014). Good debt or bad debt : Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, **65**(4), 782–796.

MAN X., LUO T. & LIN J. (2019). Financial sentiment analysis (fsa) : A survey. In *2019 IEEE International Conference on Industrial Cyber Physical Systems (ICPS)*, p. 617–622 : IEEE.

MISHEV K., GJORGJEVIKJ A., VODENSKA I., CHITKUSHEV L. T. & TRAJANOV D. (2020). Evaluation of sentiment analysis in finance : from lexicons to transformers. *IEEE access*, **8**, 131662–131682.

SCHOON E. W. (2022). Operationalizing legitimacy. *American Sociological Review*, **87**(3), 478–503.

SOHANGIR S., WANG D., POMERANETS A. & KHOSHGOFTAAR T. M. (2018). Big data : Deep learning for financial sentiment analysis. *Journal of Big Data*, **5**(1), 1–25.

STONE P. J., DUNPHY D. C. & SMITH M. S. (1966). The general inquirer : A computer approach to content analysis.

SUCHMAN M. C. (1995). Managing legitimacy : Strategic and institutional approaches. *Academy of management review*, **20**(3), 571–610.

SUDDABY R., BITEKTINE A. & HAACK P. (2017). Legitimacy. *Academy of Management Annals*, **11**(1), 451–478.

TOST L. P. (2011). An integrative model of legitimacy judgments. *Academy of management review*, **36**(4), 686–710.

VAN DE KAUTER M., BREESCH D. & HOSTE V. (2015). Fine-grained analysis of explicit and implicit sentiment in financial news articles. *Expert Systems with applications*, **42**(11), 4999–5010.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł. & POLOSUKHIN I. (2017). Attention is all you need. *Advances in neural information processing systems*, **30**.

ZHANG L., XIAO K., ZHU H., LIU C., YANG J. & JIN B. (2018). Caden : A context-aware deep embedding network for financial opinions mining. In *2018 IEEE International Conference on Data Mining (ICDM)*, p. 757–766 : IEEE.

ZIMMERMAN M. A. & ZEITZ G. J. (2002). Beyond survival : Achieving new venture growth by building legitimacy. *Academy of management review*, **27**(3), 414–431.

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

État de l'art sur la coréférence

Fabien Lopez[1]
(1) Univ. Grenoble Alpes, CNRS, Grenoble INP*, LIG, 38000 Grenoble, France
* Institute of Engineering Univ. Grenoble Alpes
fabien.lopez@univ-grenoble-alpes.fr

RÉSUMÉ

La résolution des liens de coréférences est une tâche importante du TALN impliquant cohérence et compréhension d'un texte. Nous présenterons dans ce papier une vision actuelle de l'état de l'art sur la résolution des liens de coréférence depuis 2001 et l'avènement des modèles neuronaux pour cette tâche. Cela comprend les corpus disponibles en français, les méthodes d'évaluation ainsi que les différentes architectures et leur approche. Enfin nous détaillerons les résultats, témoignant de l'évolution des méthodes de résolutions des liens de coréférences.

ABSTRACT

State-of-the-art on coreference resolution

Coreference resolution is an important task in NLP involving coherence and comprehension of a text. In this paper we will present a current view of the state of the art on coreference link resolution since 2001 and the advent of neural models for this task. This includes corpora available in French, evaluation methods as well as different architectures and their approach. Finally, we will detail the results, showing the evolution of coreference link resolution methods.

MOTS-CLÉS : État de l'art, résolution de coréférences, anaphores.

KEYWORDS: State-of-the-art, coreference resolution, anaphora.

1 Introduction

La résolution automatique de liens de coréférences est une tâche du traitement automatique des langues (TAL) qui consiste à la détection, dans un texte, des mentions qui réfèrent à une même entité du monde réel ou imaginaire.

Une **entité**, du latin *entis* signifiant "étant", est "ce qui est" sur le plan physique ou en tant qu'objet de pensée. Une **mention**, du latin *mentio* signifiant "rappel en mémoire", est une expression linguistique permettant de faire référence à une entité. On appelle **coréférence** le lien entre au moins deux mentions faisant référence à la même entité. La **résolution de coréférences** est le fait de résoudre ces liens de coréférences.

La coréférence s'inscrit dans un processus de compréhension du texte tel que décrit par (Hobbs, 1986) et a été mis en avant en partie depuis (Vilain *et al.*, 1995a).

Exemple 1. "*Joe Biden* a décidé [...]. *Le président* a [...]" est une coréférence.
"*Joe Biden* a décidé [...]. *Il* a [...]" est une anaphore.

En se référant à l'exemple 1, on peut différencier la coréférence de l'**anaphore**. L'anaphore est la relation

Actes des 16e Rencontres Jeunes Chercheurs en RI (RJCRI) et 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL), pages 101–113.

entre deux mentions dont l'une des mentions fait référence à l'autre. La seconde mention nécessite la première pour être comprise contrairement à la coréférence où les deux mentions se suffisent à elles-mêmes.

On dit alors que *Il* est une **mention anaphorique** de *Joe Biden* et que *Joe Biden* est un **antécédent** de *Il*. Le type d'anaphore le plus souvent utilisé est l'anaphore pronominale (Lappin & Leass, 1994) tel qu'illustrée dans l'exemple 1.

Deux mentions coréférentes (ou anaphoriques) s'inscrivent dans une **chaîne de coréférences** aussi appelée **entité**.

Exemple 2. *"*Adèle *voulait un* chat. Elle *a fini par en acheter* un.

Dans l'exemple 2 : Les mentions sont "Adèle", "chat", "Elle" et "un" alors que les chaînes de coréférences sont *{"*Adèle*", "*Elle*"}* et *{"*chat*", "*un*"}*.

Les chaînes de coréférences jouent un rôle important dans la **cohésion** et la **cohérence** des documents. C'est-à-dire respectivement la structure du texte, la façon dont les textes sont liés entre eux, et la logique du texte, c'est-à-dire que le texte ne se contredit pas. Par exemple avec la phrase "*Le* sac *ne rentre pas dans le* casier. Il *est trop petit*" avec comme mentions : "*sac*", "*casier*", "*Il*" ; les chaînes de coréférences {"*sac*", "*Il*"} et {"*casier*"} ne sont pas cohérentes alors que les chaînes {"*sac*"} et {"*casier*", "*Il*"} ont un sens logique : le casier est trop petit pour accueillir le sac.

La résolution des liens de coréférences s'inscrit comme étant l'une des nombreuses tâches du TAL. Au cours des dernières années, la résolution automatique des liens de coréférences a connu une forte amélioration, en partie grâce à l'utilisation des méthodes d'apprentissage profond. Ce papier a pour but de recenser l'état de l'art sur la résolution des liens de coréférences à l'aide des méthodes d'apprentissage automatique avec un point de vue orienté vers le français. Il existe d'ores et déjà de multiples études recensant l'évolution de l'état de l'art sur la coréférence et sur l'anaphore (Ng, 2010; Sukthanker *et al.*, 2018; Stylianou & Vlahavas, 2021) abordant plus en détails certaines parties que nous ne pouvons développer ici.

Nous aborderons dans la section 2 les différents jeux de données disponibles en français ainsi que le corpus de référence pour les différents modèles. Nous détaillerons dans la section 3 différentes architectures permettant une certaine vision de la tâche de résolution automatique des liens de coréférences. Dans la section 4, nous décrirons la plupart des différentes métriques proposées à ce jour pour enfin, dans la section 5, présenter l'évolution des résultats suites aux différentes architectures proposées.

2 Les jeux de données

Une des problématiques autour des méthodes d'apprentissage neuronale est la quantité et la qualité des exemples contenus dans les corpus utilisés car influençant directement la qualité du modèle. Dans cette section nous détaillerons les deux principaux corpus annotés pour la coréférence en français que sont le corpus ANCOR (Muzerelle *et al.*, 2011, 2014) et le corpus DEMOCRAT (Lattice *et al.*, 2019). Nous évoquerons également les corpus ParCorFull 2.0 (Lapshinova-Koltunski *et al.*, 2022) et OntoNotes 5.0 (Weischedel *et al.*, 2013). ParCorFull 2.0 est un corpus annoté sur la coréférence aligné sur 4 langues, incluant le français, (plus anglais, allemand et portugais) tandis que OntoNotes 5.0 (aussi connu comme le corpus CoNLL-2012) est le corpus majoritairement utilisé afin de comparer les différentes méthodes d'apprentissage artificiel bien qu'étant en anglais. Il sera brièvement abordé afin de connaître les données avec lesquelles les résultats seront comparés en section 5.

2.1 ANCOR et DEMOCRAT

Le corpus ANCOR (Muzerelle *et al.*, 2011, 2014) est le premier corpus de grande taille annoté pour la résolution automatique de liens de coréférences en français. Il se base sur quatre sous-corpus :

- *ELSO_CO2* et *ESLO_ANCOR* (Eshkol-Taravella *et al.*, 2011), basé sur une partie retranscrite du corpus oral ESLO. Cette partie correspond à de l'oral spontané et plus particulièrement à des entretiens sociolinguistiques. ELSO_CO2 se compose de 35 000 mots pour 2,5 heures tandis que ELSO_ANCOR comporte 417 000 mots pour 25 heures.
- *OTG*, qui correspond aussi à de l'oral spontané et notamment à des dialogues en présentiel entre des individus et le personnel d'accueil de l'Office du Tourisme de Grenoble sur une durée de 2 heures pour 26 000 mots.
- *Accueil_UBS* qui correspond également à de l'oral spontané et spécifiquement à des dialogues par téléphone recueillis auprès du standard téléphonique d'une université pour une durée d'une heure et 10 000 mots

Ce corpus contient donc la transcription de 30,5 heures de parole pour 488 000 mots avec en tout plus de 110 000 mentions et 50 000 relations entre les mentions. Les types de relations pris en compte sont :

- La coréférence directe : les mentions sont des groupes nominaux avec le même lexique de têtes de mentions.
- La coréférence indirecte : les mentions sont des groupes nominaux avec un lexique de têtes de mentions différent.
- L'anaphore pronominale : la mention anaphorique est un pronom
- L'anaphore de pontage : la mention anaphorique nécessite son antécédent pour être comprise bien qu'elle ne fasse pas directement référence à la même entité (exemple de (Sukthanker *et al.*, 2018) : "J'étais sur le point d'acheter une *robe* lorsque j'ai vu une tâche sur *la dentelle*").
- L'anaphore de pontage pronominale : similaire à l'anaphore de pontage mais où la mention anaphorique est un pronom.

Le corpus DEMOCRAT (Lattice *et al.*, 2019) est proposé afin d'augmenter la quantité de textes annotés pour le français. Il est composé en couvrant des catégories de textes non couvertes par le corpus ANCOR. DEMOCRAT contient ainsi 58 textes d'environ 10 000 mots chacun pour un total de 689 000 mots, 198 000 expressions référentielles réparties en 20 000 chaînes de coréférences d'au moins deux mentions et un ensemble de singleton. Il s'agit d'un corpus diachronique qui couvre des textes écrits entre le XI^{e} et le XXI^{e}siècle. Parmi ces textes, 26 d'entre eux sont des portions d'œuvres de fiction telles que *Le ventre de Paris* d'Émile Zola, des fictions complètes si elles sont assez courtes comme *La morte amoureuse* de Théophile Gautier mais aussi des textes plus anciens comme *La chanson de Roland* et *La Vie de Sainte Bathilde*, datant du XI^{e} siècle et dont l'auteur n'est pas certain. Les 32 textes restant se composent de traités didactiques, de textes juridiques (code civile des français), journalistiques ou d'articles d'encyclopédie (articles wikipédia).

Contrairement au corpus ANCOR principalement axé sur les anaphores, le corpus DEMOCRAT annote aussi les singletons, c'est-à-dire les mentions n'apparaissant qu'une seule fois dans le texte.

2.2 ParCorFul 2.0

Reprenant et augmentant le corpus ParCorFull (Lapshinova-Koltunski *et al.*, 2018), lui-même basé sur le corpus ParCor (Guillou *et al.*, 2014), ParCorFull 2.0 (Lapshinova-Koltunski *et al.*, 2022) est un corpus annoté en coréférence aligné en Anglais, Allemand, Français et Portugais. Il peut ainsi servir à la fois en résolution automatique de liens de coréférence et en traduction. Il se base principalement sur des TED

	TED Talks			Médias			Total		
Langue	txt	nb phrases	nb mots	txt	nb phrases	nb mots	txt	nb phrases	nb mots
anglais	20	3 277	70 736	19	464	10 798	39	3 741	81 534
allemand	20	2 829	66 783	19	281	10 602	39	3 110	77 385
français	20	1 959	76 229	-	-	-	20	1 959	76 229
portugais	9	1 488	27 898	11	309	6 522	20	1 797	34 420
Total	69	9 553	241 646	49	1 054	27 922	118	10 607	269 568

TABLE 1 – Répartition des textes (*txt*), du nombre de phrases (*nb phrases*) et du nombre de mots (*nb mots*) pour les différents supports dans les différentes langues pour le corpus ParCorFull 2.0 (Lapshinova-Koltunski *et al.*, 2022).

Langue	nb mention	nb chaîne
Anglais	7 279	2 319
Allemand	7 634	2 425
Français	9 009	4 744
Portugais	4 269	1 208
Total	28 191	10 696

TABLE 2 – Répartition du nombre de mentions (*nb mention*) et du nombre de chaînes de coréférence (*nb chaîne*) par langue.

Talks mais aussi sur des médias (voir Table 1).

Comme on peut le voir dans la Table 2, le nombre de mentions est bien moindre par rapport à ANCOR et DEMOCRAT. Le principal apport de ParCorFull est son alignement dans les différentes langues. Ce corpus se basant sur le corpus ParCor, il utilise des textes provenant de TED Talks de IWSLT2013 (Guillou, 2012). Il utilise également des textes de IWSLT2014 (Lapshinova-Koltunski *et al.*, 2018) ainsi que des textes de IWSLT17 (Lapshinova-Koltunski *et al.*, 2022).

2.3 OntoNote 5.0

Proposé par (Weischedel *et al.*, 2013) et connu comme le corpus CoNLL-2012, OntoNote 5.0 recense plus de 2.9 millions de mots répartis sur les 3 langues que sont l'anglais, le chinois et l'arabe. La composition du corpus est décrite en table 3 tandis que la répartition des entités, liens et mentions dans les corpus d'entraînement, de validation et de test sont présentées dans la table 4. OntoNotes 5.0 est le corpus de référence pour la comparaison des différents modèles de résolution de liens de coréférences grâce à son grand nombre d'exemples. Cependant certains choix lors de sa conception peuvent prêter à débat, par exemple ce dataset n'est pas annoté pour les singletons.Dans la section suivante, nous présentons l'évolution des différentes approches pour la résolution automatique des liens de coréférences.

3 Approches

Les systèmes de résolution automatique de coréférences étaient d'abord statistiques (Soon *et al.*, 2001; Ng & Cardie, 2002) avant de s'orienter vers une approche de type apprentissage profond qui, grâce aux modèles basés sur les réseaux de neurones, a une meilleure capacité de généralisation (Lee *et al.*, 2018;

Type de document	anglais	chinois	arabe
Fil d'actualité	625	250	300
Nouvelles radiodiffusées	200	250	-
Conversation radiodiffusées	200	150	-
Données tirées du web	300	150	-
Conversation téléphonique	120	100	-
Nouveau/Ancien Testament	300	-	-
Total	1 745	900	300

TABLE 3 – Répartition du nombre de mots (en milliers) dans les différentes langues suivant les différents types de données.

Langue	Type	Entraînement	Validation	Test	Total
Anglais	Entités	35 143	4 546	4 532	44 221
	Liens	120 417	14 610	15 232	150 259
	Mentions	155 560	19 156	19 764	194 480
Chinois	Entités	28 257	3 875	3 559	35 691
	Liens	74 597	10 308	9 242	94 147
	Mentions	102 854	14 183	12 801	129 838
Arabe	Entités	8 330	936	980	10 246
	Liens	19 260	2 381	2 255	23 896
	Mentions	27 590	3 313	3 235	334 138

TABLE 4 – Répartition du nombre d'entités, de mentions et de liens entre les mentions pour l'anglais, le chinois et l'arabe pour les corpus de d'entraînement, de validation et de test.

Wu *et al.*, 2020; Miculicich & Henderson, 2022). Nous détaillerons les différents systèmes de résolution automatique de liens de coréférences suivant deux grands axes : le premier orienté sur la résolution des liens de coréférences entre les mentions deux à deux tandis que le second s'orientera sur les systèmes cherchant à symboliser l'entité à laquelle se réfèrent les mentions.

3.1 Approche orientée sur les mentions

L'approche orientée sur la résolution des liens de coréférences entre les mentions deux à deux est la plus simple à mettre en place. Elle commence avec des méthodes dites basées sur les paires de mentions.

L'idée de l'approche basée sur les paires de mentions (Soon *et al.*, 2001; Ng & Cardie, 2002) consiste à résoudre le lien de coréférence entre deux mentions dans le texte et les annoter le cas échéant. Une fois le texte traité, on peut ainsi, par transitivité, construire une chaîne de coréférence. Avec l'exemple 3, le modèle va tour à tour chercher un lien de coréférence entre les mentions {*Hillary Clinton*, *Bill Clinton*}, {*Bill Clinton*, *Clinton*} et {*Clinton*, *il*}.

Exemple 3. "Hillary Clinton *et* Bill Clinton *ont quitté la maison blanche.*
Clinton *a déclaré qu'*il *ne voulait pas parler de son voyage*"

(Soon *et al.*, 2001) proposent un premier modèle statistique basé sur une fonction de classement binaire dont l'objectif est de lier la mention avec le premier antécédent candidat précédant la mention si celui-ci est jugé comme un choix suffisamment adéquat. Cette méthode sera reprise par la suite, en particulier par (Ng & Cardie, 2002) qui listera tous les antécédents candidats précédents et choisira le meilleur.

Cette méthode est cependant limitée en terme de résultats. Son principal inconvénient étant de ne considérer que les deux mentions courantes. Ainsi une mention ambiguë entre deux chaînes de coréférences pourrait entraîner la fusion des deux chaînes pourtant incompatibles entre elles. Avec l'exemple 3, si le système trouve un lien de coréférence entre *Hillary Clinton* et *Clinton* alors il pourra produire la chaîne de coréférence {*Hillary Clinton*, *Bill Clinton*, *Clinton*, *il*}, mettant ainsi *Hillary Clinton* et *il* dans la même chaîne.

Essayant d'ajouter plus d'information dans la résolution des liens de coréférences, une nouvelle approche basée sur un classement des scores de coréférences des mentions est proposée. Le score de coréférence des mentions est une valeur cherchant à quantifier la similarité entre deux mentions. Dans les approches basées sur un classement des scores de coréférence des mentions (Denis & Baldridge, 2007, 2008; Rahman & Ng, 2009; Durrett & Klein, 2013; Wiseman *et al.*, 2015), l'objectif est de comparer les scores de coréférences d'une mention avec toutes les autres mentions afin de ne sélectionner que le meilleur antécédent possible dans l'ensemble des antécédents candidats du document.

Reprenant l'approche de (Ng & Cardie, 2002), (Denis & Baldridge, 2007) proposent d'utiliser une fonction apprise afin de retrouver l'antécédent auquel se réfère un pronom avant de le généraliser pour toutes les mentions avec (Denis & Baldridge, 2008). Relativement simple et toujours sujette aux erreurs, cette approche a laissé place à une variante plus coûteuse au niveau calculatoire mais limitant ce genre de problèmes. Afin de limiter ce problème, (Rahman & Ng, 2009) proposent l'introduction de ϵ comme un potentiel antécédent qui représentera l'absence d'antécédent. Ainsi, si le système ne trouve aucun antécédent convenable, il pourra lier la mention avec ϵ. Bien que limitant le problème de la détection de faux liens, ce type d'approche y est toujours sensible.

Alors que les différents systèmes proposés jusqu'alors s'inscrivaient dans une chaîne de traitement, (Lee *et al.*, 2017) proposent le premier système de résolution de liens de coréférences entièrement neuronal et appris de bout-en-bout, détectant les mentions et résolvant les liens de coréférences avec le même système. Leur système incorpore une méthode dite de tête souple (de l'anglais *soft-head*) qui permet de choisir la tête de mention, c'est-à-dire le mot portant l'information principale de la mention. Pour effectuer cette sélection, le système utilise un mécanisme d'attention tel que proposé par (Bahdanau *et al.*, 2014). Ce système a posé de nouvelles bases pour la résolution de liens de coréférences, proposant une architecture entièrement neuronale. Il sera repris et amélioré par (Lee *et al.*, 2018) que nous détaillerons dans la Section 3.2.

3.2 Approche orientée sur les entités

L'approche orientée sur les mentions ayant pour défaut de ne prendre en compte que les deux mentions courantes, (Clark & Manning, 2015) proposent une façon d'apporter de l'information sur les chaînes de coréférences dans leur entièreté dans leur prise de décision. Ce type d'approche a été grandement repris par la suite car plus instinctif et donnant de meilleurs résultats.

Un premier type d'approche orientée sur les entités est donc proposé par (Clark & Manning, 2015) utilisant une approche basée sur l'ensemble des paires de mentions de la chaîne de coréférence afin de symboliser les caractéristiques représentant cette chaîne. Cette méthode a ensuite été reprise par (Wiseman *et al.*, 2016) qui proposent une façon de calculer les caractéristiques des chaînes de coréférences à l'aide de réseaux de neurones récurrents. (Wiseman *et al.*, 2016) fut aussi repris par (Clark & Manning, 2016) qui proposèrent une méthode de calcul de score de similarité entre les chaînes de coréférences grâce à l'empilement de plusieurs modèles.

Donnant suite à (Lee *et al.*, 2017), (Lee *et al.*, 2018) ont incorporé entre autres l'approche basée sur la représentation des entités à leur précédente architecture. De plus, (Lee *et al.*, 2018) implémentent

une nouvelle façon de sélectionner les mentions potentielles, plus permissive, mais ajoute une couche supplémentaire à un système de contrôle permettant de mieux limiter le coût calculatoire dans son ensemble tout en obtenant de meilleurs résultats.

D'autres types d'approches ont vu le jour utilisant les nouvelles bases apportées par (Lee *et al.*, 2017). Par exemple, (Wu *et al.*, 2020) proposent CorefQA, un modèle basé sur les méthodes de questions/réponses afin de modéliser le problème de résolution des liens de coréférences tout en permettant de générer des données supplémentaires, alors que (Miculicich & Henderson, 2022) proposent un modèle de Transformer de type Graph2Graph permettant de prendre des décisions au niveau du document afin d'utiliser plus d'informations.

Après avoir passé en revue les différentes approches, nous présentons dans la section suivante les méthodes d'évaluation

4 Méthodes d'évaluation

Afin de comparer les différents modèles proposés, il est nécessaire de quantifier les performances des modèles, c'est-à-dire de les évaluer. Pour que la comparaison soit juste et équitable, elle doit être effectuée avec les mêmes données d'entraînement, de validation et de test ainsi que la même mesure. Cette mesure se fait à l'aide d'une métrique qui à pour but de quantifier la qualité de la sortie produite par un modèle. Pour produire une "bonne" métrique, plusieurs points évoqués par (Luo, 2005) et (Moosavi & Strube, 2016) doivent être respectés :

- Discrimination : une métrique doit être capable de discriminer une sortie correcte d'une mauvaise sortie du modèle.
- Granularité : cette métrique doit avoir le même degré de granularité sur l'ensemble de sa plage de valeurs (usuellement de 0 à 1).
- Interprétabilité : la métrique doit pouvoir être aisément interprétable, par exemple, un score élevé signifie une bonne résolution des coréférences alors qu'un score bas correspond à une mauvaise résolution des coréférences.

Bien que ces caractéristiques étant triviales, la plupart des métriques de résolution de liens de coréférences à ce jour sont mises en défaut dans certains cas.

Les différentes métriques servant à l'évaluation de la résolution de coréférences proposées à ce jour se classent selon 4 catégories : basée sur les mentions, basée sur un alignement optimal, basée sur les liens de coréférences ou basée sur les liens de coréférences mais ayant conscience des entités concernées (Sukthanker *et al.*, 2018).

On gardera les notations vu précédemment et on définit en plus $\mathcal{T}$: l'ensemble des données étiquettées et $\mathcal{R}$: l'ensemble des données prédites. On utilisera $|.|$ pour la cardinalité d'une chaîne de coréférence.

4.1 MUC

Implémentée par (Vilain *et al.*, 1995b), MUC est la première métrique proposée pour l'évaluation de la résolution de coréférence. Elle se place dans la catégorie des métriques basées sur les liens. L'idée principale est d'utiliser le nombre minimal de liens nécessaires pour relier toutes les mentions d'une chaîne de coréférences entre elles. La précision et le rappel sont alors définis comme suit :

$$Precision(\mathcal{T},\mathcal{R}) = \sum_{r \in \mathcal{R}} \frac{|r| - |partition(r,\mathcal{T})|}{|r| - 1} \quad (1) \qquad Rappel(\mathcal{T},\mathcal{R}) = \sum_{t \in \mathcal{T}} \frac{|t| - |partition(t,\mathcal{R})|}{|t| - 1} \quad (2)$$

Avec $|partition(r,\mathcal{T})|$: le nombre d'éléments de $\mathcal{T}$ ayant une intersection non-vide avec r. Cette métrique a cependant plusieurs défauts. En particulier, en prenant deux petites et deux grandes entités, c'est-à-dire des chaînes de coréférences avec peu de mentions et d'autres avec un nombre conséquent de mentions, alors une erreur fusionnant les deux petites entités entre elles aura le même impact qu'une erreur fusionnant les deux grandes entités, ce qui est contre-intuitif : la fusion de deux grandes entités représente une erreur plus importante car impactant plus de mentions. En allant dans un cas extrême, (Moosavi & Strube, 2016), avec le corpus CoNLL-2012, ont lié toutes les mentions de $\mathcal{T}$ ensemble, obtenant un $Rappel = 100$ et une $Precision = 78,44$ et une F_1-mesure $= 87,91$, soit un résultat meilleur que les modèles états de l'art actuels (voir Section 5). Un autre défaut repproché à MUC est la non considération des singletons, c'est-à-dire une mention non coréférente avec aucune autre mention.

4.2 B-Cubed

Implémentée par (Bagga & Baldwin, 1998), la métrique B-Cubed (aussi notée B^3) est proposée afin de prendre en compte la taille des entités. Elle rentre dans la catégorie des métriques basées sur les mentions. Pour se faire, un coefficient est introduit afin de calculer une moyenne pondérée sur l'ensemble des chaînes de coréférences.

$$Precision = \frac{1}{\sum_{r_j \in \mathcal{R}} |r_j|} \sum_{r_j \in \mathcal{R}} \sum_{t_i \in \mathcal{T}} \frac{|r_j \cap t_i|^2}{|r_j|} \quad (3) \qquad Rappel = \frac{1}{\sum_{t_i \in \mathcal{T}} |t_i|} \sum_{t_i \in \mathcal{T}} \sum_{r_j \in \mathcal{R}} \frac{|t_i \cap r_j|^2}{|t_i|} \quad (4)$$

Ne se basant pas sur les liens de coréférences résolus mais sur les mentions, B^3 subit **l'effet d'identification des mentions** qui fait qu'une mention coréférente, détectée comme étant coréférente mais étant placée dans la mauvaise chaîne de coréférence, améliorera les résultats rendant ceux-ci contre-intuitifs et non fiables.

4.3 CEAF

Introduite par (Luo, 2005), cette métrique entre dans la catégorie des métriques utilisant un alignement optimal de $\mathcal{T}$ dans $\mathcal{R}$ noté $g^*(\cdot)$. De plus, *CEAF* utilise une métrique de similarité notée ϕ pour calculer la précision et le rappel comme suit :

$$Precision = \frac{\sum_{t_i \in \mathcal{T}^*} \phi(t_i, g^*(t_i))}{\sum_{r_j \in \mathcal{R}} \phi(r_j, r_j)} \quad (5) \qquad Rappel = \frac{\sum_{t_i \in \mathcal{T}^*} \phi(t_i, g^*(t_i))}{\sum_{t_i \in \mathcal{T}} \phi(t_i, t_i)} \quad (6)$$

Ainsi une fonction ϕ différente donnera une précision et un rappel différents. Habituellement on parle de CEAF_m pour $\phi(t_i,r_j) = |t_i \cap r_j|$ et CEAF_e pour $\phi(t_i,r_j) = \frac{2*|t_i \cap r_j|}{|t_i| + |r_j|}$.

Utilisant directement les mentions, les différentes variantes de CEAF sont toutes sujettes au problème d'identification des mentions auquel s'ajoute le problème de la non considération de la taille des entités. Enfin, à cause de l'utilisation de l'alignement optimal, si ce dernier n'aligne pas une mention correcte de $\mathcal{T}$ à sa valeur dans $\mathcal{R}$ alors cette mention ne comptera pas comme une mention correcte.

4.4 MELA

Elle fut introduite par (Denis & Baldridge, 2009) à l'occasion des événements CoNLL-2011 et CoNLL-2012 (d'où le nom usuel de score CoNLL). MELA est une moyenne des F1-mesures des métriques MUC, B-Cubed et CEAF$_e$. Bien que la plus utilisée à ce jour, certains arguent que la moyenne de 3 métriques biaisées ne peut donner des résultats fiables (Moosavi & Strube, 2016). Elle se calcule suivant la formule :

$$CoNLL = \frac{MUC_{F_1} + B^3_{F_1} + CEAF_{F_1}}{3} \tag{7}$$

4.5 BLANC

Mettant en lumière les différents défauts des métriques précédentes, (Recasens & Hovy, 2011) proposent BLANC, une métrique basée sur la mesure Rand-index (Rand, 1971), afin d'obtenir une meilleure interprétation de la F1-mesure. Pour se faire, (Recasens & Hovy, 2011) se basent sur les liens de coréférence ainsi que les liens de "non-coréférence", c'est-à-dire si le modèle a réussi à ne pas relier deux mentions entre elles s'il ne fallait pas les relier. On notera ainsi *rc* comme les bons liens de coréférence, *wc* les mauvais liens de coréférence, *rn* les bons liens de non-coréférence et *wn* : les mauvais liens de non-coréférence. (Recasens & Hovy, 2011) calculent une précision et un rappel pour les liens de coréférences (noté respectivement P_c et R_c) et ainsi que pour les liens de non-coréférence (noté respectivement P_n et R_n) :

$$\begin{array}{lll} P_c = \frac{rc}{rc+wc} & P_n = \frac{rn}{rn+wn} & Precision = \frac{P_c+P_n}{2} \\ R_c = \frac{rc}{rc+wn} & R_n = \frac{rn}{rn+wc} & Recall = \frac{R_c+R_n}{2} \\ F_c = \frac{2P_cR_c}{P_c+R_c} & F_n = \frac{2P_nR_n}{P_n+R_n} & Fmeasure = \frac{F_c+F_n}{2} \end{array} \tag{8}$$

L'un des principaux défauts de BLANC est qu'en augmentant le nombre de mentions coréférentes alors le nombre global de mentions augmente et le nombre de mentions non-coréférentes augmente bien plus, rendant la métrique peu sensible à la résolution des bons liens de coréférences. En plus, (Moosavi & Strube, 2016) affirme que l'utilisation des liens de non-coréférences rend BLANC plus sensible que les autres métriques existantes au problème d'identification des mentions.

4.6 LEA

Proposée par (Moosavi & Strube, 2016), cette métrique se base sur les liens de coréférence résolus tout en incorporant la taille de l'entité dans un facteur d'importance représenté par le cardinal de l'entité. La précision et le rappel sont alors calculés tel que :

$$Precision = \frac{\sum_{r_j \in \mathcal{R}} (|r_j| * \sum_{t_i \in \mathcal{T}} \frac{link(r_j \cap t_i)}{link(r_j)}}{\sum_{r_z \in \mathcal{R}} |r_z|} \tag{9}$$

$$Rappel = \frac{\sum_{t_i \in \mathcal{T}} (|t_i| * \sum_{r_j \in \mathcal{R}} \frac{link(t_i \cap r_j)}{link(t_i)}}{\sum_{t_z \in \mathcal{T}} |t_z|} \tag{10}$$

	MUC			B^3			CEAF$_{\phi_4}$			CoNLL-2012
Nom	R	P	F_1	R	P	F_1	R	P	F_1	CoNLL
(Durrett & Klein, 2013)	72,9	65,9	69,2	63,6	52,5	57,5	54,3	54,4	54,3	60,3
(Clark & Manning, 2015)	76,1	69,4	72,6	65,6	56,0	60,4	59,4	53,0	56,0	63,0
(Wiseman *et al.*, 2015)	76,2	69,3	72,6	66,1	55,8	60,5	59,4	54,9	57,1	63,4
(Wiseman *et al.*, 2016)	77,5	69,8	73,4	66,8	57,0	61,5	62,1	53,9	57,7	64,2
(Clark & Manning, 2016)	78,9	69,8	74,1	70,1	57,0	62,86	62,5	55,8	59,0	65,3
(Lee *et al.*, 2017)	78,4	73,4	75,8	68,6	61,8	65,0	62,7	59,0	60,8	67,2
(Lee *et al.*, 2018)	81,4	79,5	80,4	72,2	69,5	70,8	68,2	67,1	67,6	73,0
(Joshi *et al.*, 2020)	85,8	84,8	85,3	78,3	77,9	78,1	76,4	74,2	75,3	79,6
(Wu *et al.*, 2020)	**88,6**	**87,4**	**88,0**	82,4	**82,0**	**82,2**	**79,9**	**78,3**	**79,1**	**83,1**
(Miculicich & Henderson, 2022)	85,9	86,0	85,9	79,3	79,4	79,3	76,4	75,9	76,1	80,5
(Chai & Strube, 2022)	87,2	85,3	86,3	80,7	78,6	79,6	78,2	75,2	77,6	80,9

TABLE 5 – Résultats sur l'ensemble de test CoNLL-2012.

5 Évolution des résultats

On peut voir ci-dessus (Table 5), l'évolution des résultats sur la résolution de coréférence avec les métriques MUC, B^3, CEAF$_{\phi_4}$ ($CEAF_e$). La F$_1$-mesure (F_1) est calculée suivant l'équation 11 avec P la précision et R le rappel de la métrique.

$$F_1 = \frac{2 \times P * R}{P + R} \tag{11}$$

L'évolution des résultats au fil des années a été grandement marquée par l'utilisation de modèles entièrement neuronaux (Lee *et al.*, 2017). Cependant d'autres éléments entrent en jeu, l'utilisation des modèles dont l'approche est basée sur les entités semble être une approche apportant de meilleurs résultats.

6 Conclusion

Cet état de l'art ne pouvant entrer dans les détails de chaque article mentionné, nous cherchions à apporter une vision en une dizaine de pages, sur l'évolution de la résolution de liens de coréférence depuis (Soon *et al.*, 2001) jusqu'à (Miculicich & Henderson, 2022). Il propose également une bibliographie sur les corpus travaillant sur le français, le corpus de référence utilisé à ce jour, les approches des différents systèmes ainsi que les métriques utilisées pour mesurer leur qualité.

Depuis 2013, l'utilisation des modèles neuronaux a permis une forte amélioration des résultats (+23 points en score CoNLL) mais ne peut encore être considérée comme une tâche résolue. En effet, elle laisse encore place à de possibles améliorations, par exemple :

— différentes méthodes de représentation des entités.
— l'augmentation du nombre de données à travers de nouveaux corpus ou un enrichissement des existants.

Par ailleurs, la résolution des liens de coréférences étant centrale dans la compréhension et l'analyse d'un document, nous sommes convaincus que l'utilisation d'autres tâches du domaine du TAL peuvent aider à l'amélioration de la résolution des liens de coréférences comme l'intégration de modules de résolution des liens de coréférences peuvent améliorer les performances d'autres tâches.

Remerciements

Ce travail a été supporté par le projet CREMA (Coreference REsolution into MAchine translation) financé par l'Agence Nationale de la Recherche (ANR), numéro de contrat ANR-21-CE23-0021-01. Par ailleurs, nous remercions les relecteurs anonymes pour leurs conseils instructifs et détaillés.

Références

BAGGA A. & BALDWIN B. (1998). Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference*, volume 1, p. 563–566 : Citeseer.

BAHDANAU D., CHO K. & BENGIO Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv :1409.0473*.

CHAI H. & STRUBE M. (2022). Incorporating centering theory into neural coreference resolution. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 2996–3002, Seattle, United States : Association for Computational Linguistics. DOI : 10.18653/v1/2022.naacl-main.218.

CLARK K. & MANNING C. D. (2015). Entity-Centric Coreference Resolution with Model Stacking. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 1405–1415 : Association for Computational Linguistics. DOI : 10.3115/v1/P15-1136.

CLARK K. & MANNING C. D. (2016). Improving Coreference Resolution by Learning Entity-Level Distributed Representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 643–653 : Association for Computational Linguistics. DOI : 10.18653/v1/P16-1061.

DENIS P. & BALDRIDGE J. (2007). A ranking approach to pronoun resolution. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, p. 1588–1593, San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

DENIS P. & BALDRIDGE J. (2008). Specialized models and ranking for coreference resolution. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, p. 660–669, Honolulu, Hawaii : Association for Computational Linguistics.

DENIS P. & BALDRIDGE J. (2009). Global joint models for coreference resolution and named entity classification. *Procesamiento del lenguaje natural, ISSN 1135-5948, Nº. 42, 2009, pags. 87-96*, **42**.

DURRETT G. & KLEIN D. (2013). Easy victories and uphill battles in coreference resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, p. 1971–1982, Seattle, Washington, USA : Association for Computational Linguistics.

ESHKOL-TARAVELLA I., BAUDE O., MAUREL D., HRIBA L., DUGUA C. & TELLIER I. (2011). Un grand corpus oral " disponible " : le corpus d'Orléans 1 1968-2012. *Revue TAL*, **53**(2), 17–46. HAL : halshs-01163053.

GUILLOU L. (2012). Improving pronoun translation for statistical machine translation. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, p. 1–10, Avignon, France : Association for Computational Linguistics.

GUILLOU L., HARDMEIER C., SMITH A., TIEDEMANN J. & WEBBER B. (2014). Parcor 1.0 : A parallel pronoun-coreference corpus to support statistical mt. In *9th International Conference on Language Resources and Evaluation (LREC), MAY 26-31, 2014, Reykjavik, ICELAND*, p. 3191–3198 : European Language Resources Association.

HOBBS J. (1986). *Resolving Pronoun References*, In *Readings in Natural Language Processing*, p. 339–352. Morgan Kaufmann Publishers Inc. : San Francisco, CA, USA.

JOSHI M., CHEN D., LIU Y., WELD D. S., ZETTLEMOYER L. & LEVY O. (2020). SpanBERT : Improving Pre-training by Representing and Predicting Spans. *Transactions of the Association for Computational Linguistics*, **8**, 64–77. DOI : 10.1162/tacl_a_00300.

LAPPIN S. & LEASS H. J. (1994). An algorithm for pronominal anaphora resolution. *Comput. Linguist.*, **20**(4), 535–561.

LAPSHINOVA-KOLTUNSKI E., FERREIRA P. A., LARTAUD E. & HARDMEIER C. (2022). ParCorFull2.0 : a parallel corpus annotated with full coreference. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 805–813, Marseille, France : European Language Resources Association.

LAPSHINOVA-KOLTUNSKI E., HARDMEIER C. & KRIELKE P. (2018). ParCorFull : A Parallel Corpus Annotated with Full Coreference. p.6̃.

LATTICE, LILPA, ICAR & IHRIM (2019). Democrat. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

LEE K., HE L., LEWIS M. & ZETTLEMOYER L. (2017). End-to-end Neural Coreference Resolution.

LEE K., HE L. & ZETTLEMOYER L. (2018). Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 2 (Short Papers)*, p. 687–692, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/N18-2108.

LUO X. (2005). On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, p. 25–32, Vancouver, British Columbia, Canada : Association for Computational Linguistics.

MICULICICH L. & HENDERSON J. (2022). Graph refinement for coreference resolution. In *Findings of the Association for Computational Linguistics : ACL 2022*, p. 2732–2742, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.findings-acl.215.

MOOSAVI N. S. & STRUBE M. (2016). Which Coreference Evaluation Metric Do You Trust ? A Proposal for a Link-based Entity Aware Metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 632–642 : Association for Computational Linguistics. DOI : 10.18653/v1/P16-1060.

MUZERELLE J., LEFEUVRE A., ANTOINE J.-Y., SCHANG E., MAUREL D., VILLANEAU J. & ESHKOL I. (2011). ANCOR, premier corpus de français parlé d'envergure annoté en coréférence et distribué librement. In ATALA, Éd., *TALN'2013, 20e conférence sur le Traitement Automatique des Langues Naturelles*, p. 555–563, Les Sable d'Olonne, France. HAL : hal-01016562.

MUZERELLE J., LEFEUVRE A., SCHANG E., ANTOINE J.-Y., PELLETIER A., MAUREL D., ESHKOL I. & VILLANEAU J. (2014). ANCOR_Centre, a Large Free Spoken French Coreference Corpus : description of the Resource and Reliability Measures. In ELRA, Éd., *LREC'2014, 9th Language Resources and Evaluation Conference.*, p. 843–847, Reyjavik, Iceland. HAL : hal-01075679.

NG V. (2010). Supervised noun phrase coreference research : The first fifteen years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, p. 1396–1411, Uppsala, Sweden : Association for Computational Linguistics.

NG V. & CARDIE C. (2002). Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, p. 104–111, Philadelphia, Pennsylvania, USA : Association for Computational Linguistics. DOI : 10.3115/1073083.1073102.

RAHMAN A. & NG V. (2009). Supervised models for coreference resolution. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, p. 968–977, Singapore : Association for Computational Linguistics.

RAND W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, **66**, 846–850.

RECASENS M. & HOVY E. (2011). Blanc : Implementing the rand index for coreference evaluation. *Natural language engineering*, **17**(4), 485–510.

SOON W. M., NG H. T. & LIM D. C. Y. (2001). A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, **27**(4), 521–544. DOI : 10.1162/089120101753342653.

STYLIANOU N. & VLAHAVAS I. (2021). A neural entity coreference resolution review. *Expert Systems with Applications*, **168**, 114466. DOI : 10.1016/j.eswa.2020.114466.

SUKTHANKER R., PORIA S., CAMBRIA E. & THIRUNAVUKARASU R. (2018). Anaphora and Coreference Resolution : A Review.

VILAIN M., BURGER J., ABERDEEN J., CONNOLLY D. & HIRSCHMAN L. (1995a). A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Conference on Message Understanding - MUC6 '95*, p.̃45 : Association for Computational Linguistics. DOI : 10.3115/1072399.1072405.

VILAIN M., BURGER J., ABERDEEN J., CONNOLLY D. & HIRSCHMAN L. (1995b). A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Conference on Message Understanding - MUC6 '95*, p.̃45 : Association for Computational Linguistics. DOI : 10.3115/1072399.1072405.

WEISCHEDEL R., PALMER M., MARCUS M., HOVY E., PRADHAN S., RAMSHAW L., XUE N., TAYLOR A., KAUFMAN J., FRANCHINI M., EL-BACHOUTI M., BELVIN R. & HOUSTON A. (2013). OntoNotes Release 5.0. DOI : 11272.1/AB2/MKJJ2R.

WISEMAN S., RUSH A. M., SHIEBER S. & WESTON J. (2015). Learning Anaphoricity and Antecedent Ranking Features for Coreference Resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 1416–1426 : Association for Computational Linguistics. DOI : 10.3115/v1/P15-1137.

WISEMAN S., RUSH A. M. & SHIEBER S. M. (2016). Learning Global Features for Coreference Resolution. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 994–1004 : Association for Computational Linguistics. DOI : 10.18653/v1/N16-1114.

WU W., WANG F., YUAN A., WU F. & LI J. (2020). CorefQA : Coreference resolution as query-based span prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 6953–6963, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.622.

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Etudes sur la géolocalisation de tweets

Thibaud Martin[1]
(1) Institut National des Sciences Appliquées de Lyon, 7 Av. Jean Capelle O, 69100 Villeurbanne, France
thibaud.martin@insa-lyon.fr

RÉSUMÉ

La géolocalisation de textes non structurés est un problème de recherche consistant à extraire un contexte géographique d'un texte court. Sa résolution passe typiquement par une recherche de termes spatiaux et de la désambiguïsation.

Dans cet article, nous proposons une analyse du problème, ainsi que deux méthodes d'inférence pour déterminer le lieu dont traite un texte :

1. Comparaison de termes spatiaux à un index géographique
2. Géolocalisation de textes sans information géographique à partir d'un graphe de co-occurrence de termes (avec et sans composante temporelle)

Nos recherches sont basées sur un corpus de 10 millions de tweets traitant de lieux français, dont 57 830 possèdent une coordonnée géographique.

ABSTRACT

Unstructured text geoparsing

Unstructured text geoparsing is a search problem in which we want to infer a geographical context from a short text. This usually involves the search of spatial terms as well as their disambiguation in order to obtain a set of coordinates.

In this article, we propose an analysis of the aforementioned problem, as well as two inference methods to solve it :

1. Comparing spatial terms to a gazetteer
2. Geoparsing text which does not contain geographical terms (e.g : cities, regions, countries, etc.) with a term co-occurrence graph (with and without temporal contextualisation)

Our research is based on a dataset of 10 million tweets pertaining to french places, in which 57 830 have a set of geographical coordinates.

MOTS-CLÉS : TALN, géolocalisation de textes, fouille de données, Twitter.

KEYWORDS: NLP, text geolocation, data mining, Twitter.

Actes des 16e Rencontres Jeunes Chercheurs en RI (RJCRI) et 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL), pages 114–130.

1 Introduction

Depuis plus d'une décennie, le partage d'informations médiatiques a évolué pour comporter de nouveaux canaux de communication. Historiquement, un article de presse était rédigé par un journaliste qui consultait directement des personnes qui avaient un lien avec le sujet en question. Aujourd'hui, la présence de réseaux sociaux tels que Twitter a ajouté un nouvel intermédiaire à cette communication : l'information est transmise par une petite quantité d'utilisateurs, puis est ensuite diffusée par effet boule de neige (likes, retweets, etc.), jusqu'à ce qu'elle soit suffisamment populaire pour être retranscrite dans un article.

Cette dynamique de communication fait l'objet de plus d'une décennie de recherches dans les domaines de la détection d'évènements, du suivi de catastrophes (ex : pandémies, phénomènes météorologiques, etc.), et de la géolocalisation d'utilisateurs. Le sujet commun de ces travaux est l'extraction de contextes géographiques à partir d'une agrégation de sources. Il s'agit d'un processus complexe, car les informations sur des localisations sont rares - par exemple, environ 1% à 3% des tweets mondiaux contiennent au moins une méta-donnée à caractère géographique (Qazi *et al.*, 2020) - et sont présentées à des échelles hétérogènes (ex : pays, région, ville, point d'intérêt).

La géolocalisation de textes est un problème de recherche dont l'objectif est d'extraire, à partir d'un texte et d'une quantité minimale de métadonnées, le ou les lieux qui sont cités, afin d'inférer une zone géographique qui pourra ensuite être traitée en aval par des processus de fouille de données. Cette méthode est décomposée en deux étapes :

— Une extraction de termes à caractère géographiques et / ou un rapprochement de certains termes communs à un évènement dynamique pour en inférer une localisation approximative
— Une désambiguïsation des différents termes géographiques en coordonnées (latitude / longitude) ou contours géographiques. Au-delà d'un simple géocodage, il faut ici retrouver la localisation exacte en fonction du contexte du message. Par exemple, la phrase « Je suis en train de visiter Paris » contient le terme *Paris*, mais 11 villes possèdent ce nom dans le monde.

Dans cet article, nous proposons une première analyse du problème de la géolocalisation de textes non structurés à partir d'un corpus de tweets que nous avons collecté et sélectionné. pour ce faire, nous avons appliqué plusieurs méthodes inspirées des constats relevés dans d'autres articles de recherche (cf. références), afin de donner plus de clarté sur les difficultés à surmonter pour détecter avec une forte précision la localisation d'un texte.

Tout d'abord, nous introduisons le sujet avec un état de l'art sur les jeux de données et méthodes utilisées par des travaux de recherche récents. Dans la section 3, nous présentons le jeu de tweets que nous avons généré et utilisé pour les différentes méthodes présentées dans la section 5. Enfin, nous apportons dans les parties 6 et 7 les résultats de nos expérimentations et ce que nous avons pu en tirer.

2 Etat de l'art

La géolocalisation basée sur des textes est principalement basée sur des jeux de données provenant de réseaux sociaux. Twitter est la source la plus fréquente dans les travaux de recherches actuels, puisqu'il permet d'avoir accès à une grande quantité de posts avec de la géolocalisation, même si leur quantité reste limitée face à au nombre total de tweets (Cheng *et al.*, 2010; Qazi *et al.*, 2020).

Prévue pour la détection de localisations d'utilisateurs de réseaux sociaux à partir de leurs propriétés et contenus, la géolocalisation de textes est maintenant utilisée dans d'autres domaines tels que la détection d'évènements (Hui *et al.*, 2021), d'interaction entre groupes (Kumar *et al.*, 2019), ou encore du suivi de comportements dynamiquement temporels comme des pandémies (Qazi *et al.*, 2020).

Parmi les méthodes proposées pour résoudre ce problème, de premières approches se basent sur l'usage de méta-données disponibles dans les tweets, telles que les coordonnées GPS ou les localisations déclarées par les utilisateurs dans leurs posts et leur profil (Zohar, 2021; Qazi *et al.*, 2020). Ces informations deviennent cependant limitées en raison de changements dans le fonctionnement de Twitter (Zhang *et al.*, 2022; Kruspe *et al.*, 2021), qui ne permettent plus de déclarer une position précise sans mettre à jour un post en utilisant l'API.

Une méthode des études réalisées sur le traitement de contenus textuels est l'extraction toponymique (Qazi *et al.*, 2020), qui consiste à récupérer des termes pertinents (ex : villes, régions) en comparant chaque terme d'un message à un index géographique suite à un filtrage des données non-pertinentes.

La popularisation de l'apprentissage supervisé a permis de réaliser des avancées sur la précision des méthodes historiques, avec en particulier l'utilisation de réseaux de neurones type CNN (Mahajan & Mansotra, 2021), RNN (Kumar *et al.*, 2019) et Bi-LSTM (Mahajan & Mansotra, 2021; Lau *et al.*, 2017), et plus récemment de transformers (Li *et al.*, 2022). Ces modèles permettent d'atteindre dans certains cas une précision au niveau du point d'intérêt, contrairement à l'identification de villes au mieux précédemment.

D'autres méthodes se basent sur la co-occurrence de termes dans des textes (Ozdikis *et al.*, 2018). L'idée est de pouvoir retrouver dans un texte, pour un ensemble de termes liés à des évènements dynamiques (ex : hôpital, théâtre, etc.), des localisations ayant été citées dans d'autres textes où ils apparaissent. Cela amène à l'utilisation de graphes de connaissance (Rossi *et al.*, 2020; Hui *et al.*, 2021).

3 Jeu de données

Notre jeu de données contient une agrégation de 10 millions de tweets avec l'ensemble de leurs métadonnées collectés depuis début 2020. Initialement centré sur le sujet du vélo dans la ville de Lyon, il fut ensuite étendu à l'ensemble du territoire français afin de pouvoir l'utiliser dans d'autres projets.

Twitter propose dans les métadonnées de ses messages deux propriétés à caractère géographique :

— coordinates : paire de coordonnées (latitude / longitude) définissable par l'utilisateur en mettant à jour un post par le biais de l'API de Twitter
— place : lieu nommé que l'utilisateur peut choisir à partir d'une liste fournie par Twitter lors de la création d'un tweet. En fonction de son type, l'objet retourné contient le nom du lieu, son pays et son contour (bounding box)

Cependant, l'usage de ces propriétés est très marginal par rapport au nombre total de tweets présents sur la plateforme. Sur notre jeu de données, seuls 57 000 tweets sont précisément géolocalisés (~0,5%) et ~500 000 tweets ont un champ « place » non nul (~5%).

Au-delà des considérations quantitatives, il est impossible de savoir quelle information un utilisateur

souhaite communiquer en utilisant l'une de ces métadonnées : il pourrait s'agir du sujet du message, ou bien de la localisation de la personne au moment où elle crée son tweet. Dans la figure 1, nous avons répertorié l'ensemble des 57 830 tweets de notre jeu de données possédant la propriété coordinates (retweets exclus). Les points les plus foncés sur la carte représentent les lieux avec la plus grande quantité de tweets.

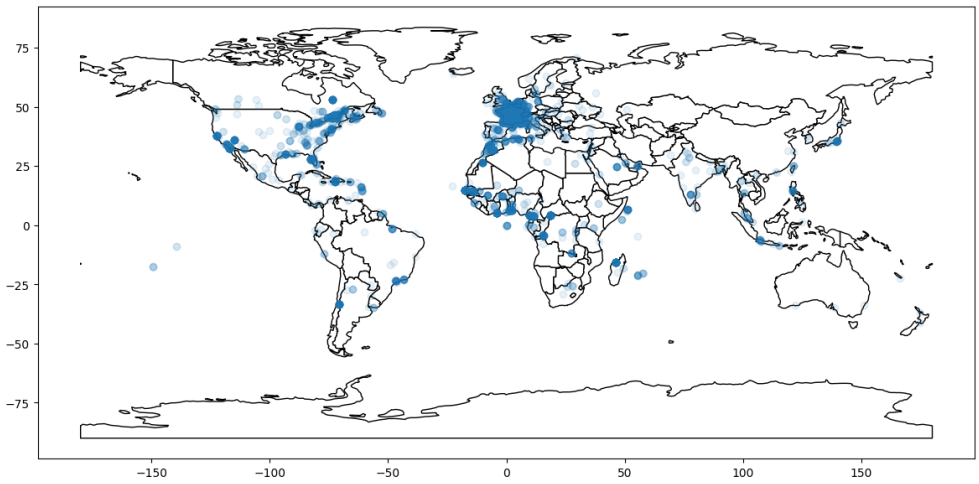


FIGURE 1 – Localisation des tweets dans notre jeu de données (coordonnées uniquement)

Nous avons ainsi noté que, malgré le fait que la grande majorité des tweets de notre jeu de données traitent de lieux en France métropolitaine, il est possible de remarquer des géolocalisations à l'étranger. Par exemple, le post https://twitter.com/PhotosAlain/status/1587845842182053890 traite du Moulin Rouge à Paris, alors que la géolocalisation indique la ville de Madrid en Espagne.

wEnfin, nous avons remarqué que la taille moyenne des tweets a évolué dans l'histoire, suite à de réguliers changements sur leur taille maximale. Historiquement fixée à 140 caractères, la limite fut augmentée en 2017 pour atteindre 280 caractères, et d'ici le printemps 2023 certains utilisateurs pourront aller jusqu'à 4000 [1]. Néanmoins, la figure 2 montre que la majorité des tweets se situent environ entre 8 et 22 mots. Cela peut poser un problème dans le cadre de la géolocalisation de ces textes, puisque la probabilité qu'un mot à caractère géographique apparaisse est plus faible.

4 Pré-traitement du corpus

Notre première version de jeu de données est contenu dans un fichier JSON unique, avec un objet par tweet. Pour nos premières expérimentations, nous n'avons gardé que les objets possédant une propriété *coordinates* non nulle, soit au total 57 830 lignes. Une première analyse des messages montre qu'il y a une grande quantité de duplications (en réalité, des retweets), que nous enlevons pour garder une instance unique par message :

1. https://twitter.com/elonmusk/status/1627388350612004865

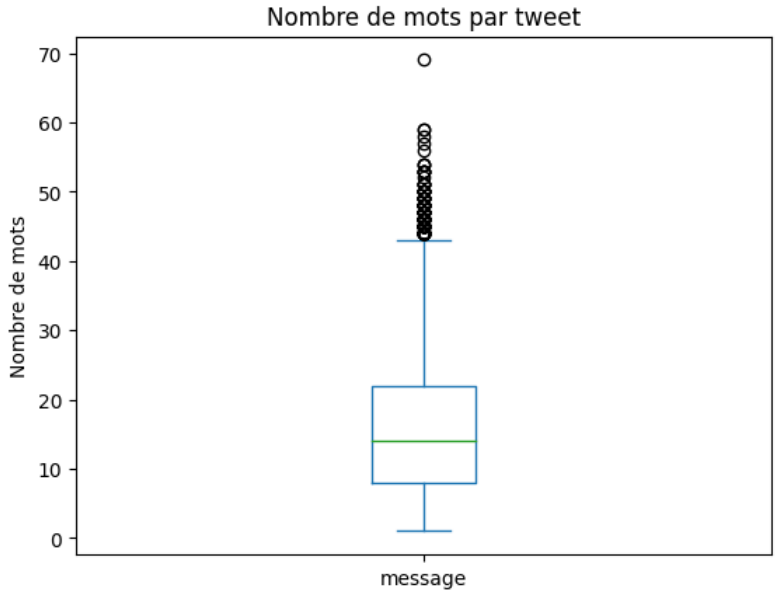


FIGURE 2 – Nombre de mots par tweet dans notre jeu de données

Nombre de tweets	tweets uniques	Message le plus fréquent	Fréquence
57 830	36 638	Vient de publier une photo à Paris France	3174

TABLE 1 – Analyse statistique des messages du jeu de données

Nous filtrons ensuite les métadonnées que nous n'avons pas considérées comme utiles pour notre expérimentation. Restent alors :

- Le message du tweet
- La date de création
- La propriété *coordinates*
- La propriété place (si elle existe)

Les messages du jeu de données sont principalement en français et en anglais, et contiennent des termes et des caractères qui ne sont pas pertinents dans le cadre d'une géolocalisation de texte. Nous avons choisi d'enlever les éléments suivants dans nos tweets :

- URL de redirection vers une page tierce (https ://t.co/[...])
- Emojis et caractères unicodes spéciaux
- Ponctuation (`!"#$%&'()*+,-./:;<=>?@[]^_`{}\~`)
  - Nous avions initialement décidé de garder les caractères ' et -, car beaucoup de villes françaises en font usage. Cela a changé suite aux implémentations de l'index géographique.
- Retour chariot (`\n` et `\r`)
- Mots vides (en anglais : *stop words*) en français et anglais
- tweets dont les coordonnées données ne sont pas en France

Après l'ensemble des étapes de pré-traitement, le jeu de données final est composé de 32 670 tweets. Par la suite, nous souhaitons aborder les hashtags, dont le format est plus compliqué à traiter puisque les mots sont attachés ensembles. Pour l'instant, nous les considérons comme des mots, ce qui permet tout d'analyser les termes "simples" à un mot (ex : #lyon, #paris).

5 Méthodologie

5.1 Extraction des termes géographiques

Dans cette partie, nous avons cherché à déterminer le nombre de tweets de notre jeu de données qui contiennent au moins un terme géographique (ex : nom d'une ville, d'un pays, d'un point d'intérêt, etc.). Dans le problème de la géolocalisation de texte non structuré, il s'agit d'instances plus simples que les autres, car nous pouvons directement appliquer des méthodes d'inférences sur ces termes qui auraient été extraits au préalable.

Pour ce faire, nous avons utilisé le module SpaCy [2] qui permet, entre autres, d'utiliser des méthodes de TALN classiques telles que la reconnaissance d'entités nommées. Notre objectif était d'extraire les entités de type *LOC* qui correspondent à une localisation géographique, à l'aide du modèle *fr_core_news_sm* [3] de SpaCy qui est entraîné sur des textes en français. Sur nos **32 670** tweets pré-traités, **18 601** contiennent au moins une entité *LOC*, soit environ 57% du jeu de données.

Certaines entités *LOC* ne sont cependant pas faciles à désambiguïser dans leur entièreté (ex : adresse exacte, combinaison de lieux comme « Paris Île-de-France France »). Ainsi, nous avons aussi tokenisé ces entités afin que, s'il n'est pas possible de géocoder l'entité composée de plusieurs mots, un essai soit réalisé sur les mots individuels.

Pour pallier aux 43% de tweets sans entité *LOC*, nous analysons aussi le message des tweets mot par mot. La structure des tweets étant libre grammaticalement, il est possible que SpaCy ait du mal à détecter certains termes comme étant des lieux. Au final, la figure 3 résume les étapes suivies afin d'extraire les termes géographiques de nos tweets.

5.2 Index géographique

Pour désambiguïser les termes géographiques que nous avons extraits dans la partie précédente, nous utilisons une base de données Elasticsearch [4] qui fait correspondre un terme à des coordonnées géographiques. Nous disposons au total de 13 070 831 noms de lieux qui proviennent de deux sources :

— Geonames [5], un dictionnaire gratuit couvrant 253 régions pour un total de 12 355 522 localisations
— Toutes les entités de DBpedia [6] contenant au moins un nom et une paire latitude / longitude, soit 715 309 lieux

La nature des tweets fait que les fautes d'orthographe sont courantes. En effet, les utilisateurs sont libres d'écrire ce qu'ils souhaitent, ce qui peut poser problème lors de l'identification de lieux. Nous requêtons ainsi notre base de données de deux manières :

— Recherche de terme exact (sensible à la casse)
— Recherche floue avec une distance d'édition variable en fonction de la taille de la chaîne de caractères en entrée (entre 3 et 6)

2. https://spacy.io/
3. https://spacy.io/models/fr#fr_core_news_sm
4. https://www.elastic.co/
5. https://www.geonames.org/
6. https://www.dbpedia.org/

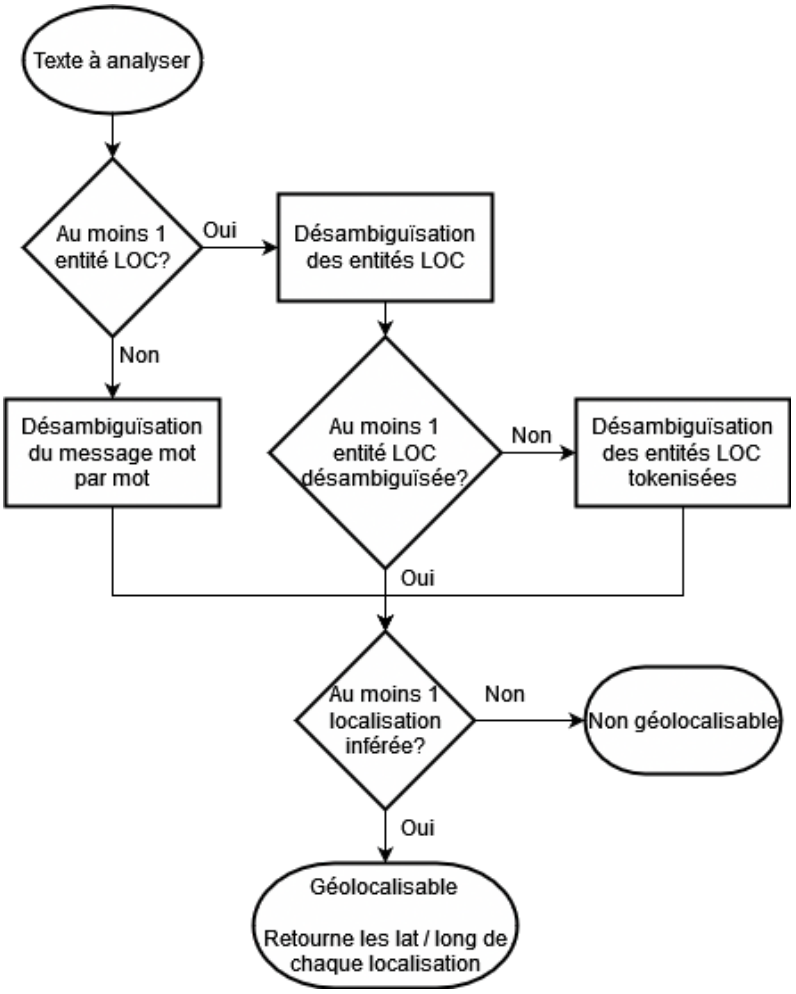


FIGURE 3 – Diagramme du processus d'extraction des entités géographiques

Chaque résultat retourné par Elasticsearch contient un score de confiance que nous utilisons pour choisir la localisation la plus pertinente. Nous appliquons un *boost* de 2 aux résultats retournés par la recherche de terme exacte afin qu'ils puissent apparaître au-dessus des autres.

Dans le cas où plusieurs termes seraient géolocalisés dans un tweet, nous retournons le centroïde composé des différentes paires de coordonnées données par l'index géographique. De plus, nous calculons un nouveau score de confiance afin de donner une indication sur la proximité de chaque point. En effet, plus les points sont dispersés, et moins le centroïde donné en résultat sera pertinent.

$$\text{confiance} = \max\left(1 - \frac{\sum_{i=0}^{n} \text{geodesic}(\text{centroid}, p_i)}{\|p\| - 1000}, 0\right) \tag{1}$$

Avec :

- **p** : ensemble de vecteurs de position (lat, long) inférées à partir du texte donné en entrée
- **centroid** : vecteur de position (lat, long) déterminé à partir des positions **p**
- **geodesic(p1, p2)** : fonction de calcul d'une distance géodésique entre 2 vecteurs de position (lat / long)
- Une distance maximale est utilisée pour avoir une borne supérieure de notre confiance. Pour toute distance ≥ 1000km, le score est de 0. Cette distance arbitraire correspond à la longueur

moyenne de la France, et sera revue dans la suite de nos travaux pour qu'elle ait plus de cohérence.

Par la suite, nous souhaitons remplacer le calcul de centroïdes par des calculs évitant les valeurs extrêmes, afin d'éviter que des positions aberrantes par rapport aux autres ne viennent fausser le résultat final.

5.3 Graphe de co-occurrence de termes

Nous avons constaté dans la partie sur l'extraction de termes géographiques qu'environ 43% des tweets de notre jeu de données ne contiennent pas d'entité *LOC* (localisation) identifiables par SpaCy. Au-delà des limitations du module, cela est aussi explicable par le fait que beaucoup de textes ne contiennent pas de termes géographiques.

Pour tenter de géolocaliser ces textes, nous proposons l'usage d'un graphe qui répertorie des co-occurrences entre termes, indépendamment de leur type (géographique ou non). Cette co-occurrence est définie par la présence de deux termes distincts dans un même texte. L'objectif est de trouver des liens entre des termes non géographiques et des lieux qui pourraient être cités dans des textes différents de celui que nous cherchons à géolocaliser.

Pour ce faire, nous avons alimenté un graphe Neo4J [7] en extrayant les noms et noms propres dans chaque texte de notre jeu de données. Pour cela, nous avons utilisé la fonctionnalité de POS-Tagging de SpaCy (codes POS : NOUN et PROPN). Ensuite, nous utilisons notre index géographique pour déterminer quels termes correspondent à une localisation. Enfin, nous générons le graphe avec deux types de sommets contenant des propriétés différentes :

— Entité géographique : nom, latitude, longitude
— Entité non géographique : nom

Nous créons ensuite une arête par co-occurrence. Puisque deux termes peuvent être co-occurrents dans plusieurs textes différents, nous identifions chaque arête par la date de création du tweet concerné.

Au final, notre graphe de co-occurrence comporte :

— 795 entités géographiques
— 24 580 entités non géographiques
— 11 141 arêtes entre entités géographiques
— 553 018 arêtes entre entités non géographiques
— 122 123 arêtes d'une entité géographique à une non géographique

5.3.1 Approche avec le graphe global

Dans cette situation, nous ne prenons pas en compte les différentes dates si plusieurs co-occurrences sont présentes entre deux sommets. Dans le cas où le terme donné serait identifié comme une entité géographique, nous retournons directement ses coordonnées (latitude, longitude).

Au contraire, si le terme est une entité non géographique, nous utilisons un *BFS* (Breadth First Search) avec une profondeur maximale de 5 pour explorer son voisinage. L'algorithme s'arrête lorsqu'une

7. https://neo4j.com/

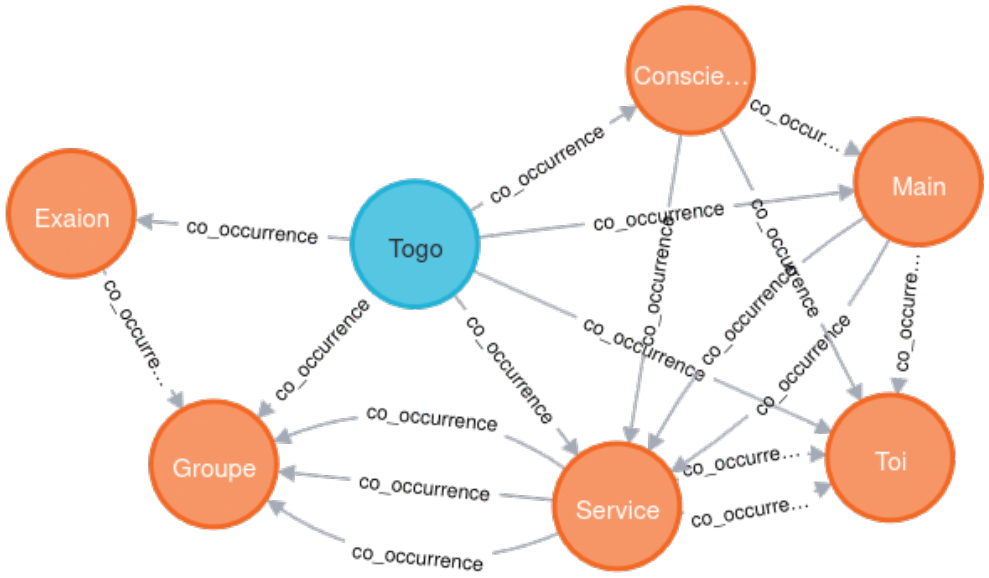


FIGURE 4 – Exemple de graphe de co-occurrence de termes (Bleu = Entité géographique, Rouge = Entité non géographique)

première entité géographique est trouvée, puis retourne ses coordonnées géographiques et le nombre de sommets traversés pour estimer une confiance sur le résultat :

$$\text{confiance} = 1 - \frac{\text{profondeur}}{\text{PROFONDEUR\_MAX} - 1} \tag{2}$$

5.3.2 Approche avec composante temporelle

Pour rendre les résultats de l'approche précédente plus pertinente, nous souhaitons faire le lien avec le domaine de la détection d'évènement. Ainsi, nous introduisons une méthode basée sur une proximité temporelle, qui favorise les co-occurrences de termes les plus « fraîches ». Notre intuition se base sur le fait que certains lieux dépendent d'évènements : par exemple, la « Fête des lumières » est très fréquemment associée à la ville de Lyon puisqu'elle se déroule dans cette ville.

Pour obtenir des résultats plus précis, nous envisageons, pour chaque terme à analyser dans le graphe, à réaliser une découpe pour ne garder que les sommets et les arêtes avec une date comprise entre t_0 (date de création du tweet correspondant) et $t_0 - x$ (x étant un paramètre). En complément, nous souhaitons aussi assigner à chaque arête restante un poids de « fraîcheur » afin de pouvoir appliquer l'algorithme de recherche de Dijkstra et privilégier les termes avec une co-ocurrence proche du tweet à analyser. Le calcul de ces poids pourrait s'apparenter à une gaussienne définie dans l'intervalle $[t_0 - x, t_0]$.

5.4 Approche incrémentale

Malgré le pourcentage de textes géolocalisables très élevé pour chaque proposition (cf. résultats), il peut être dans certains cas nécessaire d'avoir un rappel maximal, en dépit d'une précision plus faible. Pour ce cas de figure, nous proposons une approche incrémentale qui consiste, pour chaque

tweet, à exécuter toutes les méthodes citées précédemment une par une, jusqu'à ce qu'au moins une localisation soit extraite dans une paire type d'entité / méthode donnée. L'ordre est décidé par les précisions mesurées individuellement au préalable.

```
def inferer_localisation(texte, liste_methodes)
    localisations = []

    for methode in liste_methodes:
        for type_entite in ["LOC", "LOC_TOKENS", "TEXT_TOKENS"]:
            entites = texte[type_entite]
            if len(entites) > 0:
                for entite in entites:
                    localisation_inferee = methode(entite)
                    if localisation_inferee is not None:
                        localisations.append(localisation_inferee)

            if len(localisations) > 0:
                break

        if len(localisations) > 0:
            break
```

Algorithme 1 – Code Python de l'approche incrémentale

6 Evaluation - Résultats

Les programmes correspondants aux méthodes présentées ci-dessus ont été exécutés sur une machine avec un Intel® Core™ i5-8600K et 24 GiB de RAM. Pour les 32 670 tweets de notre jeu de données, le temps d'exécution moyen est de 5 min 30 pour les différents types d'index géographique. Pour le graphe de co-occurrences, ce temps est proportionnel à sa taille :

— Pour un petit graphe (5125 sommets, 18 118 arêtes) : ~5 minutes
— Pour un graphe de plus grande taille (75 953 sommets, 1 333 974 arêtes) : ~2 heures

Pour le processus d'évaluation, nous avons utilisé les métriques suivantes (toutes les distances sont exprimées en km) :

— % Géolocalisable : pourcentage de tweets où au moins une localisation a pu être extraite du texte
— Dist. moyenne : distance moyenne entre la localisation inférée et la vérité (coordonnées du tweet)
— Q(0.1), ..., Q(0.5) : quantiles des distances entre la localisation inférée et la vérité (coordonnées du tweet)
— Acc @k km : pourcentage de tweets dont la distance entre la localisation inférée et la vérité (coordonnées du tweet) est inférieure ou égale à k. En général, une Acc @10 km ou moins équivaut à la précision d'une ville, et une Acc entre 50 et 100km correspond à un département.

Dans cette partie, les méthodes suivront le nommage suivant :

— **Index géo** : index géographique
  — *DBPedia* : utilisation des localisations de DBpedia dans l'index
  — *Geonames* : utilisation des localisations de Geonames dans l'index

    - *Exact* : Recherche de terme exact (casse, orthographe)
    - *Fuzzy* : Recherche de terme floue avec une distance d'édition variable en fonction de la taille de la chaîne de caractères en entrée (entre 3 et 6)
- **Graphe de co-occurrences** (**uniquement la version sans composante temporelle**)
- **Approche incrémentale**
    - *Entités LOC* : extraction des termes géographiques sur les entités *LOC* (localisation) détectés par SpaCy dans le texte
    - *Texte tokenisé* : extraction des termes géographiques sur le texte mot par mot
    - *LOC tokenisé* : extraction des termes géographiques sur les entités *LOC* mot par mot

6.1 Résultats des méthodes

Les tableaux 2 et 3 présentent les résultats des méthodes citées précédemment. L'approche incrémentale est divisée pour montrer la précision qui peut être obtenue pour chaque type d'entité utilisé.

Méthode	% Géolocalisable	Dist. moyenne	Q(0.1)	Q(0.25)	Q(0.5)	Distance max.
Index géo, DBpedia, Exact	0,958	**175,841**	**0,105**	**0,798**	**3,642**	17 016,5
Index géo, DBpedia + Geonames, Exact	0,998	1814,32	0,174	15,535	1107,81	18 916,3
Index géo, DBpedia + Geonames, Fuzzy	**0,999**	1956,97	0,174	104,044	1438,21	19 031,3
Graphe de co-occurrences	0,954	294,115	0,207	3,592	96,697	**10 907,7**
Approche incrémentale (Entités *LOC*)	0,4	**283,939**	**0**	**0**	**0**	19 031,3
Approche incrémentale (Texte tokenisé)	**1**	388,026	0,955	3,133	69,36	**17 016,5**
Approche incrémentale (*LOC* + *LOC* tokenisé)	0,565	354,502	**0**	**0**	0,105	19 031,3
Approche incrémentale (tous types)	**1**	341,202	0,105	1,251	4,284	**17 016,5**

TABLE 2 – Statistiques des distances par rapport à la vérité (distances en km)

Méthode	Acc @1km	Acc @5km	Acc @10km	Acc @50km	Acc @100km
Index géo, DBpedia, Exact	**0,27**	**0,573**	**0,62**	**0,694**	**0,73**
Index géo, DBpedia + Geonames, Exact	0,137	0,22	0,23	0,261	0,275
Index géo, DBpedia + Geonames, Fuzzy	0,13	0,198	0,208	0,234	0,245
Graphe de co-occurrences	0,136	0,301	0,326	0,41	0,505
Approche incrémentale (Entités *LOC*)	**0,738**	**0,806**	**0,816**	**0,847**	**0,853**
Approche incrémentale (Texte tokenisé)	0,105	0,37	0,414	0,476	0,551
Approche incrémentale (*LOC* + *LOC* tokenisé)	0,606	0,744	0,763	0,802	0,816
Approche incrémentale (tous types)	0,228	0,531	0,579	0,653	0,688

TABLE 3 – Précision des méthodes

Nous pouvons constater que la méthode d'index géographique avec une recherche de termes exacts et l'usage de DBpedia comme unique source de localisations produit les meilleures précisions de manière générale. L'ajout de Geonames permet de géolocaliser plus de tweets, mais la distance entre

la vérité et les termes inférés est trop élevée pour être pertinente : dans le tableau 3, seulement 27,5% des tweets inférés par une recherche de termes exacte possèdent une précision inférieure à 100km (Acc @100km), tandis que la recherche de termes floue atteint 24,5%.

La méthode permettant d'identifier le plus de localisations dans les tweets est l'approche incrémentale (% géolocalisable = 100 %). Malgré une précision sensiblement plus faible que notre meilleure méthode, les résultats restent assez comparables (avec une Acc @5km égale à ~53%, contrairement à ~57% pour l'index géographique), ce qui permet d'avoir un bon équilibre entre les deux critères.

En comparaison avec les différents types de termes tokenisés, les entités *LOC* donnent les meilleurs résultats avec ~74% de précision $\leq$ 1km, ce qui laisse à croire que le NER de SpaCy garantit d'extraire des termes de qualité. Cependant, nous constatons que cela ne concerne que 40% du jeu de données.

Les localisations inférées se concentrent aussi sur une liste restreinte de villes en France. Par exemple, pour la méthode *Index géo, DBpedia, Exact* qui retourne 42 649 lieux, les cinq les plus inférés sont :

— Paris (24 593, soit ~58%)
— Lyon (5046, soit ~12%)
— France (4644, soit ~11%)
— Bordeaux (2047, soit ~5%)
— Publier (761, soit ~2%), qui est potentiellement un faux positif : Publier est une ville situé à la frontière franco-suisse, mais aussi un terme non-géographique très commun dans le contexte d'un tweet

Cela signifie que ~88% des localisations inférées font partie de cette liste.

Nous nous sommes aussi intéressés au score de confiance calculé à partir du calcul de centroïde pour les tweets avec ≥ 2 localisations identifiées, afin de connaître l'effet de plusieurs critères sur la qualité des inférences.

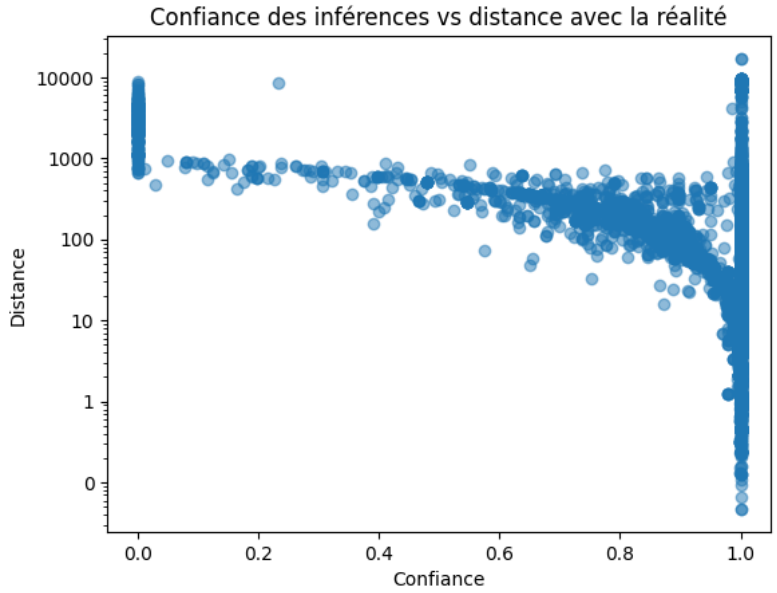


FIGURE 5 – Comparaison entre la confiance des inférences et la distance avec la réalité

La figure 5 montre une corrélation bien observable entre le score de confiance obtenu et la distance

entre la localisation inférée d'un tweet et la vérité. Nous pouvons cependant observer qu'aux deux extrémités de l'intervalle, nous retrouvons les tweets avec les distances les plus élevées. Cela peut s'expliquer de deux manières :

— Les inférences avec une confiance de 0 sont les moins cohérentes, puisque les localisations utilisées pour calculer le centroïde sont très éparpillées
— Les inférences avec une confiance de 1 et une distance très éloignée avec la vérité sont des erreurs de localisations. Un travail supplémentaire sur le calcul du score de confiance pourrait diminuer cet effet

Afin de pousser cette analyse, nous avons souhaité savoir si le score de confiance était impacté par le nombre de localisations que nous avons pu extraire dans un tweet.

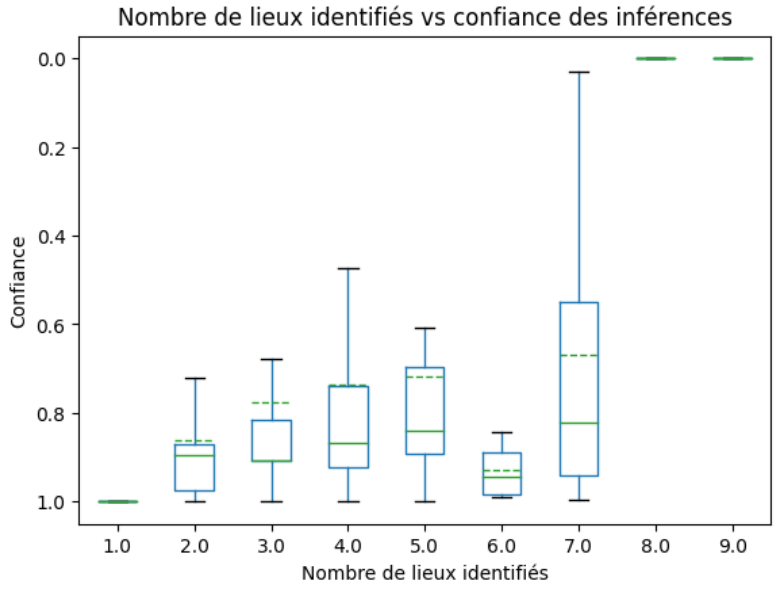


FIGURE 6 – Représentation de la confiance des inférences en fonction du nombre de lieux identifiés dans un texte

La figure 6 confirme que plus nous avons de localisations extraites dans un tweet donné, plus le centroïde calculé pour l'inférence a des chances d'être peu précis. Cela peut être causé par l'apparition plus probable de positions aberrantes (ex : [« Paris », « Musée du Louvre », « Jardin des Tuilleries », **« New York »**] qui auront tendance à fausser le résultat.

A cet effet, les observations sur les figures 5 et 6 nous confortent dans l'idée de remplacer le calcul d'un centroïde par l'utilisation de clusters pour les inférences avec ≥ 2 localisations identifiées, afin de concentrer le résultat à un ensemble de lieux qui sont proches.

6.2 Résultats de l'extraction d'entités géographiques

Afin d'appréhender l'efficacité de notre méthode d'extraction des entités géographiques, nous avons souhaité connaître le nombre de tweets géolocalisables par type d'entité et méthode d'inférence sur l'approche incrémentale. Le tableau 4 résume une exécution sur nos 32 670 tweets :

— chaque case indique le nombre de tweets dont au moins une localisation a été identifiée pour la méthode (ligne) et le type d'entité (colonne) testés

— il n'y a pas de remise. Lorsqu'une localisation est identifiée pour un tweet, le type d'entité et la méthode sont retenues, et l'exécution ne passe pas à la suite. Le tableau se lit de gauche à droite, puis de haut en bas (cf. code de l'approche incrémentale).

Méthode	Entité *LOC*	*LOC* + *LOC* tokenisé	Texte tokenisé
Index géo, DBpedia, Exact	9515	6609	15 167
Index géo, DBpedia + Geonames, Exact	52	158	1121
Index géo, DBpedia + Geonames, Fuzzy	0	0	32
Graphe de co-occurrences	0	0	16

TABLE 4 – Nombre de tweets géolocalisables par méthode et type d'entité (sur un corpus de 32 670 textes)

Nous remarquons que le nombre de tweets dont au moins une entité *LOC* extraite par SpaCy est géolocalisable est assez faible comparé aux autres types de termes, ce qui est cohérent avec les 40% de textes géolocalisables constatés ci-dessus. Nous nous sommes donc demandés si cela est causé par un manque de performance de SpaCy ou un problème venant de nos méthodes.

Au total, le nombre de tweets sans entité *LOC* est de 14 069, soit ~43% du jeu de données pré-traité. Parmi ces textes, le pourcentage de géolocalisation dépend de la méthode utilisée (cf. résultats des méthodes). Le tableau 5 résume le nombre d'entités *LOC* qui ont pu être géolocalisées par type de méthode (exécutée individuellement).

Méthode	Entités *LOC* avec géolocalisation	Entités *LOC* sans géolocalisation
Index géo, DBpedia, Exact	219	5424
Index géo, DBpedia + Geonames, Exact	702	4941
Index géo, DBpedia + Geonames, Fuzzy	1253	4390
Graphe de co-occurrences	940	4703

TABLE 5 – Nombre d'entités *LOC* géolocalisables ou non par méthode

Ces résultats démontrent que, malgré la grande quantité d'entités *LOC* uniques disponibles, peu d'entre elles sont véritablement géolocalisées par nos méthodes. Voici quelques exemples non-exhaustifs :

— Hôtel Novotel Paris Tour Eiffel
— Palais Tokyo Ville Paris
— Stade Groupama Lyon
— Palais Luxembourg invalides

Ce qui indique que des localisations « complexes » ne sont pas géolocalisables par notre index géographique actuel. Ceci est une des raisons pourquoi nous utiliserons un index plus performant par la suite (cf. index géographique).

7 Discussion

7.1 Limites de notre jeu de données

L'agrégation des tweets que nous avons employé pour générer notre jeu de données apporte des biais sur le fonctionnement de nos méthodes et les résultats de nos méthodes. En effet, la majorité des textes parlent uniquement de lieux en France, et les inférences réalisées par nos différentes méthodes se concentrent sur un petit sous-ensemble de régions et villes (cf. résultats).

De plus, les hashtags ne peuvent pas être pris en compte en l'état actuel. En effet, nous utilisons soit SpaCy soit une fonction de tokénisation pour traiter les différents termes dans notre corpus, or elles ne sont pas capables de traiter un ensemble de mots attachés, dont la casse est généralement aléatoire (souvent en titlecase ou minuscule). Nous avons pour objectif de nous concentrer sur le sujet dans la suite de nos travaux.

Par ailleurs, nous avons pu montrer que le NER de SpaCy possède des limites sur notre jeu de données qui sont causées par le contexte multilingue (anglais + français), mais aussi par la nature des tweets qui possèdent une grammaire et une orthographe très libre. D'autres NER comme Flair [8] et TwitterNER [9] doivent être explorés afin de potentiellement pouvoir améliorer l'extraction des entités géographiques.

Enfin, le fait de n'avoir que des paires de coordonnées géographiques nous empêche de mieux décrire les différents niveaux de précision de géolocalisation qui sont possibles. En effet, il est possible qu'un tweet parle de la France en général, donc les coordonnées GPS de la vérité et de la localisation inférée par nos méthodes ait une grande distance, ce qui fausse nos métriques actuelles. Nous souhaitons donc utiliser, où cela est possible, différents niveaux de précision (ex : ville, région, pays) qui peuvent être évaluées avec des contours géographiques, comme la propriété *place* proposée dans certains tweets (cf. les questionnements sur les données utilisables et la partie sur l'index géographique)

7.2 Limites du graphe de co-occurrences

Nous avons pu remarquer dans la partie résultats des méthodes que le graphe de co-occurrence que nous proposons n'atteint pas des performances similaires à notre meilleur type d'index géographique et notre approche incrémentale. Tout d'abord, nous n'avons pas encore pu tester la version avec composante temporelle, qui devrait supposément garantir une meilleure précision en ajoutant plus de localisations inférées à partir de termes non géographiques.

Malgré tout, nous concluons qu'il n'y a pas assez de « chemins » de co-occurrence utiles, ce qui veut dire que nous n'avons pas la capacité d'inférer plus de localisations qu'avec la méthode de l'index géographique. Pour pallier ce problème, nous suggérons de générer le graphe à partir d'un jeu de données différent, composé de textes plus longs (ex : articles de presse, Wikipedia) que des tweets qui sont majoritairement très courts (cf. jeu de données). Cela devrait permettre d'avoir une plus grande richesse de co-occurrences sur des données de meilleure qualité syntaxique.

8. https://huggingface.co/flair/ner-french
9. https ://github.com/napsternxg/TwitterNER

7.3 Questionnement sur les données utilisables

Notre étude du problème de géolocalisation de textes non structurés s'est concentré uniquement sur des tweets, comme pour la majorité des travaux dans le domaine (Cheng *et al.*, 2010; Hui *et al.*, 2021; Kruspe *et al.*, 2021; Lau *et al.*, 2017; Li *et al.*, 2022; Mahajan & Mansotra, 2021; Zohar, 2021; Zhang *et al.*, 2022). Cependant, il serait intéressant de pouvoir géolocaliser d'autres types de sources comme les articles de presse, dont la disponibilité en ligne et leur qualité d'information surpasse la contribution d'utilisateurs de réseaux sociaux.

Les éditions ne proposent généralement pas de système permettant de facilement récupérer les lieux qui sont traités dans leurs articles, ce qui rend pertinent l'usage des mêmes méthodes utilisées sur Twitter pour extraire des localisations. Malheureusement, la quantité de méta-données mises à disposition est moindre : sur Twitter, nous pouvons utiliser les propriétés *coordinates* et *place* afin d'avoir une idée de la localisation des posts qui les possèdent.

A cet effet, notre problème peut être traité de deux manières :

— Développer des méthodes spécifiques à des tweets, en utilisant les méta-données fournies
— Se généraliser à des textes de tailles variables (court ou long), avec un minimum de méta-données disponibles

La première solution nous permettrait d'utiliser des informations supplémentaires pour améliorer la précision des localisations inférées :

— La propriété *place*, qui permet de connaître le type de lieu (point d'intérêt, ville, région, etc.) et fournit un contour géographique (quand cela est applicable) afin de décider si une localisation inférée par nos méthodes se situe bien dans le lieu en question
— Les hashtags (cf. limites de notre jeu de données)
— La localisation de l'utilisateur qui publie le tweet, mais aussi celle des utilisateurs mentionnés dans le texte

8 Conclusion

Dans cet article, nous avons pu évaluer la performance de certaines méthodes communes dans le domaine de la géolocalisation de textes non structurés, à partir d'un jeu de données basé sur des tweets traitant de lieux en France. Les résultats nous encouragent dans la poursuite de leur utilisation, même si des améliorations notables sont à réaliser : par exemple, agréger une plus grande quantité de tweets pour avoir des données plus hétérogènes, et utiliser un index géographique plus complet pour mieux désambiguïser des termes géographiques.

Dans la suite de nos travaux, nous souhaitons tester les méthodes basées sur le principe du plongement de mots et les réseaux de neurones, ainsi que de développer notre propre méthode basée sur un graphe de co-occurrences à composante temporelle.

Références

CAILLAUT G., GRACIANNE C., AUCLAIR S., ABADIE N. & TOUYA G. (2022). Annotation sémantique pour la géolocalisation d'entités spatiales dans des tweets. In *PFIA Résilience et IA*, Saint-Etienne, France. HAL : hal-03682484.

CHENG Z., CAVERLEE J. & LEE K. (2010). You are where you tweet : A content-based approach to geo-locating twitter users. *International Conference on Information and Knowledge Management, Proceedings*, p. 759–768. DOI : 10.1145/1871437.1871535.

HARANG R. & BUSCALDI D. (2021). Apprentissage par transfert avec BERT pour la géolocalisation des Tweets. In *Atelier Deep Learning pour le traitement automatique des langues, EGC 2021*, Montpellier, France. HAL : hal-03166986.

HUI B., CHEN H., YAN D. & KU W.-S. (2021). Edge : Entity-diffusion gaussian ensemble for interpretable tweet geolocation prediction. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, p. 1092–1103. DOI : 10.1109/ICDE51399.2021.00099.

KRUSPE A., HÄBERLE M., HOFFMANN E. J., RODE-HASINGER S., ABDULAHHAD K. & ZHU X. X. (2021). Changes in twitter geolocations : Insights and suggestions for future usage. *W-NUT 2021 - 7th Workshop on Noisy User-Generated Text, Proceedings of the Conference*, p. 212–221. DOI : 10.48550/arxiv.2108.12251.

KUMAR S., ZHANG X. & LESKOVEC J. (2019). Predicting dynamic embedding trajectory in temporal interaction networks. **10**. DOI : 10.1145/3292500.3330895.

LAU J. H., CHI L., TRAN K.-N. & COHN T. (2017). End-to-end network for twitter geolocation prediction and hashing. DOI : 10.48550/arxiv.1710.04802.

LI M., LIM K. H., GUO T. & LIU J. (2022). A transformer-based framework for poi-level social post geolocation. DOI : 10.48550/arxiv.2211.01336.

MAHAJAN R. & MANSOTRA V. (2021). Predicting geolocation of tweets : Using combination of cnn and bilstm. *Data Science and Engineering*, **6**, 402–410. DOI : 10.1007/s41019-021-00165-1.

OZDIKIS O., RAMAMPIARO H. & NØRVÅG K. (2018). Spatial statistics of term co-occurrences for location prediction of tweets. In G. PASI, B. PIWOWARSKI, L. AZZOPARDI & A. HANBURY, Éds., *Advances in Information Retrieval*, p. 494–506, Cham : Springer International Publishing. DOI : 10.1007/978-3-319-76941-7_37.

QAZI U., IMRAN M. & OFLI F. (2020). Geocov19 : A dataset of hundreds of millions of multilingual covid-19 tweets with location information. DOI : 10.48550/ARXIV.2005.11177.

ROSSI E., CHAMBERLAIN B., FABRIZIO T., TWITTER F., TWITTER D. E., MONTI F. & TWITTER M. B. (2020). Temporal graph networks for deep learning on dynamic graphs. DOI : 10.48550/arxiv.2006.10637.

ZHANG J., DELUCIA A. & DREDZE M. (2022). Changes in tweet geolocation over time : A study with carmen 2.0. p. 1–14.

ZOHAR M. (2021). Geolocating tweets via spatial inspection of information inferred from tweet meta-fields. *International Journal of Applied Earth Observation and Geoinformation*, **105**, 102593. DOI : 10.1016/j.jag.2021.102593.

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

IR-SenTransBio: Modèles Neuronaux Siamois pour la Recherche d'Information Biomédicale

Safaa Menad[1]
(1) Univ. Rouen Normandie, LITIS UR4108, 76000, Rouen
safaa.menad1@univ-rouen.fr

RÉSUMÉ

L'entraînement de modèles transformeurs de langages sur des données biomédicales a permis d'obtenir des résultats prometteurs. Cependant, ces modèles de langage nécessitent pour chaque tâche un affinement (fine-tuning) sur des données supervisées très spécifiques qui sont peu disponibles dans le domaine biomédical. Dans le cadre de la classification d'articles scientifiques et les réponses aux questions biomédicales, nous proposons d'utiliser de nouveaux modèles neuronaux siamois (sentence transformers) qui plongent des textes à comparer dans un espace vectoriel. Nos modèles optimisent une fonction objectif d'apprentissage contrastif auto-supervisé sur des articles issus de la base de données bibliographique MEDLINE associés à leurs mots-clés MeSH (Medical Subject Headings). Les résultats obtenus sur plusieurs benchmarks montrent que les modèles proposés permettent de résoudre ces tâches sans exemples (zero-shot) et sont comparables à des modèles transformeurs biomédicaux affinés sur des données supervisés spécifiques aux problèmes traités. De plus, nous exploitons nos modèles dans la tâche de la recherche d'information biomédicale. Nous montrons que la combinaison de la méthode BM25 et de nos modèles permet d'obtenir des améliorations supplémentaires dans ce cadre.

ABSTRACT

IR-SenTransBio : Siamese Neural Networks for Biomedical Information Retrieval

Training transformers models on biomedical data has yielded promising results. However, these language models require fine-tuning on very specific supervised data for each task, which are not widely available in the biomedical domain. In the context of document classification and question answering in the biomedical domain, we propose to use new Siamese neural models (sentence transformers) that embed texts to be compared in a vector space. The proposed models optimize an objective self-supervised contrastive learning function on articles from the MEDLINE bibliographic database associated to their MeSH (Medical Subject Headings) keywords. The obtained results on several benchmarks show that the proposed models can solve these tasks without examples (zero shot) and are comparable to biomedical transformers fine-tuned on supervised data specific to the problem at hand. Moreover, our models are exploited in biomedical information retrieval task. We show that the combination of BM25 and our models improves biomedical information retrieval.

MOTS-CLÉS : Modèles de Langage · Transformeurs · Apprentissage Contrastif · Modèles Neuronaux Simaois · Apprentissage sans Exemple · Apprentissage auto-supervisé · Recherche d'Information · Classification de Documents · Réponses aux Questions · Textes Biomédicaux.

KEYWORDS: Language Models · Transformers · Contrastive Learning · Siamese Neural Networks · Zero-shot Learning · Self-supervised Learning · Information Retrieval . Document Classification ·

Actes des 16e Rencontres Jeunes Chercheurs en RI (RJCRI) et 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL), pages 131–142.

1 Introduction

Le développement de modèles transformeurs pré-entraînés, tels que BERT (Bidirectional Encoder Representations from Transformers) (Devlin *et al.*, 2019), a permis d'améliorer les performances du traitement automatique du langage (TAL). L'abondance de données biomédicales disponibles, comme les articles scientifiques, a aussi rendu possible l'entraînement de ces modèles sur des corpus de textes (p. ex. documents, dossiers cliniques de patients) pour des applications biomédicales de prédiction (Alsentzer *et al.*, 2019; Lee *et al.*, 2020; Liu *et al.*, 2021). Ces modèles de langage nécessitent cependant un affinement (fine-tuning) pour chaque tâche sur des données supervisées très spécifiques et rarement disponibles, ce qui limite fortement leur usage en pratique. Comme la plupart des tâches de TAL biomédical (p. ex. extraction de relations, classification de documents, questions-réponses) peuvent se réduire au calcul d'une mesure de similarité sémantique entre deux textes (p. ex. catégorie/résumé d'un article, requête/résultats, question/réponse), nous proposons ici de construire un nouveau modèle transformeur siamois (sentence transformeur) IR-SenTransBio (Information Retrieval-Sentence Transformer in Biomedical data) pré-entraîné qui plonge des paires de textes sémantiquement liés (longs et courts) dans un même espace de représentation vectoriel. En plus d'être applicable à plusieurs types de tâches de TAL, un modèle siamois a aussi l'avantage de permettre de gagner du temps lors de son utilisation en précalculant les représentations vectorielles des textes. Par exemple, en recherche documentaire, un modèle siamois peut permettre de précalculer et d'indexer les représentations vectorielles des textes du corpus ciblé pour n'en calculer que la représentation des requêtes lorsqu'elles sont soumises au moteur, contrairement aux modèles transformeurs "affinés" qui prennent en entrée la combinaison de toutes les paires de textes à comparer. Grâce à ce modèle, nous souhaitons : i) éviter les coûts engendrés par l'étiquetage des données, les calculs d'entraînement et d'affinement ; et ii) réduire considérablement ceux de la prédiction en proposant un modèle auto-supervisé de référence directement applicable à un large éventail de tâches biomédicales.

Dans ce contexte, nous comparons plusieurs modèles transformeurs siamois que nous avons entraînés sur des paires de textes formées, d'une part, de résumés du corpus d'articles biomédicaux PubMed [1], et d'autre part, des mots-clés MeSH (Medical Subject Headings) [2] qui leur sont associés. Nous utilisons une fonction objectif d'apprentissage contrastif auto-supervisé. Étant donnée une paire de textes (résumé, mots-clés), le modèle doit prédire laquelle, parmi un ensemble d'autres paires de textes échantillonnées au hasard, lui est réellement associée dans PubMed. Nous montrons ensuite expérimentalement sur plusieurs benchmarks biomédicaux que sans affinement pour une tâche spécifique, notre meilleur modèle siamois pré-entraîné permet, sans exemples d'apprentissage (zero shot), de classer des documents et de répondre à des questions, et cela avec des résultats comparables aux modèles transformeurs biomédicaux ou encore généralistes affinés sur des données supervisées spécifiques aux problèmes traités. En dernier lieu, nous mettons en pratique ces modèles en les exploitant pour la recherche d'information biomédicale. Pour évaluer notre approche, nous l'appliquons sur deux corpus biomédicaux, à savoir TREC-COVID (Voorhees *et al.*, 2021) et NFCorpus (Boteva *et al.*, 2016). Nous montrons aussi que la combinaison de la méthode BM25 avec nos modèles permet d'améliorer les performances de recherche d'information.

1. https://ftp.ncbi.nlm.nih.gov/pubmed/baseline/
2. Le MeSH https://www.nlm.nih.gov/mesh/ est un thésaurus spécialisé du domaine biomédical composé de 30 000 descripteurs utilisés pour l'indexation d'articles PubMed.

Cet article est structuré de la manière suivante : dans la section 2, nous proposons une revue de la littérature qui explore les performances des transformeurs dans la tâche de recherche d'information (RI). La section 3 présente en détail les modèles transformeurs pré-entraînés et leur application dans les modèles siamois. Nous décrivons nos propres modèles siamois dans la section 4. Les résultats obtenus sur des benchmarks de référence sont exposés dans la section 5. Dans la section 6, nous détaillons la tâche de recherche d'information biomédicale et analysons les performances de nos modèles dans ce domaine. Enfin, nous concluons et ouvrons des perspectives de recherche dans la section 7.

2 Travaux similaires

L'efficacité des modèles de langage dans la tâche de recherche d'information a été abordée par plusieurs études. Ces travaux se sont intéressés à évaluer les performances des modèles existants dans la recherche, la classification et la récupération d'informations pertinentes. Des approches basées sur des modèles tels que BERT, Sentence-BERT et d'autres transformeurs ont été explorées pour améliorer la précision et la pertinence des résultats.

Dans (Nguyen *et al.*, 2022), les auteurs créent un nouveau corpus en langue vietnamienne et proposent un système de question-réponse à deux étapes pour ce corpus. Le système récupère des documents pertinents en se basant sur la question d'entrée à l'aide d'un moteur de recherche basé sur le TF-IDF. Dans la deuxième étape, le système utilise un modèle S-BERT affiné pour identifier la phrase la plus pertinente dans les documents récupérés et pour extraire la réponse à la question.

Dans (Yang *et al.*, 2020), les auteurs utilisent également les bi-encodeurs pour la tâche de la RI et montrent que cette approche permet d'obtenir de très bons résultats sur différents benchmarks dans un domaine général.

(Tinn *et al.*, 2021) propose l'application des transformeurs tels que BERT et GPT sur des tâches de TAL biomédicales comme la reconnaissance d'entités nommées et l'extraction des relations. Les résultats qu'ils ont obtenu montrent l'avantage de réajustement de ces modèles sur ces tâches.

Dans (Soni & Roberts, 2020), les auteurs évaluent les performances de plusieurs modèles pré-entraînés et de modèles affinés sur une tâche de question-réponse clinique en utilisant l'ensemble de données MedQuAD. Ils comparent les performances de ces modèles à plusieurs méthodes de base et analysent l'impact de la sélection de l'ensemble de données sur les performances. Ils cherchent à identifier les ensembles de données de pré-entraînement et d'affinement les plus efficaces pour cette tâche.

Les auteurs de (Chakraborty *et al.*, 2020) proposent un modèle pré-entraîné sur le corpus BREATHE (corpus médical) pour la tâche de question-réponse. Ils montrent que l'entraînement de ces modèles sur des données biomédicales leur permettent de dépasser les modèles de base comme BERT.

3 Les Transformeurs

Les transformeurs sont des réseaux neuronaux basés sur le mécanisme d'auto-attention multi-têtes améliorant l'efficacité de l'apprentissage des modèles de grande taille. Il est composé d'un encodeur qui transforme le texte d'entrée en vecteur, et d'un décodeur qui transforme ce vecteur en texte en

sortie. Le mécanisme d'attention fournit de meilleures performances grâce à la modélisation des liens entre les éléments d'entrée et de sortie. Un modèle de langage pré-entraîné (MLP) est un réseau neuronal entraîné sur une grande quantité de données non annotées de manière non supervisée. Le modèle est ensuite transféré pour une tâche de TAL cible (downstream task), où un ensemble de données annotées plus petit et spécifique à la tâche est utilisé pour affiner le MLP permettant ainsi de construire le modèle final capable d'exécuter la tâche cible (ajustement d'un MLP).

3.1 Modèles pré-entraînés

Les modèles de langage pré-entraînés, tels que BERT, ont conduit à des gains impressionnants dans de nombreuses tâches de TAL. Les travaux existants traitent des données généralistes. Dans les tâches de TAL biomédicales, le pré-entraînement sur les textes de PubMed par exemple a permis d'obtenir de meilleures performances (Beltagy *et al.*, 2019; Lee *et al.*, 2020; Peng *et al.*, 2019a). L'approche standard de pré-entraînement d'un modèle biomédical débute avec un modèle généraliste et poursuit le pré-entraînement en utilisant un corpus biomédical. Par exemple, BioBERT (Lee *et al.*, 2020) utilise pour cela les résumés extraits de PubMed et les articles en texte intégral de PubMed Central (PMC). BlueBERT (Peng *et al.*, 2019b) utilise à la fois le texte de PubMed et les notes cliniques MIMIC-III (Medical Information Mart for Intensive Care) (Johnson *et al.*, 2016). SciBERT (Beltagy *et al.*, 2019) constitue une exception, le pré-entraînement est réalisé à partir de zéro, en utilisant la littérature scientifique.

3.2 Modèles siamois

Les transformeurs de paires de phrases (sentence-transformers) sont des modèles développés pour la tâche de calcul d'un score de similarité entre deux phrases (p. ex.calcul de similarité sémantique entre phrases, recherche d'informations, reformulation de phrases etc). Ces transformeurs sont basés sur deux architectures : i) les cross-encodeurs, qui traitent la concaténation de la paire ; et ii) les modèles siamois bi-encodeurs, qui encodent en vecteur chacun des éléments de la paire. Par exemple, Sentence-BERT (Reimers & Gurevych, 2019) est un bi-encodeur basé sur BERT permettant de générer des plongements de phrases sémantiquement significatifs à utiliser dans des comparaisons de similarité textuelle. Pour chaque entrée (voir figure 1), le modèle produit un vecteur de taille fixe (u et v). La fonction objectif est choisie de façon à ce que l'angle entre les deux vecteurs u et v soit d'autant plus faible que les entrées sont similaires. Plus précisément, la fonction objectif utilise le cosinus de l'angle : $cos(u, v) = \frac{u.v}{||u||||v||}$, si $cos(u, v) = 1$ alors les phrases sont similaires et si $cos(u, v) = 0$ alors les phrases n'ont aucune relation sémantique. D'autres modèles de plongement de phrases ont été développés (Gao *et al.*, 2021; Wang *et al.*, 2021; Cohan *et al.*, 2020). Parmi eux MiniLM-L6-v25 [3] qui est un bi-encodeur basé sur une version simplifiée de MiniLM (Wang *et al.*, 2020). Ce modèle, rapide et de petite taille, a permis d'obtenir de bonnes performances sur différentes tâches pour 56 corpus (Muennighoff *et al.*, 2022).

3. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

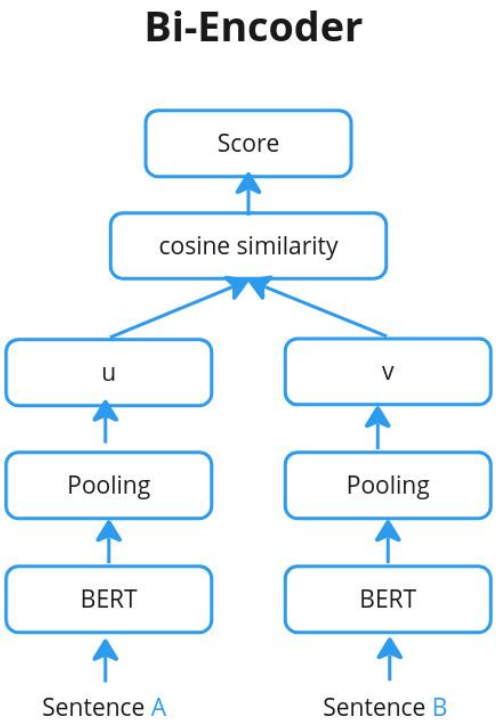


FIGURE 1 – Un encodeur siamois.

4 Modèles de Langage Proposés

Les transformeurs siamois donnent de bons résultats dans des domaines généralistes, mais pas dans les domaines de spécialité, comme le domaine biomédical (Muennighoff *et al.*, 2022). Nous proposons ici de nouveaux modèles siamois pré-entraînés sur le corpus PubMed. Les transformeurs siamois ont été initialement conçus pour transformer des phrases (de taille similaire) en vecteurs. Nous proposons dans notre approche de transformer dans le même espace vectoriel les termes MeSH, les titres et les résumés des articles PubMed en entraînant un modèle de transformeur siamois sur ces données que nous avons préparées. Nous voulons nous assurer qu'il y a une correspondance dans cet espace vectoriel entre le texte court et le texte long. Nous avons donc entraîné nos modèles avec des paires d'entrées (titre, terme MeSH) et (résumé, terme MeSH). À partir de ces données, nous avons construit deux types de modèles : le premier type est nos propres transformeurs siamois (BioSTransformers) construits à partir d'un transformeur pré-entraîné sur des données biomédicales et le second est un transformeur siamois déjà pré-entraîné sur des données généralistes (BioS-MiniLM).

BioSTransformers. Pour ce type, nous nous sommes inspirés du modèle Sentence-BERT (Reimers & Gurevych, 2019) en remplaçant BERT par d'autres transformeurs. Nous avons utilisé des transformeurs qui ont été entraînés sur des données biomédicales (bio-transformeurs). Pour construire les transformeurs siamois (BioSTransformers), nous avons ajouté une couche de pooling et modifié la fonction objectif. Ensuite, nous les avons entraîné sur nos données. La couche de pooling calcule le vecteur moyen des vecteurs de sortie du transformeur (token embeddings). Les deux textes en entrée passent successivement dans le transformeur produisant deux vecteurs u et v en sortie du pooling, qui sont par la suite utilisées par la fonction objectif. Parmi les bio-transformeurs disponibles, nous avons

sélectionné les modèles les plus performants : BlueBERT (Peng *et al.*, 2019b), PubMedBERT (Gu *et al.*, 2022) et BioELECTRA (Kanakarajan *et al.*, 2021). Ces modèles ont été entraînés sur PubMed, à part BlueBERT qui a également été entraîné sur des notes cliniques.

BioS-MiniLM. Pour ce modèle nous avons adapté un transformeur siamois pré-entraîné sur des données généralistes. Plusieurs modèles généraux de sentence-transformer déjà pré-entraînés sont disponibles [4]. Ils diffèrent en taille, vitesse et performance. Parmi ceux qui obtiennent les meilleures performances, nous avons affiné MiniLM-L6-v2 (voir section 3) qui a été pré-entraîné sur 32 corpus généralistes (Reddit comments, S2ORC, WikiAnswers etc.).

Fonction objectif. Pour un transformeur de paires de phrases classique on dispose de données supervisées sous forme de triplets (phrase 1, phrase 2, score de similarité entre les deux phrases). Dans notre cas cependant, nous ne disposons d'aucun score pour les résumés ou les titres et leurs termes MeSH correspondants. Nous considérons donc qu'un résumé, un titre et les termes MeSH associés au même article (identifié par un PMID) sont similaires (le score est égal à 1) et inversement, qu'un résumé ou un titre avec des termes MeSH qui ne sont pas associés au même article ne sont pas similaires (le score est égal à 0).

Nous utilisons une fonction objectif d'apprentissage contrastif auto-supervisé basée sur la fonction de perte de classement négatif multiple (Henderson *et al.*, 2017) dite MNRL (Multiple Negative Ranking Loss) dans le package Sentence-Transformers [5]. Cette fonction permet à un modèle d'apprendre à partir de données non étiquetées en utilisant la comparaison de paires d'exemples similaires et différents. Elle vise à maximiser la similarité entre les représentations de deux exemples similaires et à minimiser la similarité entre les représentations de deux exemples différents. La MNRL n'a besoin que des paires positives en entrée (le titre ou le résumé et un terme MeSH associé à l'article dans notre cas). Pour une paire positive (titre_i ou résumé_i, MeSH_i), la MNRL considère que chaque paire (titre_i ou résumé_i, MeSH_j) avec $i \neq j$ dans le même batch est négative. Comme un article peut être associé à plusieurs termes MeSH, nous avons fait en sorte dans la génération des batchs qu'un résumé (ou un titre) associé à un terme MeSH dans PubMed ne soient jamais considérés comme une paire négative.

5 Expérimentations et Résultats

5.1 Expérimentations

Dans un premier temps, pour tester les différents transformeurs et la fonction objectif à retenir, nous n'avons utilisé que les titres et nous avons réduit le nombre de termes MeSH. Au total 1 402 termes MeSH et 3,79 millions de paires (titre, MeSH) ont été sélectionnées. Pour la validation nous avons utilisé 18 940 articles avec leurs titres et termes MeSH.

Dans un second temps, une fois sélectionnés les modèles transformeurs et la fonction objectif MNRL, nous avons évalué nos modèles BioSTransformers et BioS-MiniLM sur les paires (titre, MeSH) et (résumé, MeSH) générés à partir de tous les termes MeSH de PubMed. Ayant constaté qu'il n'était pas nécessaire d'utiliser toutes les paires des 35 millions d'articles de PubMed, nous avons sélectionné

4. https://huggingface.co/sentence-transformers
5. https://www.sbert.net/docs/package_reference/losses.html#multiplenegativesrankingloss

6,75 millions de paires pour le fine-tuning. Un total de 18 557 articles sert à la validation.

Les deux tâches de TAL, ainsi que les données utilisées sont décrites ci-après :

1. La classification de documents : le corpus Hallmarks of Cancer (HoC) est constitué de 1 852 résumés de publications PubMed annotés manuellement par des experts selon une taxonomie qui est composée de 37 classes. Chaque phrase du corpus se voit attribuer zéro à plusieurs classes (Hanahan & Weinberg, 2000) ;
2. Questions-réponses (QA) :
   (a) PubMedQA est un corpus pour les réponses aux questions spécifiques à la recherche biomédicale. Il contient un ensemble de questions, ainsi qu'un champ annoté indiquant si le texte contient la réponse à la question de recherche (Jin *et al.*, 2019) ;
   (b) BioASQ est un corpus qui contient plusieurs tâches de QA avec des données annotées par des experts, y compris des questions oui/non, de liste et de résumés. Nous nous concentrons sur le type de questions oui/non (tâche 7b) (Nentidis *et al.*, 2019).

Nous considérons les deux tâches précédentes comme un problème de similarité de textes : pour chaque requête nous considérons les k résultats les plus proches, k étant le nombre de résultats attribués à la requête par l'expert (gold standard). La similarité entre la requête et les résultats est mesurée par la similarité cosinus entre le vecteur de la requête et les vecteurs des résultats. Dans une tâche de classification, la requête est la catégorie et les résultats sont les documents classés dans cette catégorie. Dans une tâche de questions-réponses, la requête est la question et les résultats sont une réponse.

5.2 Résultats

Nos modèles sont évalués selon le score F1 utilisé dans les benchmarks HoC, PubmedQA et BioASQ dans (Gu *et al.*, 2022). Les résultats obtenus par nos modèles transformeurs siamois sans exemple (sans fine-tuning) sont synthétisés dans le Tableau 1, avec en gras les meilleurs scores.

Corpus/modèle	BioS-MiniLM	S-BioELECTRA	S-PubMedBERT	S-BlueBERT
HoC	0,492	**0,499**	0,489	0,468
PubMedQA	0,649	0,675	**0,729**	0,652
BioASQ	0,747	0,694	**0,751**	0,713

TABLE 1 – Résultats d'évaluation de nos modèles sur différents benchmarks selon le F1 score.

Le Tableau 2 résume les résultats obtenus sur les mêmes tâches par des modèles affinés spécifiquement à ces tâches (Gu *et al.*, 2022). Pour chaque benchmark, ces modèles sont affinés avec les données supervisées disponibles dans chaque cas. Ces résultats montrent que les modèles que nous proposons permettent de réaliser ces tâches avec des résultats comparables à des modèles biomédicaux affinés sur des données supervisés spécifiques aux problèmes traités, mais que nous n'avons pas utilisés dans notre approche sans exemple.

Pour le benchmark HoC, les résultats obtenus par notre meilleur modèle S-BioELECTRA sont très en dessous des résultats obtenus par PubMedBERT+aff (0,499 vs. 0,823). En effet, les modèles de (Gu *et al.*, 2022) ont été affinés spécifiquement pour chaque tâche, notamment la classification des documents, en modifiant l'architecture du modèle et en ajoutant des couches spécifiques pour

Corpus/modèle	BERT +aff	RoBERTa +aff	BioBERT +aff	SciBERT +aff	ClinicalBERT +aff	BlueBERT +aff	PubMedBERT +aff
HoC	0,802	0,797	0,815	0,812	0,807	0,805	**0,823**
PubmedQA	0,516	0,528	**0,602**	0,574	0,491	0,484	0,558
BioASQ	0,744	0,752	0,841	0,789	0,685	0,687	**0,876**

TABLE 2 – Résultats d'évaluation des modèles affinés (+aff) spécifiquement à ces tâches sur différents benchmarks selon le F1 score.

chaque cas. En revanche, pour le benchmark PubMedQA, les résultats obtenus par notre meilleur modèle S-PubMedBERT dépassent les résultats obtenus par BioBERT+aff (0,729 vs. 0,602). Enfin, pour le benchmark BioASQ, les résultats obtenus par notre meilleur modèle S-PubMedBERT sont comparables aux résultats obtenus par les modèles affinés même si PubMedBERT+aff donne de meilleurs résultats (0,751 vs. 0,876). Tout cela a été obtenu sans réadapter l'architecture de nos modèles pour chaque tâche et sans les affiner sur les données spécifiques aux benchmarks cités.

Les modèles BioSTransformers obtiennent de meilleurs résultats que le BioS-MiniLM, cela s'explique par le fait que le modèle BioS-MiniLM a été pré-entraîné sur des données généralistes, tandis que les autres modèles ont été pré-entraînés sur des données biomédicales spécialisées. Cela démontre l'importance de la phase de pré-entraînement.

6 Recherche d'Information Biomédicale

Après avoir évalué nos modèles sur deux tâches de NLP dans le domaine biomédical, dans cette section, nous allons les appliquer sur la tâche de la RI. Bien que les méthodes neuronales ont surpassé les approches traditionnelles de la RI telles que TF-IDF (Term Frequency - Inverse Document Frequency) et BM25 dans des domaines généralistes, elles demeurent cependant insuffisantes dans le domaine biomédical. Dans cette partie, nous proposons d'améliorer la recherche d'informations biomédicale avec nos modèles siamois proposés.

Expérimentations

Nous considérons la tâche de la RI comme une recherche de proximité entre les représentations d'une requête et des documents. Nous utilisons nos modèles pour plonger les deux entrées et calculer le score de similarité. Nous testons nos modèles sur deux corpus biomédicaux TREC-COVID (Voorhees *et al.*, 2021) et NFCorpus (Boteva *et al.*, 2016).

TREC-COVID. Le jeu de données TREC-COVID est une collection d'articles scientifiques liés à la COVID-19. Il a été créé dans le cadre de la tâche de RI de la conférence Text REtrieval Conference (TREC) COVID, qui visait à soutenir la communauté scientifique dans sa réponse à la pandémie de COVID-19 par des systèmes de RI efficaces. Il est composé de 171 000 articles et de 50 requêtes.

NFCorpus. NFCorpus est un ensemble de données d'extraction de texte intégral en anglais pour la RI biomédicale qui concerne la nutrition. Il contient un total de 3 244 requêtes en langage naturel

(extraites du site NutritionFacts.org) avec 169 756 jugements de pertinence extraits automatiquement pour 9 964 documents médicaux provenant principalement de PubMed.

Corpus/modèle	BM25	BioS-MiniLM	BM25+BioS-MiniLM	BM25+S-BERT
TREC-COVID	0,616	0,478	0,616	**0,656**
NFCorpus	**0,300**	0,282	0,283	0,262

TABLE 3 – Résultats d'évaluation de nos modèles sur les données biomédicales selon NDCG@10.

Le Tableau 3 recense les résultats obtenus avec nos modèles et des modèles de base selon la métrique d'évaluation NDCG@10 (Wang *et al.*, 2013).

Il Compare les performances de nos modèles avec le modèle BM25+S-BERT utilisé comme un modèle de reclassement qui donne les meilleurs résultats avec TREC-COVID dans le benchmark BEIR [24]. Dans BM25+S-BERT et BM25+BioS-MiniLM, les 100 meilleurs résultats donnés par la fonction de classement BM25 sont réordonnés en utilisant le modèle S-BERT [6].

D'après le tableau nous pouvons déduire que BM25 (baseline) obtient de bons résultats sur les deux jeux de données biomédicales. Cependant, cette méthode se base sur une recherche lexicale qui dépend de la fréquence des termes et ne capte pas la sémantique des phrases. Notre modèle ne dépasse pas la BM25. Cela peut s'expliquer par le fait que notre modèle est entraîné sur la récupération de documents à la base de mots clés très spécifiques (MeSH) et non pas de requêtes en langage naturel général. Cependant, sur le corpus TREC-COVID, BM25 et notre méthode peuvent se compléter : leur utilisation conjointe permet de légèrement améliorer les résultats et de bénéficier de l'avantage des deux méthodes. BM25+S-BERT obtient mieux car le modèle S-BERT a été pré-entraîné sur le corpus MSMARCO [7], un corpus volumineux spécialement conçu pour l'entraînement des modèles dédiés à la tâche de recherche d'information. Contrairement à notre modèle qui a été pré-entraîné sur un corpus généraliste qui n'est pas dédié à la RI. Le principal avantage de notre méthode est qu'elle permet de calculer les représentations des documents et de les indexer auparavant. Le modèle calcule la représentation de la requête soumise au moteur et récupère directement les documents les plus pertinents dans un temps réduit.

7 Conclusion

Dans cette étude, nous avons proposé de nouveaux modèles neuronaux siamois pour créer des représentations vectorielles de textes. Ces représentations sont ensuite utilisées pour calculer la similarité entre les textes. Nous avons testé nos modèles sur les deux tâches de classification et de question-réponse dans le domaine biomédical et nous avons montré qu'ils permettaient d'obtenir des résultats comparables à ceux des modèles qui ont été spécifiquement conçus pour ces tâches et entraînés sur des données correspondantes. Nous avons ensuite appliqué nos modèles sur la tâche de recherche d'information biomédicale, en utilisant deux corpus couramment utilisés dans ce domaine. Nous avons montré que l'utilisation de nos modèles en combinaison avec une méthode lexicale traditionnelle BM25 conduit à de meilleurs résultats et permet de bénéficier de l'avantage des deux types de méthodes. Ces résultats comparables et encourageants nous permettent d'envisager d'étendre notre approche en entraînant nos modèles sur des données biomédicales contenant des mots clés différents

6. sentence-transformers/msmarco-distilbert-base-v4
7. https ://github.com/microsoft/MSMARCO-Passage-Ranking

de ceux actuellement utilisés. Nous prévoyons également d'ajuster nos modèles spécifiquement pour la tâche de RI en incluant des étapes supplémentaires dans le processus d'entraînement. Enfin, nous étendrons la RI à d'autres données et à la recherche de dossiers patients pour permettre la constitution de cohortes dans les études cliniques.

Remerciements

Ce travail de thèse est fait sous la supervision de Fatima Lina SOUALMIA (TIBS, LITIS, Université de Rouen Normandie) et Saïd ABDEDDAIM (TIBS, LITIS, Université de Rouen Normandie).

Références

ALSENTZER E., MURPHY J., BOAG W., WENG W.-H., JINDI D., NAUMANN T. & MCDERMOTT M. (2019). Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, p. 72–78, Minneapolis, Minnesota, USA : Association for Computational Linguistics. DOI : 10.18653/v1/W19-1909.

BELTAGY I., LO K. & COHAN A. (2019). SciBERT : A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 3615–3620.

BOTEVA V., GHOLIPOUR D., SOKOLOV A. & RIEZLER S. (2016). A full-text learning to rank dataset for medical information retrieval. In *Advances in Information Retrieval : 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20–23, 2016. Proceedings 38*, p. 716–722 : Springer.

CHAKRABORTY S., BISONG E., BHATT S., WAGNER T., ELLIOTT R. & MOSCONI F. (2020). Biomedbert : A pre-trained biomedical language model for qa and ir. In *Proceedings of the 28th International Conference on Computational Linguistics*, p. 669–679.

COHAN A., FELDMAN S., BELTAGY I., DOWNEY D. & WELD D. S. (2020). Specter : Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 2270–2282.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, p. 4171–4186.

GAO T., YAO X. & CHEN D. (2021). Simcse : Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 6894–6910.

GU Y., TINN R., CHENG H., LUCAS M., USUYAMA N., LIU X., NAUMANN T., GAO J. & POON H. (2022). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, **3**(1), 1–23. DOI : 10.1145/3458754.

HANAHAN D. & WEINBERG R. A. (2000). The hallmarks of cancer. *Cell*, **100**(1), 57–70.

HENDERSON M., AL-RFOU R., STROPE B., SUNG Y.-H., LUKÁCS L., GUO R., KUMAR S., MIKLOS B. & KURZWEIL R. (2017). Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv :1705.00652*.

JIN Q., DHINGRA B., LIU Z., COHEN W. & LU X. (2019). PubMedQA : A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 2567–2577.

JOHNSON A. E., POLLARD T. J., SHEN L., LEHMAN L.-W. H., FENG M., GHASSEMI M., MOODY B., SZOLOVITS P., ANTHONY CELI L. & MARK R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific data*, **3**(1), 1–9.

KANAKARAJAN K. R., KUNDUMANI B. & SANKARASUBBU M. (2021). BioELECTRA : Pretrained biomedical text encoder using discriminators. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, p. 143–154, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.bionlp-1.16.

LEE J., YOON W., KIM S., KIM D., KIM S., SO C. H. & KANG J. (2020). BioBERT : a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, **36**(4), 1234–1240.

LIU F., SHAREGHI E., MENG Z., BASALDELLA M. & COLLIER N. (2021). Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 4228–4238.

MUENNIGHOFF N., TAZI N., MAGNE L. & REIMERS N. (2022). Mteb : Massive text embedding benchmark. *arXiv preprint arXiv :2210.07316*.

NENTIDIS A., BOUGIATIOTIS K., KRITHARA A. & PALIOURAS G. (2019). Results of the seventh edition of the BioASQ challenge. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, p. 553–568 : Springer.

NGUYEN N. T.-H., HA P. P.-D., NGUYEN L. T., VAN NGUYEN K. & NGUYEN N. L.-T. (2022). Spbertqa : A two-stage question answering system based on sentence transformers for medical texts. In *Knowledge Science, Engineering and Management : 15th International Conference, KSEM 2022, Singapore, August 6–8, 2022, Proceedings, Part II*, p. 371–382 : Springer.

PENG Y., YAN S. & LU Z. (2019a). Transfer learning in biomedical natural language processing : An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, p. 58–65.

PENG Y., YAN S. & LU Z. (2019b). Transfer learning in biomedical natural language processing : An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, p. 58–65, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/W19-5006.

REIMERS N. & GUREVYCH I. (2019). Sentence-BERT : Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 3982–3992, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1410.

SONI S. & ROBERTS K. (2020). Evaluation of dataset selection for pre-training and fine-tuning transformer language models for clinical question answering. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 5532–5538.

TINN R., CHENG H., GU Y., USUYAMA N., LIU X., NAUMANN T., GAO J. & POON H. (2021). Fine-tuning large neural language models for biomedical natural language processing. *arXiv preprint arXiv :2112.07869*.

VOORHEES E., ALAM T., BEDRICK S., DEMNER-FUSHMAN D., HERSH W. R., LO K., ROBERTS K., SOBOROFF I. & WANG L. L. (2021). Trec-covid : constructing a pandemic information retrieval test collection. In *ACM SIGIR Forum*, volume 54, p. 1–12 : ACM New York, NY, USA.

WANG K., REIMERS N. & GUREVYCH I. (2021). Tsdae : Using transformer-based sequential denoising auto-encoderfor unsupervised sentence embedding learning. In *Findings of the Association for Computational Linguistics : EMNLP 2021*, p. 671–688.

WANG W., WEI F., DONG L., BAO H., YANG N. & ZHOU M. (2020). Minilm : Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, **33**, 5776–5788.

WANG Y., WANG L., LI Y., HE D. & LIU T.-Y. (2013). A theoretical analysis of ndcg type ranking measures. In *Conference on learning theory*, p. 25–54 : PMLR.

YANG Y., CER D., AHMAD A., GUO M., LAW J., CONSTANT N., ABREGO G. H., YUAN S., TAR C., SUNG Y.-H. *et al.* (2020). Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, p. 87–94.

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

L'évaluation de la traduction automatique du caractère au document : un état de l'art

Mariam Nakhlé [1, 2]
(1) Univ. Grenoble Alpes, CNRS, Grenoble INP*, LIG, 38000 Grenoble, France
(2) Lingua Custodia, France
mariam.nakhle@univ-grenoble-alpes.fr,
mariam.nakhle@linguacustodia.com

RÉSUMÉ

Ces dernières années l'évaluation de la traduction automatique, qu'elle soit humaine ou automatique, a rencontré des difficultés. Face aux importantes avancées en matière de traduction automatique neuronale, l'évaluation s'est montrée peu fiable. De nombreuses nouvelles approches ont été proposées pour améliorer les protocoles d'évaluation. L'objectif de ce travail est de proposer une vue d'ensemble sur l'état global de l'évaluation de la Traduction Automatique (TA). Nous commencerons par exposer les approches d'évaluation humaine, ensuite nous présenterons les méthodes d'évaluation automatiques tout en différenciant entre les familles d'approches (métriques superficielles et apprises) et nous prêterons une attention particulière à l'évaluation au niveau du document qui prend compte du contexte. Pour terminer, nous nous concentrerons sur la méta-évaluation des méthodes.

ABSTRACT

The evaluation of machine translation from character to document : state of the art

The human and automatic evaluation of machine translation have undergone great disruption in recent years. In the face of significant advances in neural machine translation, evaluation has shown to be unreliable. Multiple new approaches have been proposed to improve evaluation protocols. The aim of this work is to provide an overview of the global state of Machine Translation evaluation. We will start by outlining the approaches and methods used in human evaluation, next, we will present automatic evaluation methods while distinguishing between families of approaches (string-based and learned metrics) and we will pay particular attention to context-sensitive document-level evaluation. Finally, we will focus on the meta-evaluation of the methods.

MOTS-CLÉS : évaluation de la traduction automatique, traduction automatique, traduction automatique au niveau du document.

KEYWORDS: machine translation evaluation, machine translation, document-level machine translation.

1 Introduction

Nous présentons un état de l'art de l'évaluation de la traduction automatique (TA), en particulier dans le contexte de l'arrivée des approches neuronales dans la TA. La traduction automatique a connu de grands progrès ces dernières années, notamment suite à l'introduction du modèle Transformer

∗. Institute of Engineering Univ. Grenoble Alpes

Actes des 16e Rencontres Jeunes Chercheurs en RI (RJCRI) et 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL), pages 143–159.

(Vaswani *et al.*, 2017), ce qui a permis une grande amélioration de la qualité des moteurs de traduction. Face à cette nouvelle génération de modèles de traduction, l'évaluation de la TA manque de robustesse (Bojar *et al.*, 2018).

Le score BLEU (Papineni *et al.*, 2002), la méthode d'évaluation automatique la plus connue, a été conçue pendant l'ère de la traduction statistique pour un usage avec de multiples phrases de référence. Avec l'arrivée des approches neuronales, ce score (de la même manière que les approches similaires) s'est montré peu fiable face aux nouveaux systèmes. Le score BLEU était controversé même avant l'arrivée de la TA neuronale et cette nouvelle génération de systèmes n'a fait qu'aggraver les critiques de la métrique. Les nouveaux systèmes ont un comportement - forces et faiblesses - très différent des systèmes statistiques du début du siècle. En effet, ils sont capables de formuler des phrases de plus en plus naturelles, employer des synonymes divers et des constructions syntaxiques variées, c'est pourquoi une simple comparaison de la traduction automatique avec la traduction de référence (comme le fait BLEU) n'est plus suffisante. Le développement des systèmes de traduction capables de traduire des segments plus longs que la phrase requiert une évaluation qui va au-delà de la phrase également. De nouvelles approches sont donc nécessaires.

La conférence annuelle sur la TA, appelée WMT, essaie de fournir une réponse à ce besoin en organisant chaque année une tâche commune (*shared-task*) sur les métriques automatiques (Freitag *et al.*, 2022, 2021b; Mathur *et al.*, 2020b). Ainsi, de nombreuses nouvelles métriques ont émergé. Face à cette pluralité de nouvelles approches, la communauté fait face un nouveau défi : quelle est la plus performante et comment mesurer cette performance ? Plusieurs travaux se sont proposés de comparer leur performance et fiabilité (Mathur *et al.*, 2020a; Kocmi *et al.*, 2021; Freitag *et al.*, 2022). Selon les résultats les plus récents, la méthode qui semble montrer les meilleures performances est le score COMET (Rei *et al.*, 2020), qui est en train de devenir le nouveau score standard.

Une autre famille d'approches propose de répondre au problème du contexte dans l'évaluation. Vu qu'il a été démontré que des erreurs de traduction passent inaperçues sans le contexte, ces méthodes ciblent l'évaluation sur les phénomènes linguistiques, préalablement identifiés comme problématiques, et dont la traduction correcte dépend du contexte. L'évaluation se fait en utilisant des ensembles de test spécialement conçus pour la tâche. Ces ensembles s'appellent des *test-suites* et ils servent à étudier la performance des modèles de traduction sur le phénomène linguistique ciblé. Généralement un taux de traductions correctes est calculé, qui sert de mesure d'évaluation complémentaire, surtout pour l'évaluation des systèmes de TA sensibles au contexte. Il s'agit par exemple de la traduction des pronoms anaphoriques ou des mots polysémiques (Vojtěchová *et al.*, 2019; Rios *et al.*, 2018; Müller *et al.*, 2018).

Devant cette pluralité d'approches d'évaluation, il peut être difficile de choisir la plus adaptée. Plusieurs méthodes ont été essayées pendant des années et de nouvelles approches émergent tous les ans. Comme une évaluation solide est la base du développement, nous proposons d'offrir une vue globale sur l'état de l'art du domaine, en présentant les différentes approches étudiées et une réflexion sur les zones d'améliorations. Nous organisons l'article en trois parties : dans la section 2 nous décrivons les méthodes d'évaluation humaine, dans la section 3 nous présentons les méthodes d'évaluation automatique en les classant selon le type d'approche. Nous faisons la distinction entre les méthodes qui évaluent les phrases isolées (sous-section 3.1) et les approches qui évaluent les phrases dans leur contexte (sous-section 3.2). Dans la sous-section 3.3, nous montrons comment évaluer les métriques automatiques et finalement dans la section 4 nous présentons une discussion sur le sujet.

2 Évaluation humaine

Déterminer la qualité de la traduction est une tâche compliquée qui n'a pas de réponse unique. Contrairement à certaines tâches du traitement automatique des langues où l'évaluation est plutôt simple à effectuer (comme par exemple la reconnaissance vocale), il est difficile de déterminer la qualité de la traduction automatique et cela même pour les humains. Plusieurs traductions d'une seule phrase peuvent être correctes et pour les évaluateurs (surtout les non-professionnels) il est compliqué de distinguer la meilleure. Plusieurs aspects jouent un rôle sur le choix : objectif, style, contexte, domaine, etc. du texte à traduire. Pourtant un protocole d'évaluation humaine clair est crucial pour l'obtention de résultats fiables et par conséquent pour le développement de la TA. Dans la recherche du meilleur protocole d'évaluation, plusieurs approches ont été essayées et étudiées :

- **Mesures d'effort de post-édition** : Ces approches mesurent l'effort nécessaire pour post-éditer une traduction automatique jusqu'à ce qu'elle soit correcte. Typiquement trois dimensions sont considérées : le temps passé, l'effort technique (nombre de frappes nécessaires) et l'effort cognitif (Koponen, 2016). De nos jours, cette piste n'est plus beaucoup étudiée.
- **Évaluation directe** : Il s'agit de l'attribution d'un score d'appréciation, typiquement sur l'échelle de Likert allant de 1 à 5 (où 1=le pire score et 5=le score parfait) (Bojar *et al.*, 2016), ou bien sur une échelle continue (Graham *et al.*, 2013). Ce score est unique pour toute la phrase, il a comme objectif d'englober tous les aspects pertinents (qui peuvent varier selon le cas d'usage) tels que la fluidité, la précision, l'exactitude terminologique, le style, etc. C'est la métrique utilisée dans la conférence WMT à partir de 2017 jusqu'à 2021. Le désavantage de cette technique est que les valeurs d'accord inter-annotateurs sont souvent basses. Ce type d'évaluation est appelée en anglais *direct assessment*.
  - **Scores basés sur la fluidité et l'adéquation** : Cela est une sous-catégorie de l'évaluation directe. Les évaluateurs donnent un score pour la fluidité (*fluency*) de la traduction et un autre pour l'adéquation (*adequacy*). Le score attribué l'est généralement sur une échelle de Likert de 1 à 5. Cette approche a été utilisée à WMT en 2006 et 2007.
- **Classement comparatif** : Il s'agit de comparer deux ou plusieurs sorties candidates. Les évaluateurs classent les traductions de la meilleure à la pire (classer deux sorties comme ayant le même niveau est d'habitude possible). Cette métrique a été utilisée à WMT de 2007 à 2016. Elle a comme désavantage la croissance du nombre de phrases à lire, surtout en comparant plusieurs systèmes. Un autre désavantage de cette approche est que le score reste relatif et ne peut pas être interprété en valeur absolue (Bojar *et al.*, 2016). Ce type d'évaluation est appelé en anglais *relative ranking*.
- **Annotation de type *Multidimensional Quality Metrics* (MQM)** : Introduit par Lommel *et al.* (2014) et popularisé par Freitag *et al.* (2021a), cette approche consiste à annoter les erreurs dans le texte traduit. Chaque erreur est liée à un poids selon sa gravité. Le score final est calculé sur la base des erreurs repérées. L'avantage de cette approche est qu'elle permet de définir un ensemble d'erreurs ainsi que leurs poids selon le cas d'usage, ce qui rend l'évaluation plus spécifique puisqu'elle prend en compte les critères concrets du domaine et de l'application. Elle était utilisée à WMT en 2021 et 2022.

L'évaluation humaine est considérée la plus fiable. Cependant, durant ces dernières années, il s'est avéré que rien que le fait d'effectuer une évaluation avec des évaluateurs humains n'est pas la garantie d'une évaluation de qualité. En effet, elle est souvent subjective et de mauvais protocoles d'évaluation peuvent la rendre peu fiable. L'annonce de la parité de la traduction automatique et humaine de

(Hassan *et al.*, 2018) a généré un grand débat concernant l'évaluation humaine de la TA. De nombreux auteurs ont répondu en démontrant que cette annonce n'était qu'une fausse conclusion liée aux défauts du protocole d'évaluation humaine (Läubli *et al.*, 2018; Toral *et al.*, 2018; Graham *et al.*, 2019). Lors de l'évaluation humaine, les évaluateurs n'avaient pas accès au contexte des phrases à évaluer, ce qui a rendu certaines erreurs impossibles à détecter (Läubli *et al.*, 2018; Freitag *et al.*, 2021a). Également, certaines phrases utilisées comme phrases sources dans le jeu de test étaient elles-mêmes des traductions, ce qui rend la traduction plus facile dû à l'effet « translationese » (Graham *et al.*, 2019). Dans le contexte de nouvelles approches neuronales où les modèles devenaient de plus en plus robustes, cette évaluation n'était plus suffisante et les organisateurs de WMT ont annoncé qu'il faudrait trouver de nouveaux moyens pour évaluer (Bojar *et al.*, 2018). À partir de 2021, la conférence a changé de protocole d'évaluation et a adopté l'approche d'annotation MQM telle que proposée dans Freitag *et al.* (2021a).

L'idée principale derrière cette approche est qu'implicitement, l'attribution de score par un évaluateur se fait à travers l'analyse des erreurs dans la traduction. C'est pourquoi les auteurs proposent de rendre explicite cette analyse lors de l'évaluation grâce à l'annotation MQM. Ils ont rendu publiques des annotations MQM des jeux de test issus de la conférence WMT pour permettre de futurs travaux [1]. Certaines implémentations de l'évaluation MQM introduisent des arbres de décision pour guider l'évaluateur, rendant ainsi l'évaluation un processus plus logique et conscient, ce qui permet d'obtenir des résultats plus standardisés et moins subjectifs.

Dans un effort de compréhension de la significativité statistique de l'évaluation humaine, Wei *et al.* (2022) ont analysé un ensemble de 1728 campagnes d'évaluation provenant de l'évaluation interne chez Microsoft (Kocmi *et al.*, 2021) et sont arrivés à la conclusion que dans la majorité des cas, les résultats ne sont pas statistiquement significatifs, c'est-à-dire que l'évaluation ne permet pas de déterminer si un système est plus performant qu'un autre. Ils concluent que la différence entre les modèles étant petite, les échantillons évalués sont insuffisants et un nombre d'échantillons (et donc un budget) beaucoup plus élevé serait nécessaire pour aboutir à une évaluation statistiquement significative. Nous pouvons nous poser la question si une telle analyse statistique sur l'évaluation MQM donnerait des conclusions similaires. Freitag *et al.* (2021b) ont analysé les annotations MQM et ont trouvé que malgré l'accord inter-annotateur plutôt bas, les classements finaux des systèmes restent cohérents parmi les évaluateurs.

Nous voyons alors que l'évaluation humaine a quelques défauts. Ceci est d'autant plus problématique vu que, comme nous allons le présenter dans la section suivante, l'évaluation humaine sert de référence de qualité (*gold standard*) pour évaluer et, selon les cas, entraîner les approches automatiques. Dans la section suivante, nous présentons les approches d'évaluation automatique.

3 Évaluation automatique

Une grande partie de l'évaluation de la TA est faite par des méthodes automatiques, c'est-à-dire qui ne nécessitent pas d'évaluateurs humains pour obtenir un jugement sur la qualité du système de traduction. Ces méthodes sont souvent appelées métriques, pourtant cela n'est pas dans le sens mathématique du terme. Leur avantage principal est leur coût réduit (surtout grâce aux ensembles de tests publics et par rapport à l'évaluation humaine où les évaluateurs sont payés) et leur rapidité. Ces méthodes sont largement utilisées pendant la phase de développement des systèmes. Nous allons

1. github.com/google/wmt-mqm-human-evaluation

diviser les approches en deux catégories : (1) les méthodes qui évaluent la traduction isolée de son contexte et (2) les méthodes au niveau du document, qui cherchent à évaluer la traduction dans son contexte. Il existe un grand nombre de méthodes d'évaluation de la génération de texte. Même si la traduction automatique fait partie des tâches de la génération de texte, dans le cadre de ce travail, nous allons nous contenter de présenter les méthodes les plus utilisées en évaluation de la traduction automatique.

3.1 Évaluation automatique au niveau de la phrase

La majorité des métriques automatiques existantes opèrent au niveau de la phrase. C'est-à-dire que le contexte (les phrases précédentes ou suivantes) n'est pas pris en compte lors de l'évaluation de la traduction. Nous pouvons distinguer deux types : des métriques superficielles basées sur les chaînes de caractères et des métriques apprises.

3.1.1 Métriques superficielles basées sur les chaînes de caractères

Ces métriques opèrent au niveau superficiel de la phrase. Typiquement, elles comparent la traduction automatique (également appelée la traduction candidate) à une traduction de référence. Plus ces deux phrases se ressemblent, plus la traduction est jugée correcte. Voici une description des métriques les plus utilisées.

Les métriques basées sur la comparaison des n-grammes

ChrF (Popović, 2015) : cette métrique qui est basée sur les caractères, compare les n-grammes de caractères (et non pas de mots) entre la traduction automatique et la traduction de référence et calcule la F-mesure des correspondances.

BLEU (Papineni *et al.*, 2002) : compare les n-grammes de mots de la phrase candidate aux n-grammes de mots de la phrase de référence. Le score final est le produit de la précision des n-grammes et de la pénalité de brièveté qui pénalise les traductions plus courtes que la référence. De nombreuses implémentations sont disponibles, celle qui est recommandée est sacreBLEU (Post, 2018).

METEOR (Denkowski & Lavie, 2011) : prend en compte non seulement les formes exactes des mots, mais également les lemmes, les synonymes et les paraphrases. La métrique attribue un poids différent selon la partie du discours du mot. Comme cela requiert des ressources externes, un nombre limité de langues est supporté. Une méthode pour adapter l'approche au multilinguisme a été proposée (Elloumi *et al.*, 2015).

NIST (Doddington, 2002) : son fonctionnement est similaire au score BLEU, mais cette métrique calcule le score avec l'aide d'un poids qui est plus important quand la séquence de mots correcte est moins probable d'apparaître dans un texte, donnant ainsi plus de poids aux mots porteurs d'information.

D'autres métriques ont été proposées, nous pouvons encore citer ROUGE (Lin & Och, 2004) basée sur le calcul de rappel qui a été proposé pour l'évaluation de résumé automatique, GTM (Melamed *et al.*, 2003) basée sur la F-mesure, BLANC (Lita *et al.*, 2005), WNM (Babych & Hartley, 2004), CDER (Leusch *et al.*, 2006) et SIA (Liu & Gildea, 2006).

Les métriques basées sur la distance d'édition

CharacTER (Wang *et al.*, 2016) : cette approche est très similaire à TER, mais au lieu d'opérer au niveau des mots, elle opère au niveau des caractères.

TER (Snover *et al.*, 2006) : calcule le nombre d'éditions de mots (insertions, suppressions, déplacements et substitutions) nécessaires pour passer de la traduction candidate à la traduction de référence.

Parmi d'autres approches, nous pouvons citer WER (Nießen *et al.*, 2000) qui est similaire à TER, mais qui ne compte pas le déplacement comme une édition ou PER (Tillmann *et al.*, 1997) qui ignore l'ordre des mots et les seules éditions sont l'insertion et la suppression.

Le défaut principal de ces métriques est qu'elles sont limitées par la forme superficielle de la phrase. Cela veut dire que la traduction et la référence sont comparées sur la base de la forme uniquement. Dès que la forme d'un mot n'est pas identique à la forme de la référence ceci est considéré comme une erreur. Certaines métriques utilisent des ressources externes (comme des dictionnaires de synonymes ou de la lemmatisation) pour surmonter cette limite. Cela permet de considérer les synonymes ou les formes fléchies du mot comme une traduction correcte malgré leur forme différente. Une deuxième approche pour surmonter la difficulté des formes fléchies des mots consiste à utiliser des métriques qui opèrent au niveau des caractères. Typiquement, la flexion concerne une sous-partie du mot (comme les suffixes), c'est pourquoi comparer les caractères se montre une approche plus flexible que de comparer les mots entiers.

3.1.2 Métriques apprises

Les métriques apprises se distinguent des métriques superficielles par le fait qu'elles sont basées sur l'apprentissage automatique. Souvent, elles exploitent les plongements de mots issus des modèles pré-entraînés et ainsi elles sont plus robustes par rapport aux changements de la forme. Dans la suite nous présentons les approches les plus utilisées et/ou les mieux classées. Ces métriques sont aussi appelées *métriques neuronales* parce qu'elles sont basées sur des modèles de langues neuronaux (ceci est vrai pour toutes les métriques citées dans la suite sauf BEER).

Métriques non-supervisées

La majorité de ces approches reposent sur les calculs de similarité entre les plongements de mots des modèles pré-entraînés. Il s'agit d'apprentissage non-supervisé, parce que ces approches n'ont pas besoin de données annotées en jugements humains et elles reposent sur la logique que la distance entre les représentations de deux phrases dans l'espace vectoriel correspond à leur similarité sémantique.

MEANT (Lo, 2017) : utilise les plongements de mots (Mikolov *et al.*, 2013) et les analyses sémantiques peu superficielles qui déterminent la structure prédicat - argument entre les mots de chaque phrase. La métrique calcule la similarité lexicale et structurale entre la phrase candidate et la phrase de référence.

YISI-1 (Lo, 2019) : calcule la similarité sémantique de la phrase candidate et de référence en utilisant des plongements de mots contextuels de BERT (Devlin *et al.*, 2018). Un analyseur sémantique peut aussi être utilisé pour exploiter les structures sémantiques des deux phrases.

BERTscore (Zhang *et al.*, 2019) : utilise des plongements de mots contextuels de BERT (Devlin *et al.*, 2018) pour calculer la distance cosinus entre les vecteurs des mots de la traduction et de la référence.

Il y a également des approches non-supervisées qui ne reposent pas sur le calcul de similarité.

Prism (Thompson & Post, 2020a) : cette approche utilise un modèle de génération de paraphrases (Thompson & Post, 2020b) pour produire un score de la traduction automatique par rapport à la phrase de référence. Le modèle de paraphrases est multilingue (entraîné sur 39 langues), il peut évaluer la traduction vers toutes ces langues. Son avantage est qu'il n'a pas besoin de jugements humains pour son entraînement.

Métriques supervisées

Ces types d'approche sont entraînés à prédire les jugements humains qui sont fournis en données d'entraînement. La majorité des approches sont construites sur la base de modèles de langues, elles nécessitent alors un modèle pour la langue en question (ou un modèle multilingue) et des données annotées en jugements humains.

BEER (Stanojević & Sima'an, 2014) : cette approche est basée sur un modèle linéaire qui combine les caractéristiques linguistiques de la similarité (comme la précision, le rappel et la F-mesure de mots et de caractères) entre la traduction candidate et de référence avec les caractéristiques d'arbres de permutation (Zhang & Gildea, 2007) qui prennent en compte l'ordre de mots pour prédire le score.

COMET (Rei *et al.*, 2020) [2] : cette métrique est construite sur la base du modèle multi-langues XLM-R (Conneau *et al.*, 2019) et elle a été entraînée en utilisant les jugements d'évaluation humaine de type évaluation directe de WMT des années 2017 à 2020. Elle prend en entrée non seulement les deux traductions (candidate et référence), mais également la phrase source.

BLEURT (Sellam *et al.*, 2020) : cette métrique a été développée pour l'évaluation de la génération du langage naturel. Elle exploite le modèle anglais BERT (Devlin *et al.*, 2018), qui est d'abord affiné sur des données synthétiques et ensuite affiné une deuxième fois sur les jugements humains de l'évaluation de la TA de type évaluation directe provenant de WMT.

UNITE (Wan *et al.*, 2022) : à la différence des autres approches proposées, UNITE est un modèle qui peut servir pour l'évaluation de la TA avec (1) la source uniquement, (2) la référence uniquement, (3) la source et référence combinées. Il repose sur des modèles de langue pré-entraînés qui sont affinés sur des jugements humains.

MATESE (Perrella *et al.*, 2022) : cette métrique repose sur l'approche de l'annotation humaine MQM. Elle utilise des modèles multi-langues comme Conneau *et al.* (2019) et Liu *et al.* (2020a) pour annoter les erreurs de la phrase candidate et les classifier en erreurs majeures et mineures. Le score final est calculé selon les poids tels que défini dans le protocole MQM.

Les métriques décrites ci-dessus nécessitent une traduction de référence. Certains travaux ont également exploré la possibilité d'une évaluation sans besoin de référence (en anglais cette approche est appelée *quality estimation*) et ont proposé des variantes de leur métrique principale. Ceci est le cas pour COMET-QE (Rei *et al.*, 2021), PRISM-src (Thompson & Post, 2020a) et YISI-2 (Lo, 2019). UNITE est la seule métrique où le même modèle peut servir pour évaluer avec ou sans référence.

Les métriques automatiques doivent être évaluées pour savoir si leur score est fiable. La sous-section 3.3 décrit comment cela est fait.

Les métriques automatiques classiques (dont le score BLEU) présentent une faiblesse en évaluant des moteurs de qualité élevée dont elles sont incapables de capturer les différences subtiles (Fomicheva

2. Plus précisément le modèle wmt22-comet-da.

& Specia, 2019; Mathur *et al.*, 2020a). Des travaux plus récents aboutissent à la conclusion que les métriques neuronales ont les meilleures performances et recommandent l'usage de COMET comme métrique principale (Freitag *et al.*, 2022; Kocmi *et al.*, 2021).

Les métriques apprises montrent des résultats encourageants, pourtant elles relèvent de l'effet *boîte noire* : leur score est difficile à expliquer et elles peuvent contenir des biais dont nous ne sommes pas encore conscients. Les métriques basées sur les chaînes de caractères sont moins problématiques de ce point de vue là parce que leur score est facilement explicable.

Les travaux de Kocmi *et al.* (2021) se sont intéressés à étudier si les métriques apprises présentent des biais selon leur mode d'entraînement. Ils ont analysé les performances des métriques selon différents scénarios tel qu'en fonction du couple de langues, de l'alphabet utilisé et du domaine des documents sans pourtant découvrir de biais. Il faut noter que ces analyses ont été effectuées avec la mesure de précision par paire (présentée dans la sous-section 3.3) qui s'est montrée peu performante. Une nouvelle analyse des biais avec une mesure plus fiable serait alors intéressante. Les auteurs préviennent que la pluralité des métriques pourrait entraîner des initiatives malhonnêtes à simplement choisir celle qui donne le résultat le plus opportun, et proposent d'utiliser COMET comme standard.

Aujourd'hui, nous observons une période de transition où le BLEU est encore très utilisé, mais aussi fortement critiqué. Les années d'usage du score BLEU ont fait que ses faiblesses sont plutôt bien connues, même si l'interprétation de petites différences de scores est restée peu claire (Popescu-Belis, 2019). Il a été démontré que la communauté se fiait pendant de nombreuses années sur la comparaison des scores BLEU qui en réalité n'étaient pas comparables (Marie *et al.*, 2021), un changement est alors nécessaire. La métrique COMET commence à se montrer comme la nouvelle métrique standard, sans pourtant encore être complètement adoptée par tous les acteurs. À part le choix de la métrique, il est indispensable d'utiliser des tests de significativité statistique avant de tirer des conclusions des scores obtenus et un des tests très utilisé est le ré-échantillonnage bootstrap (*bootstrap resampling*) (Efron & Tibshirani, 1994).

3.2 Évaluation au niveau de document

Dans le domaine de la TA, une lignée importante de travaux se focalise sur la traduction au niveau du document. Ce type de traduction a pour but de franchir la limite de la phrase et de traduire des segments plus longs, assurant ainsi la prise en compte du contexte. Certains phénomènes linguistiques ont besoin d'un contexte plus large pour permettre une bonne compréhension du texte et par conséquent une bonne traduction. Parmi ces phénomènes, nous pouvons citer par exemple :

- *les pronoms anaphoriques* : le pronom doit respecter l'accord avec son antécédent qui peut se trouver dans les phrases précédentes.
- *les noms polysémiques* : le sens des mots polysémiques dépend du contexte.
- *la cohérence lexicale et terminologique* : certains termes doivent être traduits de la même façon au sein du document. C'est le cas des contrats où au début apparaissent les définitions des parties avec leurs appellations qui doivent être respectées tout au long du document.
- *la structure discursive* : la traduction des connecteurs logiques entre les phrases nécessite la connaissance de la relation logique entre les phrases.
- *la richesse lexicale* : pour ne pas répéter les mêmes expressions, l'accès aux phrases précédentes est nécessaire pour choisir des nouvelles formulations.

Les métriques d'évaluation traditionnelles ne sont pas adaptées pour évaluer les améliorations

obtenues et cela surtout parce qu'elles évaluent la traduction isolée. Aussi, la contextualisation de la traduction touche à une petite proportion du texte, les erreurs ou améliorations de la traduction de ces éléments problématiques ne sont reflétées que très peu dans le score final. De plus, lors de la traduction des pronoms anaphoriques, plus que l'accord avec la traduction de référence, c'est l'accord avec son antécédent qui compte, comme montré dans le tableau 1. Tout cela appauvrit l'évaluation et la rend moins fiable ; d'autant plus quand l'objectif est de développer et d'évaluer un système de traduction sensible au contexte.

Phrase source :	Look, a stone. I'll throw it.
Traduction automatique :	Regarde, un caillou. Je vais **le** lancer.
Traduction de référence :	Regarde, une pierre. Je vais **la** lancer.

TABLE 1 – Exemple d'une traduction correcte même si différente de la référence.

Les travaux proposant des méthodes d'évaluation au niveau du document peuvent être classés en deux catégories. Pour mesurer les améliorations sur certaines erreurs, quelques auteurs utilisent des méthodes basées sur les références (appelées *ground truth*). Par exemple, la test-suites Protest (Guillou & Hardmeier, 2016) cible l'évaluation de la traduction des pronoms anaphoriques. Wong *et al.* (2020) proposent un jeu de test pour les pronoms cataphoriques. La faiblesse de ces approches repose sur le fait qu'elles comparent la traduction à la référence, pourtant comme nous l'avons vu dans le tableau 1, cette méthode n'est pas adaptée.

La deuxième approche est constituée des méthodes basées sur les paires contrastives (en anglais *contrastive pairs*). Il s'agit de test-suites qui ciblent certains phénomènes spécifiques et où pour chaque phrase source, il est associé son contexte, une traduction correcte et une ou plusieurs traductions incorrectes. Le système est évalué sur sa capacité à donner une probabilité plus élevée à la traduction correcte (Rios *et al.*, 2018). Pour pouvoir utiliser cette approche, il est nécessaire d'avoir accès aux probabilités du modèle. Il faut noter que l'approche est basée uniquement sur la comparaison des probabilités à générer la phrase correcte et celle incorrecte. Pourtant, ceci n'est pas vraiment une évaluation de la traduction, parce que rien n'assure que le moteur générerait une de ces phrases (Popescu-Belis, 2019; Post & Junczys-Dowmunt, 2023).

Nous pouvons citer la test-suite de Lopes *et al.* (2020) pour l'évaluation de la traduction de pronoms de l'anglais vers le français, Müller *et al.* (2018) pour la même tâche pour la paire de langues de l'anglais vers l'allemand ou encore Bawden *et al.* (2018) pour évaluer la cohérence lexicale. Rios *et al.* (2018) ont proposé un ensemble pour évaluer la désambiguïsation des mots polysémiques pour les paires de langues allemand-anglais et français-anglais. De la même manière, Alves *et al.* (2022) ont proposé SMAUG, une test-suite contrastive qui cible les erreurs graves en anglais-portugais, espagnol-anglais et portugais-anglais. Il ne se focalise pas sur les erreurs liées au contexte, mais utilise la même logique que celle des paires contrastives.

Les deux approches décrites ci-dessus ne peuvent pas être utilisées comme seule métrique automatique. Leur caractère reste complémentaire et elles sont caractérisées par la faiblesse d'être coûteuses et chronophages à l'élaboration. Elles ne traitent que quelques problèmes bien précis et uniquement pour certaines langues. De nouvelles approches, à vocation plus globale sont apparues. Vernikos *et al.* (2022) proposent d'adapter des métriques neuronales existantes en encodant la phrase à évaluer avec son contexte . Il s'agit d'une extension possible à effectuer avec n'importe quelle métrique neuronale, avec uniquement l'entrée qui est modifiée (en intégrant le contexte) sans nouvel entraînement de la métrique. Les auteurs ont exemplifié leur approche avec BERTScore, Prism, COMET et COMET-src.

Liu *et al.* (2020b) ont utilisé le score BLEU en calculant les correspondances des n-grammes de mots sur l'ensemble du document et appellent cette version de la métrique *d-bleu* (*document-bleu*). Pourtant, comme l'indiquent Post & Junczys-Dowmunt (2023), il y a encore besoin d'une méthode automatique au niveau du document globale qui soit robuste et fiable.

3.3 Méta-évaluation des métriques automatiques

Une méta-évaluation est nécessaire pour connaître la qualité des métriques et savoir lesquelles donnent les scores les plus fiables. Les métriques sont évaluées en comparant les scores automatiques avec les résultats de l'évaluation humaine. L'évaluation humaine devient alors la référence d'or pour l'évaluation (et éventuellement l'entraînement) des métriques automatiques. La manière classique pour évaluer les métriques est de mesurer la corrélation entre les scores prédits par les métriques et ceux attribués par les évaluateurs humains.

Un des acteurs majeurs dans le développement et l'évaluation de nouvelles métriques automatiques de TA est la conférence WMT. Au cours des premières années de la conférence, à partir de 2007, la corrélation de rang de Spearman était initialement utilisée, mais il a été montré que cette corrélation pénalisait très fortement les désaccords même si la différence de qualité de deux systèmes de traduction était petite (Mathur *et al.*, 2020a). C'est pourquoi, à partir de 2014, elle a été remplacée par la corrélation de Pearson. Le test de Williams (Graham & Baldwin, 2014) est utilisé pour vérifier si la différence de performance de deux métriques est statistiquement significative.

Cette façon d'évaluer a récemment été mise en doute. Mathur *et al.* (2020a) démontrent que la présence des modèles dont la qualité mesurée est très distante de celle des autres modèles change considérablement la valeur de la corrélation. Ces modèles sont très faciles à distinguer des autres par les métriques, ce qui se traduit par une corrélation gonflée et donne une fausse confiance dans la fiabilité des métriques. Les auteurs recommandent de recalculer la corrélation après avoir identifié et supprimé ces systèmes de qualité très différente qui sont considérés comme « donnée aberrante ».

Kocmi *et al.* (2021) ont analysé les scores de l'évaluation des systèmes classés en trois groupes selon la paire de langues : 1. l'anglais comme langue cible, 2. l'anglais comme langue source, 3. les paires où ni la langue cible ni la langue source n'est l'anglais. Ils se sont aperçus que les scores des métriques sont sur des échelles différentes selon les paires de langues. Ils argumentent que cela rend l'usage de la corrélation inutilisable. Comme les échelles ne sont pas les mêmes, il n'est pas adéquat de calculer la moyenne des corrélations à travers les différents scénarios pour déterminer quelle métrique a le taux de corrélation le plus élevé. Les auteurs proposent une nouvelle mesure qu'ils appellent « pair-wise accuracy », la *précision par paire* qui est définie comme le pourcentage des paires de systèmes pour lesquels la métrique automatique a assigné le score le plus élevé au même système que l'évaluation humaine.

Cette méthode a été adoptée en 2021 à la conférence WMT (Freitag *et al.*, 2021b), pourtant elle s'est révélée peu discriminante, car elle ne montrait pas assez de différences significatives entre les métriques. Ainsi, plusieurs métriques ont été classées sur la même place. L'année suivante, WMT a alors proposé une nouvelle méthode (Freitag *et al.*, 2022) qui calcule le rang moyen de chaque métrique basée sur la corrélation à travers les différents scénarios, comme : la paire de langues, le domaine, le niveau d'évaluation (niveau du segment ou du système), le coefficient de corrélation utilisé (la précision par paire, le coefficient de Pearson et de Kendall), la méthode de calcul de moyennes. Le classement final est une moyenne des classements à travers les scénarios cités.

4 Discussion

Avec le nombre grandissant de métriques automatiques, la question principale est de savoir laquelle choisir et pour cela, la méta-évaluation doit apporter la réponse. Comme nous l'avons vu, la méta-évaluation est étroitement liée à (1) la méthode de mesure de corrélation et (2) les jugements d'évaluation humaine. Kocmi *et al.* (2021) ont présenté un classement de performance des métriques automatiques. Comme référence, ils ont utilisé les jugements d'évaluation directe collectés lors des évaluations internes chez Microsoft. Comme méthode de mesure de corrélation, ils ont introduit et utilisé la précision par paire. Pourtant, il s'est montré que les jugements humains ont des défauts de significativité statistique (Wei *et al.*, 2022) et que la mesure de la précision par paire n'arrive pas à bien distinguer entre les métriques (Freitag *et al.*, 2022). Tout cela montre que malgré les avancées en matière d'évaluation humaine, la procédure de la méta-évaluation des métriques présente des difficultés et doit être davantage étudiée et développée.

Même s'il semble que l'évaluation MQM est la plus fiable, la quantité de cette donnée reste petite par rapport à l'historique d'évaluation directe. En plus, cette approche d'évaluation est plus chronophage, ce qui rend son utilisation peu pratique. Cela fait que dans certains cas, les organisateurs d'évaluations préfèrent d'autres approches, comme Kocmi *et al.* (2022) qui ont combiné l'évaluation directe en échelle continue avec l'évaluation sur l'échelle de Likert qui corrèle bien avec l'évaluation MQM. Au mieux de notre connaissance, aucune analyse statistique de grande échelle telle que celle de (Wei *et al.*, 2022) n'a été faite sur les annotations MQM, ce qui se révèlerait fortement intéressant.

Du point de vue de l'évaluation humaine comme donnée d'entraînement pour les métriques automatiques, les données les plus abondantes restent les jugements d'évaluation directe. En effet, les nouvelles métriques de la famille COMET commencent à utiliser les annotations MQM pour l'entraînement, mais l'évaluation directe reste la source de données la plus riche. Pourtant, il a été démontré que ces jugements corrèlent très peu avec les jugements d'annotation MQM (considérées comme la référence la plus exacte et véridique) (Freitag *et al.*, 2021a). Nous pouvons alors supposer qu'avec la croissance d'ensemble de données annotées en MQM, les métriques automatiques à leur tour vont être améliorées.

Une vraie carence reste l'évaluation automatique sensible au contexte. L'approche des test-suites est insuffisante pour deux raisons : 1) elle n'est qu'une approche complémentaire à utiliser avec une autre métrique, 2) les ensembles de test ne ciblent que certains cas bien précis et en nombre limité de langues. L'approche de Vernikos *et al.* (2022) propose une solution, cependant elle ne fait que modifier l'entrée du modèle. Une métrique apprise entraînée avec des données annotées au niveau du document est souhaitable et peut faire l'objet de futurs travaux.

Un autre aspect à prendre en compte est l'impact des traductions de référence sur la performance des métriques, puisque la plupart des métriques les plus performantes ont besoin de ces traductions. La qualité de ces références impacte largement leur performance et il n'est pas encore clair de savoir quelles sont les caractéristiques d'une bonne référence (Freitag *et al.*, 2022). Les métriques sans références qui deviennent de plus en plus robustes (Rei *et al.*, 2021) pourraient être une solution à ce problème.

Une autre piste de travaux possibles est d'investiguer davantage si les différentes métriques présentent un biais par rapport à certaines familles de modèles de traduction (selon leur architecture ou langues). Avec le développement rapide de grands modèles de langues (*large language models*), il serait intéressant de tester la fiabilité des métriques d'évaluation sur les traductions produites par ces

modèles.

5 Conclusion

En conclusion, cet article a présenté l'état de l'art de l'évaluation de la traduction automatique, en soulignant que l'évaluation de la qualité de la traduction automatique est subjective et difficile. Nous avons présenté les différentes approches de l'évaluation humaine avec leurs avantages et inconvénients. Ensuite, nous avons présenté les approches d'évaluation automatique les plus utilisées dans le domaine de la traduction. D'après les études récentes, il est impératif d'arrêter l'usage du score BLEU comme métrique principale. Les métriques neuronales se montrent plus fiables. Notamment le score COMET a attiré beaucoup d'attention, il a été classé comme métrique la plus performante dans plusieurs études et pour cela commence à devenir la nouvelle métrique standard. Nous avons montré qu'il y a un besoin pressant de développer des métriques fiables pour évaluer la traduction tout en prenant compte de son contexte.

L'évaluation humaine est essentielle pour le développement des approches automatiques, elle est utilisée pour la méta-évaluation et (selon l'approche) pour l'entraînement. Il faut mentionner également que la méta-évaluation qui consiste à calculer la corrélation entre les scores automatiques et humains ne fait pas l'unanimité et de nouvelles méthodes ont été proposées sans encore définir de protocole. En somme, nous avons souligné l'importance de continuer à développer des protocoles d'évaluation humaine et de méta-évaluation fiables pour ainsi créer une base solide pour le développement des métriques automatiques. Une bonne évaluation étant la condition essentielle pour un développement optimal dans le domaine de la traduction automatique, cela contribuera à l'avancement de manière globale.

Remerciements

Ce travail était effectué dans le cadre d'une convention CIFRE, gérée par l'Association Nationale de la Recherche Technique, et établie entre le Laboratoire d'Informatique de Grenoble et la société Lingua Custodia. Nous tenons à remercier les encadrants Emmanuelle Esperança-Rodier, Marco Dinarelli, Hervé Blanchon et Raheel Qader pour leurs conseils et commentaires pertinents. Nous remercions également les relecteurs anonymes.

Références

ALVES D., REI R., FARINHA A. C., DE SOUZA J. G. & MARTINS A. F. (2022). Robust mt evaluation with sentence-level multilingual augmentation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, p. 469–478.

BABYCH B. & HARTLEY T. (2004). Extending the bleu mt evaluation method with frequency weightings. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, p. 621–628.

BAWDEN R., SENNRICH R., BIRCH A. & HADDOW B. (2018). Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter*

of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers), p. 1304–1313, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/N18-1118.

BOJAR O., FEDERMANN C., FISHEL M., GRAHAM Y., HADDOW B., HUCK M., KOEHN P. & MONZ C. (2018). Findings of the 2018 conference on machine translation (wmt18). In *Proceedings of the Third Conference on Machine Translation*, volume 2, p. 272–307.

BOJAR O., FEDERMANN C., HADDOW B., KOEHN P., POST M. & SPECIA L. (2016). Ten years of wmt evaluation campaigns : Lessons learnt. In *Proceedings of the LREC 2016 Workshop "Translation Evaluation–From Fragmented Tools and Data Sets to an Integrated Ecosystem*, p. 27–34.

CONNEAU A., KHANDELWAL K., GOYAL N., CHAUDHARY V., WENZEK G., GUZMÁN F., GRAVE E., OTT M., ZETTLEMOYER L. & STOYANOV V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv :1911.02116*.

DENKOWSKI M. & LAVIE A. (2011). Meteor 1.3 : Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the sixth workshop on statistical machine translation*, p. 85–91.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2018). Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*.

DODDINGTON G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, p. 138–145.

EFRON B. & TIBSHIRANI R. J. (1994). *An introduction to the bootstrap*. CRC press.

ELLOUMI Z., BLANCHON H., SERASSET G. & BESACIER L. (2015). METEOR for multiple target languages using DBnary. In *Proceedings of Machine Translation Summit XV : Papers*, Miami, USA.

FOMICHEVA M. & SPECIA L. (2019). Taking mt evaluation metrics to extremes : Beyond correlation with human judgments. *Computational Linguistics*, **45**(3), 515–558.

FREITAG M., FOSTER G., GRANGIER D., RATNAKAR V., TAN Q. & MACHEREY W. (2021a). Experts, errors, and context : A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, **9**, 1460–1474.

FREITAG M., REI R., MATHUR N., LO C.-K., STEWART C., AVRAMIDIS E., KOCMI T., FOSTER G., LAVIE A. & MARTINS A. F. (2022). Results of wmt22 metrics shared task : Stop using bleu–neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, p. 46–68.

FREITAG M., REI R., MATHUR N., LO C.-K., STEWART C., FOSTER G., LAVIE A. & BOJAR O. (2021b). Results of the wmt21 metrics shared task : Evaluating metrics with expert-based human evaluations on ted and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, p. 733–774.

GRAHAM Y. & BALDWIN T. (2014). Testing for significance of increased correlation with human judgment. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 172–176, Doha, Qatar : Association for Computational Linguistics. DOI : 10.3115/v1/D14-1020.

GRAHAM Y., BALDWIN T., MOFFAT A. & ZOBEL J. (2013). Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop*

and Interoperability with Discourse, p. 33–41, Sofia, Bulgaria : Association for Computational Linguistics.

GRAHAM Y., HADDOW B. & KOEHN P. (2019). Translationese in machine translation evaluation. *arXiv preprint arXiv :1906.09833*.

GUILLOU L. & HARDMEIER C. (2016). Protest : A test suite for evaluating pronouns in machine translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 636–643.

HASSAN H., AUE A., CHEN C., CHOWDHARY V., CLARK J., FEDERMANN C., HUANG X., JUNCZYS-DOWMUNT M., LEWIS W., LI M. *et al.* (2018). Achieving human parity on automatic Chinese to English news translation. *arXiv preprint arXiv :1803.05567*.

KOCMI T., BAWDEN R., BOJAR O., DVORKOVICH A., FEDERMANN C., FISHEL M., GOWDA T., GRAHAM Y., GRUNDKIEWICZ R., HADDOW B. *et al.* (2022). Findings of the 2022 conference on machine translation (wmt22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, p. 1–45.

KOCMI T., FEDERMANN C., GRUNDKIEWICZ R., JUNCZYS-DOWMUNT M., MATSUSHITA H. & MENEZES A. (2021). To ship or not to ship : An extensive evaluation of automatic metrics for machine translation. *arXiv preprint arXiv :2107.10821*.

KOPONEN M. (2016). *Machine Translation Post-editing and Effort : Empirical Studies on the Post-editing Process*. Thèse de doctorat, University of Helsinki, Finland.

LÄUBLI S., SENNRICH R. & VOLK M. (2018). Has machine translation achieved human parity ? a case for document-level evaluation. *arXiv preprint arXiv :1808.07048*.

LEUSCH G., UEFFING N. & NEY H. (2006). Cder : Efficient mt evaluation using block movements. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, p. 241–248.

LIN C.-Y. & OCH F. J. (2004). Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, p. 605–612.

LITA L. V., ROGATI M. & LAVIE A. (2005). Blanc : Learning evaluation metrics for mt. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, p. 740–747.

LIU D. & GILDEA D. (2006). Stochastic iterative alignment for machine translation evaluation. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, p. 539–546.

LIU Y., GU J., GOYAL N., LI X., EDUNOV S., GHAZVININEJAD M., LEWIS M. & ZETTLEMOYER L. (2020a). Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, **8**, 726–742. DOI : 10.1162/tacl_a_00343.

LIU Y., GU J., GOYAL N., LI X., EDUNOV S., GHAZVININEJAD M., LEWIS M. & ZETTLEMOYER L. (2020b). Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, **8**, 726–742. DOI : 10.1162/tacl_a_00343.

LO C.-K. (2017). Meant 2.0 : Accurate semantic mt evaluation for any output language. In *Proceedings of the second conference on machine translation*, p. 589–597.

LO C.-K. (2019). Yisi-a unified semantic mt quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2 : Shared Task Papers, Day 1)*, p. 507–513.

LOMMEL A., BURCHARDT A. & USZKOREIT H. (2014). Multidimensional quality metrics (mqm) : A framework for declaring and describing translation quality metrics. *Tradumàtica : tecnologies de la traducció*, **0**, 455–463. DOI : 10.5565/rev/tradumatica.77.

LOPES A., FARAJIAN M. A., BAWDEN R., ZHANG M. & MARTINS A. T. (2020). Document-level neural mt : A systematic comparison. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, p. 225–234, Lisbon, Portugal.

MARIE B., FUJITA A. & RUBINO R. (2021). Scientific credibility of machine translation research : A meta-evaluation of 769 papers. *arXiv preprint arXiv :2106.15195*.

MATHUR N., BALDWIN T. & COHN T. (2020a). Tangled up in bleu : Reevaluating the evaluation of automatic machine translation evaluation metrics. *arXiv preprint arXiv :2006.06264*.

MATHUR N., WEI J., FREITAG M., MA Q. & BOJAR O. (2020b). Results of the wmt20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, p. 688–725.

MELAMED I. D., GREEN R. & TURIAN J. (2003). Precision and recall of machine translation. In *Companion Volume of the Proceedings of HLT-NAACL 2003-Short Papers*, p. 61–63.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv :1301.3781*.

MÜLLER M., RIOS A., VOITA E. & SENNRICH R. (2018). A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. *arXiv preprint arXiv :1810.02268*.

NIESSEN S., OCH F. J., LEUSCH G., NEY H. *et al.* (2000). An evaluation tool for machine translation : Fast evaluation for mt research. In *LREC*.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, p. 311–318.

PERRELLA S., PROIETTI L., SCIRÈ A., CAMPOLUNGO N. & NAVIGLI R. (2022). Matese : Machine translation evaluation as a sequence tagging problem. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, p. 569–577.

POPESCU-BELIS A. (2019). Context in neural machine translation : A review of models and evaluations. *arXiv preprint arXiv :1901.09115*.

POPOVIĆ M. (2015). chrF : character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, p. 392–395, Lisbon, Portugal : Association for Computational Linguistics. DOI : 10.18653/v1/W15-3049.

POST M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation : Research Papers*, p. 186–191, Belgium, Brussels : Association for Computational Linguistics.

POST M. & JUNCZYS-DOWMUNT M. (2023). Escaping the sentence-level paradigm in machine translation. *arXiv preprint arXiv :2304.12959*.

REI R., FARINHA A. C., ZERVA C., VAN STIGT D., STEWART C., RAMOS P., GLUSHKOVA T., MARTINS A. F. T. & LAVIE A. (2021). Are references really needed ? unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, p. 1030–1040, Online : Association for Computational Linguistics.

REI R., STEWART C., FARINHA A. C. & LAVIE A. (2020). Comet : A neural framework for mt evaluation. *arXiv preprint arXiv :2009.09025*.

RIOS A., MÜLLER M. & SENNRICH R. (2018). The word sense disambiguation test suite at wmt18. In *Proceedings of the Third Conference on Machine Translation : Shared Task Papers* : Association for Computational Linguistics.

SELLAM T., DAS D. & PARIKH A. (2020). BLEURT : Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7881–7892, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.704.

SNOVER M., DORR B., SCHWARTZ R., MICCIULLA L. & MAKHOUL J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas : Technical Papers*, p. 223–231.

STANOJEVIĆ M. & SIMA'AN K. (2014). Beer : Better evaluation as ranking. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, p. 414–419.

THOMPSON B. & POST M. (2020a). Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online : Association for Computational Linguistics.

THOMPSON B. & POST M. (2020b). Paraphrase generation as zero-shot multilingual translation : Disentangling semantic similarity from lexical and syntactic diversity. In *Proceedings of the Fifth Conference on Machine Translation (Volume 1 : Research Papers)*, Online : Association for Computational Linguistics.

TILLMANN C., VOGEL S., NEY H., ZUBIAGA A. & SAWAF H. (1997). Accelerated dp based search for statistical translation. In *Eurospeech*.

TORAL A., CASTILHO S., HU K. & WAY A. (2018). Attaining the unattainable ? reassessing claims of human parity in neural machine translation. *arXiv preprint arXiv :1808.10432*.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł. & POLOSUKHIN I. (2017). Attention is all you need. *Advances in neural information processing systems*, **30**.

VERNIKOS G., THOMPSON B., MATHUR P. & FEDERICO M. (2022). Embarrassingly easy document-level mt metrics : How to convert any pretrained metric into a document-level metric. *arXiv preprint arXiv :2209.13654*.

VOJTĚCHOVÁ T., NOVÁK M., KLOUČEK M. & BOJAR O. (2019). Sao wmt19 test suite : Machine translation of audit reports. *arXiv preprint arXiv :1909.01701*.

WAN Y., LIU D., YANG B., ZHANG H., CHEN B., WONG D. & CHAO L. (2022). UniTE : Unified translation evaluation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 8117–8127, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.558.

WANG W., PETER J.-T., ROSENDAHL H. & NEY H. (2016). CharacTer : Translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation : Volume 2, Shared Task Papers*, p. 505–510, Berlin, Germany : Association for Computational Linguistics. DOI : 10.18653/v1/W16-2342.

WEI J. T.-Z., KOCMI T. & FEDERMANN C. (2022). Searching for a higher power in the human evaluation of mt. *arXiv preprint arXiv :2210.11612*.

WONG K., MARUF S. & HAFFARI G. (2020). Contextual neural machine translation improves translation of cataphoric pronouns. *arXiv preprint arXiv :2004.09894*.

ZHANG H. & GILDEA D. (2007). Factorization of synchronous context-free grammars in linear time. In *Proceedings of SSST, NAACL-HLT 2007/AMTA Workshop on Syntax and Structure in Statistical Translation*, p. 25–32.

ZHANG T., KISHORE V., WU F., WEINBERGER K. Q. & ARTZI Y. (2019). Bertscore : Evaluating text generation with bert. *arXiv preprint arXiv :1904.09675*.

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Normalisation lexicale de contenus générés par les utilisateurs sur les réseaux sociaux

Lydia Nishimwe
Inria Paris, 2 rue Simone IFF, 75012 Paris, France
lydia.nishimwe@inria.fr

RÉSUMÉ

L'essor du traitement automatique des langues (TAL) se vit dans un monde où l'on produit de plus en plus de contenus en ligne. En particulier sur les réseaux sociaux, les textes publiés par les internautes sont remplis de phénomènes « non standards » tels que les fautes d'orthographe, l'argot, les marques d'expressivité, etc. Ainsi, les modèles de TAL, en grande partie entraînés sur des données « standards », voient leur performance diminuer lorsqu'ils sont appliqués aux contenus générés par les utilisateurs (CGU). L'une des approches pour atténuer cette dégradation est la normalisation lexicale : les mots non standards sont remplacés par leurs formes standards. Dans cet article, nous réalisons un état de l'art de la normalisation lexicale des CGU, ainsi qu'une étude expérimentale préliminaire pour montrer les avantages et les difficultés de cette tâche.

ABSTRACT

Lexical normalisation of user-generated content on social media

The boom of natural language processing (NLP) is taking place in a world where more and more content is produced online. On social networks especially, textual content published by users are full of "non-standard" phenomena such as spelling mistakes, jargon, marks of expressiveness, etc. Thus, NLP models, which are largely trained on "standard" data, suffer a decline in performance when applied to user-generated content (UGC). One approach to mitigate this degradation is through lexical normalisation where non-standard words are replaced by their standard forms. In this paper, we review the state of the art of lexical normalisation of UGC, as well as run a preliminary experimental study to show the advantages and difficulties of this task.

MOTS-CLÉS : normalisation lexicale, contenus générés par les utilisateurs (CGU), réseaux sociaux, modèles de langue.

KEYWORDS: lexical normalisation, user-generated content (UGC), social media, language models.

1 Introduction

Pour développer des systèmes de traitement automatique des langues (TAL) capables de traiter les « contenus générés par les utilisateurs » (CGU), il est nécessaire de se pencher soit sur les moyens de rendre les modèles robustes aux variations linguistiques associés à ces contenus, soit sur la normalisation de ces contenus afin qu'ils ressemblent le plus possible à la langue standard sur laquelle ces modèles sont généralement entraînés. Dans cet article, nous étudions la seconde de ces deux approches et nous consacrons ainsi à la tâche de normalisation lexicale des CGU, qui consiste à remplacer les formes non standard par leurs variantes standard (« normalisées »).

Actes des 16e Rencontres Jeunes Chercheurs en RI (RJCRI) et 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL), pages 160–183.

Nous commençons par un état de l'art du domaine : nous décrivons d'abord les spécificités des CGU et les problèmes qu'ils posent pour les systèmes de TAL (section 2). Nous détaillons ensuite les méthodes proposées dans la littérature (section 3.1) pour la normalisation des CGU, mais également pour des tâches connexes telles que la normalisation phonétique, la correction post-OCR, la correction grammaticale et la normalisation des variantes dialectales ou historiques (section 3.2). Nous poursuivons avec un bref panorama des jeux de test et des métriques pour la tâche en question (section 3.3). Nous proposons ensuite une étude expérimentale préliminaire (section 4) dont le but est d'illustrer certaines des difficultés de la tâche en termes de modélisation et d'évaluation. La méthode que nous avons choisie repose sur un processus en deux étapes : (i) la détection supervisée de tokens non standards modélisée comme une tâche d'étiquetage de séquences, (ii) la normalisation des tokens détectés lors de la première étape à l'aide d'une approche couplant modélisation de langue par masquage (*masked language modelling*) et distance d'édition. L'idée sous-jacente est de choisir une forme standard qui soit à la fois appropriée en contexte et formellement similaire au mot non standard d'origine.

Nos expériences explorent le poids relatif à donner à la distance d'édition, en comparant plusieurs modèles de langue par masquage, dont certains entraînés sur des données CGU non standards. Nous menons nos évaluations au moyen de plusieurs métriques automatiques, y compris des métriques qui s'appuient sur la précision et des métriques développées pour la traduction automatique, qui ont été utilisées précédemment dans la littérature. Notre évaluation nous permet de constater que l'évaluation de la normalisation des CGU est loin d'être simple. L'approche que nous testons présente des limites évidentes, dont certaines sont mal prises en compte par les métriques usuelles.

2 Le TAL et les CGU : une relation amour-haine

2.1 Les CGU sur les réseaux sociaux

Sproat *et al.* (2001) ont utilisé le terme de « mots non standards » (*non-standard words, NSW*) pour décrire des mots et symboles (chiffres, abréviations, dates, devises monétaires, acronymes) qui ne se trouvent pas dans un dictionnaire, ou dont la prononciation ne peut se déduire des règles usuelles [1].

Avec l'expansion des messages textuels envoyés par téléphone (*Short Message Service, SMS*) au tournant du XXI$^{\text{ème}}$ siècle, d'autres phénomènes non standards sont apparus dans les textes écrits : la simplification de l'orthographe (p. ex. la suppression d'accents) et de la grammaire (p. ex. l'omission de pronoms), la substitution phonétique (`a 2m1` pour `à demain`), l'utilisation d'émoticônes, etc. Alors que les mots non standards étaient considérés comme grammaticalement corrects, ces nouveaux phénomènes n'étaient pas encore formalisés en linguistique (Aw *et al.*, 2006).

Après les SMS, les textes non standards ont connu un essor sur les réseaux sociaux, les forums de discussion, les tchats et d'autres plate-formes où les internautes interagissent. Cela a marqué l'émergence des CGU [2], qui ont été largement qualifiés de « bruités » dans le domaine du TAL. Pour

1. D'autres termes similaires employés dans la littérature sont : « mots mal formés » (*ill-formed words*, Han & Baldwin (2011)) ou encore « tokens non standards » (*non-standard tokens*, Liu *et al.* (2012)).

2. Une meilleure traduction serait « contenus *produits* par les utilisateurs », mais *générés* est globalement accepté par symétrie à l'anglais *user-generated content (UGC)*. D'autres appellations rencontrées dans la littérature sont : « langage texto » (*texting language*, Choudhury *et al.* (2007)), « textes bruités » (Formiga & Fonollosa, 2012) ou encore « textes bruités générés par les utilisateurs » (*noisy user-generated text*) (Baldwin *et al.*, 2015).

quantifier cette affirmation, Baldwin *et al.* (2013) ont mené une étude linguistique et statistique sur un corpus de CGU provenant de sources différentes et ont démontré qu'il était effectivement plus « bruité » qu'un corpus composé de textes standards édités. Par ailleurs, Eisenstein (2013) a expliqué des raisons fréquentes pour lesquelles les utilisateurs écrivent « si mal », à savoir : l'illettrisme, le nombre de caractères limité (p. ex. Twitter), le système de saisie du texte (clavier externe *versus* clavier tactile avec autocomplétion), des phénomènes pragmatiques, et certaines variables sociales.

Certains mots non standards présents dans les CGU sont propres aux réseaux sociaux utilisés, comme les hashtags (`#JeuxOlympiques`), les mentions (`@gouvernementFR`) et leur méta-langage (`RT` pour `Retweet`). De plus, l'argot évolue constamment et il y a toujours de nouveaux mots émergents (p. ex. des néologismes comme `burka + bikini` → `burkini`), ou encore l'utilisation du *leetspeak* pour censurer des jurons ou des propos offensifs (`!d10t` pour `idiot`). Un autre phénomène omniprésent est l'emploi de mots empruntés d'autres langues ou même le mélange de plusieurs langues (l'alternance codique).

Faire une liste exhaustive de tous les phénomènes non standards spécifiques aux CGU n'est pas une tâche aisée, cependant quelques tentatives ont été faites. Par exemple, Seddah *et al.* (2012) ont proposé une classification des phénomènes CGU rencontrés dans des forums de discussion et réseaux sociaux français. Ils les ont définis selon trois axes : (1) les phénomènes ergographiques qui visent à simplifier l'écriture, par exemple l'omission d'accents, la phonétisation, et certaines fautes d'orthographe (`son` pour `sont`) ; (2) les phénomènes transversaux comme la contraction (`nimp` pour `n'importe quoi`) et la segmentation typographique (`c a dire` pour `c'est-à-dire`) ; (3) les marques d'expressivité comme l'étirement des graphèmes ou de ponctuation (`superrr !!!`) et les émoticônes. Sanguinetti *et al.* (2020) se sont appuyés sur cette classification et y ont rajouté les phénomènes d'autocensure, ainsi qu'un quatrième axe des phénomènes d'influence de langues étrangères comme la translitération, la formation de nouveaux verbes et l'autocorrection. Par ailleurs, van der Goot *et al.* (2018) ont élaboré une taxonomie des spécificités CGU en anglais. Ils ont considéré trois types d'« anomalies » : (1) les anomalies non intentionnelles comme les fautes typographiques, orthographiques ou de segmentation ; (2) les anomalies intentionnelles telles que les abréviations d'expressions (`mdr` pour `mort de rire`), les répétitions, les contractions, les transformations phonétiques et l'argot ; (3) les anomalies inconnues.

2.2 L'impact négatif de CGU sur le TAL

Les modèles de TAL étant généralement entraînés sur des données standards, ils s'attendent à traiter des données du même type. En présence de phénomènes CGU, la performance de plusieurs tâches de TAL est négativement affectée, à savoir : l'analyse syntaxique (Foster, 2010 ; Seddah *et al.*, 2012), la détection de thèmes (*topic detection*) (Muñoz-García *et al.*, 2012), la tokénisation (Aminian *et al.*, 2012), la reconnaissance d'entités nommées (Moon *et al.*, 2018), l'analyse des dépendances (van der Goot & van Noord, 2018 ; van der Goot, 2019a), la traduction automatique (Belinkov & Bisk, 2017 ; Michel & Neubig, 2018 ; Rosales Núñez *et al.*, 2021a), l'analyse de sentiment (Kumar *et al.*, 2020), etc.

3 La normalisation lexicale : le chevalier blanc ?

Dans leur analyse statistique et linguistique des corpus CGU, Baldwin *et al.* (2013) ont montré que l'application de certaines tâches de TAL, comme l'identification de la langue, la normalisation lexicale et l'étiquetage morphosyntaxique, peuvent réduire le niveau de bruit dans ces corpus. Pour pallier la dégradation de performance des modèles de TAL causée par la présence de phénomènes CGU, Eisenstein (2013) a recensé deux approches principales : (1) la normalisation, qui vise à adapter les données à ce que les modèles attendent, et (2) l'adaptation de domaine, qui consiste à adapter les modèles aux données, par exemple en entraînant sur des données réelles CGU (Nguyen *et al.*, 2020) ou encore à entraîner les modèles sur des données bruitées synthétiques (Karpukhin *et al.*, 2019). Une autre approche consiste à changer l'architecture du modèle, par exemple en passant à l'échelle des caractères (Riabi *et al.*, 2021; Rosales Núñez *et al.*, 2021b), ou des segments de phrases (Rosales Núñez *et al.*, 2019a) pour la traduction automatique.

L'approche privilégiée est restée la normalisation lexicale, utilisée en amont d'autres tâches de TAL. Elle a fortement amélioré la performance dans plusieurs tâches : la reconnaissance d'entités nommées simples (Nguyen *et al.*, 2016) ou imbriquées (Plank *et al.*, 2020), l'étiquetage morphosyntaxique (van der Goot *et al.*, 2017; van der Goot & Çetinoğlu, 2021), l'analyse de dépendances (van der Goot *et al.*, 2020), ou encore la compréhension des CGU par des locuteurs non natifs (Ehara, 2021).

3.1 Méthodes

Approches modulaires L'une des approches classiques est de concevoir un système qui combine plusieurs modules, soit pour normaliser différents types de mots non standards, soit pour aborder la normalisation sous plusieurs angles (orthographe, phonétique, modèle de langue, ...). Par exemple, Sproat *et al.* (2001) ont implémenté un système qui, après avoir tokénisé le texte, classifie les tokens par un arbre de décision. Ensuite, il crée un treillis de mots à partir des étiquettes de classe, et un modèle n-grammes en déduit le meilleur candidat de normalisation. Liu *et al.* (2012) ont introduit un modèle à 4 composantes : un module de transformation de lettres, un module d'amorçage visuel, un correcteur orthographique, et un module qui combine les meilleurs candidats proposés par les 3 autres modules. Ahmed (2015) a aussi proposé une approche qui, dans un premier temps, génère un ensemble de candidats sur base d'une distance d'édition. Ensuite, cet ensemble est raffiné par une technique de similarité phonétique. Simultanément, l'algorithme de Peter Norvig pour la correction orthographique [3] est aussi appliqué sur ces candidats, et les deux résultats sont comparés. Si les deux méthodes produisent le même candidat, alors la normalisation est terminée. Sinon, un modèle 5-grammes est appliqué sur les contextes phonétiques pour sélectionner le meilleur candidat. Melero *et al.* (2016) ont proposé une stratégie qui utilise un correcteur orthographique pour détecter les mots non standards et générer des candidats, et un module de sélection constitué d'une interpolation linéaire de 4 modèles de langue encodant des informations linguistiques différentes (vraie casse, minuscules, lemmes, morphosyntaxe). van der Goot & van Noord (2017) ont conçu MoNoise, un modèle modulaire qui comprend, entre autres, un correcteur orthographique et un modèle de plongements de mots. Ils ont entraîné une forêt aléatoire pour sélectionner la meilleure normalisation parmi les candidats générés par les différents modules. van der Goot (2019b) a implémenté des améliorations au modèle MoNoise, et a démontré qu'il obtient la meilleure performance (à l'époque) de normalisation lexicale sur plusieurs langues. van der Goot (2021) a ensuite implémenté une version de MoNoise qui effectue

3. http://norvig.com/spell-correct.html

un transfert entre les différentes langues : le modèle est entraîné sur des données monolingues et annotées dans la langue source, qui sont remplacées par des données monolingues dans la langue cible pendant l'inférence.

Approches statistiques Choudhury *et al.* (2007) ont développé un modèle bigramme à base d'un modèle de Markov caché (MMC) pour corriger les erreurs dans le langage texto ; Xu *et al.* (2015) se sont appuyés sur cette approche et l'ont adapté au chinois en proposant un modèle à base de champs aléatoires conditionnels (*Conditional Random Fields, CRF*) pour segmenter les mots non standards en syllabes. Han & Baldwin (2011) ont implémenté une stratégie en trois temps. D'abord, pour chaque mot hors vocabulaire, un ensemble de candidats est généré selon des variations morphophonémiques. Ensuite, un classifieur détecte si le mot est « mal formé » à partir de cet ensemble, et le meilleur candidat de normalisation est sélectionné lors de la dernière étape. Supranovich & Patsepnia (2015) ont proposé un système à 2 composantes : un modèle à base de CRF pour détecter les mots bruités, et une étape de normalisation qui remplace les mots détectés par leurs variantes dans le lexique. Plus récemment, Jiang *et al.* (2022) ont introduit une approche de normalisation lexicale à grande échelle qui utilise des familles LSH (*Locality-Sensitive Hashing*) pour générer des candidats en se basant sur la morphologie des mots.

Méthodes à base de règles Clark & Araki (2011) ont proposé une méthode qui recherche les mots non standards dans une base de données et les remplace selon des règles définies. Baranes & Sagot (2014) ont développé une approche qui repose sur un système d'induction par analogie des règles apprises sur un corpus de fautes lexicales annotées en français. Pour la normalisation de tweets en espagnol, Ruiz *et al.* (2014) ont combiné des règles dans une étape de pré-traitement avec un modèle de distance d'édition et un modèle n-grammes pour sélectionner les candidats. Par ailleurs, Cotelo *et al.* (2015) ont proposé un schéma modulaire dont un module de calcul de distance d'édition, et d'autres modules à base de règles selon le type de mots hors vocabulaire. D'autre part, Cerón-Guzmán & León-Guzmán (2016) ont implémenté un modèle qui utilise des transducteurs finis pour générer des ensembles de candidats à partir de règles graphémiques et phonémiques. Beckley (2015) a implémenté une architecture simple qui consiste en une liste de substitutions obtenues de manière semi-supervisée, quelques composantes à base de règles, et un algorithme de Viterbi (Viterbi, 1967) sélectionnant le meilleur candidat. Kogkitsidou & Antoniadis (2016) ont proposé un modèle hybride pour la normalisation de SMS qui, d'une part, produit une représentation intermédiaire du message par l'application de grammaires locales et, d'autre part, utilise un modèle de traduction automatique à base de règles pour convertir cette représentation vers une forme standard.

La normalisation vue comme une tâche de traduction Aw *et al.* (2006) ont adapté un modèle de traduction statistique à base de segments pour « traduire de l'anglais des SMS vers l'anglais standard ». Pour la normalisation de SMS en français, Kobus *et al.* (2008) ont aussi utilisé un tel modèle de traduction. Afin d'améliorer sa performance, ils l'ont combiné avec un module inspiré de la reconnaissance automatique de la parole pour proposer des hypothèses pour les mots hors vocabulaires, et un modèle de langue pour sélectionner le meilleur candidat. Par ailleurs, Pennell & Liu (2011) ont implémenté un modèle de traduction à base de caractères pour normaliser spécifiquement les abréviations dans les SMS. Li & Liu (2012) ont proposé de combiner un correcteur orthographique avec un modèle de traduction à base de blocs de caractères générés selon des règles phonétiques chinoises. Formiga & Fonollosa (2012) ont utilisé un modèle de traduction à base de caractères pour traduire le texte « bruité » en texte « propre » en amont d'une tâche de traduction de l'anglais vers l'espagnol. Ce premier modèle produit un treillis de variantes orthographiques qui est passé en entrée d'un modèle de traduction bilingue à base de segments. Matos Veliz *et al.* (2019) ont évalué deux modèles de traduction automatique, statistique et neuronale, pour la normalisation de divers CGU en

anglais et en néerlandais. Ils ont conclu que, pour la traduction statistique, il est mieux d'entraîner le modèle de langue sous-jacent sur un corpus issu d'un domaine similaire à celui des UGC et que, pour la traduction neuronale, il est préférable d'ajouter plus de données d'entraînement que de les augmenter artificiellement. Ils ont aussi proposé d'envisager une approche modulaire pour le modèle statistique, et une technique d'augmentation de données basée sur des règles pour le modèle neuronal.

Approches par apprentissage profond Tiwari & Naskar (2017) ont proposé un modèle encodeur-décodeur de réseaux de neurones récurrents (*Recurrent Neural Network, RNN*) à mémoire court et long terme (*Long Short-Term Mermory, LSTM*) avec un méchanisme d'attention, et ils ont aussi créé des données artificielles pour entraîner ce modèle. Lourentzou *et al.* (2019) ont introduit un modèle hybride encodeur-décodeur à base de mots et de caractères, la composante à base de caractères étant entraîné sur des exemples antagonistes synthétiques. Stewart *et al.* (2019) ont utilisé un modèle de réseaux de neurones récurrents à portes (*Gated Recurrent Unit, GRU*) au niveau des mots et ont présenté de meilleures performances que les modèles au niveau des caractères et d'autres méthodes par apprentissage profond existantes.

Modèles détecteur-correcteur Une autre approche consiste à découpler la détection des mots non standards de leur normalisation lexicale. Par exemple, Leeman-Munk *et al.* (2015) ont utilisé deux modèles de réseaux de neurones à propagation avant (*Feed-forward Neural Network, FFNN*) augmentés : un « signaleur » pour identifier les tokens à normaliser et un « normalisateur » qui corrige un token à la fois. Par ailleurs, Tian *et al.* (2017) ont proposé un modèle de réseaux de neurones convolutifs (*Convolutional Neural Network, CNN*) à base de caractères pour la détection de mots non standards dans les tweets. Cette étape de détection est prévue en amont d'une normalisation lexicale.

Modèles de langue pré-entraînés Muller *et al.* (2019) ont apporté des modifications à l'architecture de BERT (Devlin *et al.*, 2019) et l'ont affiné pour la normalisation lexicale en tant que tâche de prédiction de tokens. Par ailleurs, Scherrer & Ljubešić (2021) ont proposé un système basé sur un modèle BERT affiné pour la classification de tokens qui prédit le type de transformation nécessaire pour corriger le mot non standard. Un modèle de traduction automatique à base de caractères est utilisé pour appliquer les corrections proposées par BERT. De plus, Bucur *et al.* (2021) ont aussi considéré la normalisation lexicale comme une tâche de traduction et ont proposé un modèle séquentiel au niveau de la phrase basé sur mBART (Liu *et al.*, 2020). Kubal & Nagvenkar (2021) ont quant à eux affiné un modèle BERT multilingue (Devlin *et al.*, 2019) pour la normalisation lexicale comme une tâche d'étiquetage de séquences et l'ont combiné avec une technique d'alignement de mots. Ainsi, ils ont pu utiliser le même modèle pour effectuer la normalisation sur plusieurs langues. van der Goot & Çetinoğlu (2021) ont aussi utilisé un modèle BERT multilingue dans la normalisation lexicale de CGU présentant de l'alternance codique, notamment pour l'identification de langue.

3.2 Tâches connexes

Nous avons vu que la normalisation lexicale peut être assimilée à une tâche de traduction d'une version non standard d'une langue vers sa version standard. En effet, Kobus *et al.* (2008) ont catégorisé les approches de normalisation lexicale de SMS selon trois « métaphores » : la correction orthographique, la traduction et la reconnaissance de la parole. De même, il existe aussi d'autres tâches qui sont plus ou moins connexes à la normalisation lexicale et qui peuvent lui inspirer d'autres approches, à savoir :

La normalisation phonétique Elle peut être considérée comme une sous-tâche de la normalisation lexicale puisque certains des phénomènes non standards observés sont en effet d'ordre phonétique.

D'ailleurs, certaines méthodes décrites dans la section 3.1 intègrent un module de calcul de similarité phonétique. Rosales Núñez *et al.* (2019b) ont proposé un modèle de normalisation phonétique pour améliorer la traduction des CGU du français vers l'anglais. Cette tâche est particulièrement utile pour normaliser les CGU dans des langues riches en homophonies comme le chinois (Qin *et al.*, 2021). Elle a aussi été appliquée à la correction orthographique dans les moteurs de recherche du commerce en ligne (Yang *et al.*, 2022).

La correction post-OCR et post-ASR Les textes résultant de la reconnaissance optique de caractères (*Optical Character Recognition, OCR*) doivent souvent être corrigés en post-traitement car ils contiennent des caractères mal reconnus et donc des mots non standards. De même, les transcriptions résultant de la reconnaissance automatique de la parole (*Automatic Speech Recognition, ASR*) contiennent des mots non standards provenant des phonèmes mal compris.

La correction d'erreurs grammaticales En contrepartie de la tâche normalisation lexicale, qui vise à corriger les erreurs d'ordre lexical, la correction grammaticale vise à corriger les erreurs d'ordre grammatical. Elle est aussi souvent découpée en deux sous-tâches : détection et correction. En pratique, la frontière entre erreur lexicale et erreur grammaticale n'est pas bien définie dans les CGU : certains phénomènes peuvent appartenir aux deux classes. Cependant, les annotateurs de données de normalisation lexicale essaient de se limiter à corriger les mots non standards d'un point de vue lexical, même si la phrase résultante reste agrammaticale.

La normalisation de variantes linguistiques En comparant grossièrement le langage non standard des CGU à un « dialecte » du langage standard, la tâche de normalisation lexicale peut être assimilée à celle de la normalisation de variantes linguistiques. En particulier, certains travaux sur la normalisation de dialectes (Partanen *et al.*, 2019) et des créoles (Liu *et al.*, 2022), de textes produits par des locuteurs non natifs (Sarkar *et al.*, 2020; Alam & Anastasopoulos, 2020), et de langue non contemporaine (Bawden *et al.*, 2022) peuvent s'avérer intéressants pour la normalisation lexicale.

La simplification de textes Un parallèle peut être établi entre cette tâche et la normalisation lexicale si nous considérons que le lexique non standard des CGU, difficile à comprendre en dehors d'un public restreint, doit être renvoyé vers un lexique standard, plutôt facile à comprendre et accessible à un plus grand public.

3.3 Évaluation

Bien que la normalisation lexicale soit une bonne solution pour le problème des mots non standards dans les CGU, elle reste une tâche qui est difficile à évaluer en raison du manque de ressources annotées d'une part, et du manque d'homogénéité dans le choix des conventions d'annotation et des métriques utilisées.

Ressources Malgré l'abondance de CGU sur internet, peu de données parallèles annotées pour la normalisation lexicale sont disponibles [4]. Néanmoins, la campagne d'évaluation MultiLexNorm2021 (van der Goot *et al.*, 2021) comprend des données annotées en plusieurs langues issues d'autres campagnes d'évaluations, à savoir : en danois [DA] (Plank *et al.*, 2020), en allemand [DE] (Sidarenka *et al.*, 2013), en anglais [EN] (Baldwin *et al.*, 2015), en espagnol [ES] (Alegria *et al.*, 2013), en croate [HR] (Ljubešić *et al.*, 2017a), en indonésien-anglais [ID-EN] (Barik *et al.*, 2019), en italien [IT] (van der Goot *et al.*, 2020), en néerlandais [NL] (Schuur, 2020), en slovène [SL] (Erjavec *et al.*, 2017),

4. Bikaun *et al.* (2021) ont créé un bon outil d'annotation à cet effet.

en serbe [SR] (Ljubešić *et al.*, 2017b), en turc [TR] (Çolakoğlu *et al.*, 2019), et en alternance codique turc-allemand [TR-DE] (van der Goot & Çetinoğlu, 2021). D'autres données parallèles annotées sont disponibles en anglais et en néerlandais (De Clercq *et al.*, 2014), et en japonais (Higashiyama *et al.*, 2021). Il existe aussi des données parallèles, non pas pour la normalisation lexicale en soi, mais pour l'évaluation des tâches en aval comme la traduction de CGU (Michel & Neubig, 2018; Rosales Núñez *et al.*, 2019a, 2021a).

Métriques Plusieurs métriques ont été utilisées pour évaluer la tâche de normalisation lexicale : le taux d'erreur de mots (*Word Error Rate, WER*) (Sproat *et al.*, 2001; Kobus *et al.*, 2008; Matos Veliz *et al.*, 2019) ; le taux d'erreur de phrases (*Sentence Error Rate, SER*) (Kobus *et al.*, 2008) ; le taux de couverture (Liu *et al.*, 2012) ; l'exactitude, la précision, le rappel et la F-mesure (Baldwin *et al.*, 2015) ; la précision sur les mots hors vocabulaire (Alegria *et al.*, 2013) ; et le BLEU (Aw *et al.*, 2006; Kobus *et al.*, 2008; Han & Baldwin, 2011), qui est une métrique de traduction. van der Goot (2019b) a considéré que ces métriques sont soit trop complexes pour la tâche, soit difficiles à interpréter et à comparer. Il a donc préconisé l'utilisation du « taux de réduction de l'erreur » (*Error Reduction Rate, ERR*), qui correspond à l'exactitude normalisée par le nombre de mots remplacés :

$$\text{ERR} = \frac{\%\text{exactitude} - \%\text{mots non normalisés}}{100 - \%\text{mots non normalisés}} \tag{1}$$

Cette métrique a été définie par van der Goot (2019c) et utilisée dans la campagne d'évaluation MultiLexNorm2021 (van der Goot *et al.*, 2021). Elle permet de comparer la performance d'un modèle sur plusieurs jeux de données différents, voire plusieurs langues.

4 Étude expérimentale

Dans cette partie, nous allons réaliser une étude expérimentale préliminaire pour illustrer la normalisation lexicale de CGU décrite dans l'état de l'art précédent. Nous allons comparer des modèles de langue pré-entraînés pour montrer la difficulté de la tâche (il est difficile de faire mieux qu'un modèle de base qui ne fait rien !) et celle d'en trouver une métrique adéquate.

4.1 Données

Nous avons utilisé les données parallèles de la campagne d'évaluation LexNorm2015 (Baldwin *et al.*, 2015). Elles consistent en tweets publiés en anglais, alignés avec leurs normalisations lexicales.

Jst read a tweet **lol** and **l o v e** it
↓
just read a tweet **laughing out loud** and **love** it

FIGURE 1 – Un exemple de tweet en anglais avec sa normalisation lexicale.
(Traduction : *Je viens de lire un tweet, mort de rire, et j'adore !*)

La figure 1 est un exemple de tweet normalisé. Elle illustre les trois types de normalisations lexicales distingués par les organisateurs de la campagne d'évaluation :

— la normalisation **1-1** : un mot non standard est remplacé par un mot standard (`Jst` → `just`) ;

— la normalisation **1-N** : un mot non standard est remplacé par plusieurs mots standards (`lol` → `laughing out loud`);
— la normalisation **N-1** : une séquence non standard de mots ou de sous-mots est remplacée par un seul mot standard (`l o v e` → `love`).

Les données étaient déjà prétokénisées [5] (par espaces et ponctuation), tout en tenant compte des spécificités de Twitter (liens URL, hashtags, mentions). Ces dernières n'ont pas été modifiées lors de l'annotation. Par ailleurs, les normalisations effectuées étaient toutes en minuscules (`Jst` → `just`) [6].

Les données LexNorm2015 comprennent un jeu d'entraînement et un de test. Le tableau 1 résume les statistiques [7] de ces deux jeux de données : le nombre de tweets et le pourcentage de mots normalisés.

Remarque. Les tweets étant déjà tokénisés, nous appelons « mot » toute suite de caractères délimitée par un espace. Ainsi, la séquence `l o v e` comprend quatre mots. En revanche, elle correspond à une seule occurrence dans le nombre de normalisations N-1.

Jeu de données	**# tweets**	**% mots normalisés**	*dont*		
			% 1-1	*% 1-N*	*% N-1*
Entraînement	2950	8,85	*73,25*	*26,55*	*0,20*
Test	1967	9,40	*73,92*	*25,68*	*0,40*

TABLE 1 – Statistiques des données LexNorm2015.

4.2 Modèles

Le pourcentage de mots à normaliser étant inférieur à 10% pour les données d'entraînement et de test (cf. le tableau 1), nous avons opté pour un modèle détecteur-correcteur afin de cibler la normalisation lexicale uniquement sur les mots qui la nécessitent. En plus, notre approche combine deux autres des méthodes citées dans l'état de l'art : elle est basée sur des modèles de langue par masquage (MLM) pré-entraînés, et elle est modulaire (elle inclut une distance d'édition).

4.2.1 Le détecteur

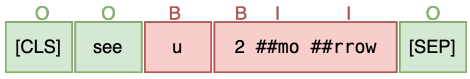
La détection de mots à normaliser peut être assimilée à une tâche de classification de tokens : pour chaque token de la phrase source, le modèle de détection doit prédire une étiquette qui correspond à sa classe d'appartenance (ici, standard ou non standard).

Concrètement, nous avons pris le MLM pré-entraîné BERT [8] (Devlin *et al.*, 2019) et nous l'avons affiné pour la tâche de classification de tokens sur le jeu d'entraînement de LexNorm2015. Nous avons utilisé le schéma d'étiquetage **B-I-O** : **B** (Beginning, *début*) pour étiqueter le premier token d'un mot non standard, **I** (Inside, *intérieur*) pour un token à l'intérieur d'un mot non standard, et **O** pour les tokens des mots standards.

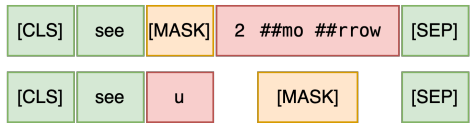
5. par le tokeniseur CMU-ARK (https://github.com/myleott/ark-twokenize-py).
6. Un choix discutable des annotateurs.
7. calculées par nous.
8. https://huggingface.co/bert-base-uncased

Prenons l'exemple de la phrase `see` **`u 2morrow`** (→ `see` **`you tomorrow`**, *à demain*). La figure 2a illustre sa tokenisation et son étiquetage B-I-O des tokens [9] par notre détecteur BERT.

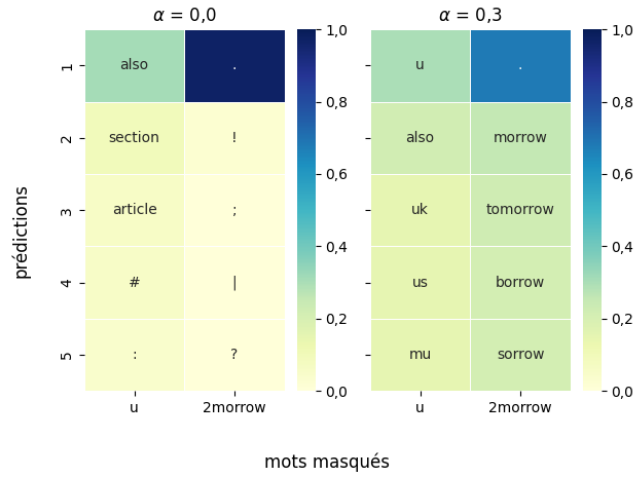
Remarque. Il est pertinent de noter qu'un « mot », tel que défini dans la section 4.1, peut être tokénisé en plusieurs tokens différents par chaque MLM. Ainsi, même si le détecteur dans la figure 2a a étiqueté quatre tokens comme non standards, ceux-ci correspondent seulement à deux mots à normaliser.



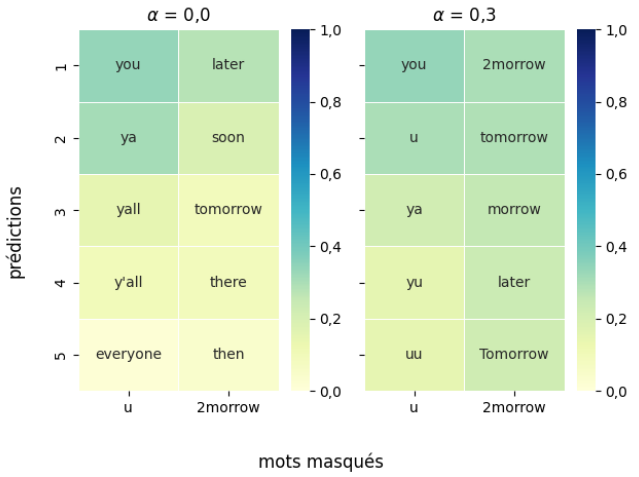
(a) Détection de mots à normaliser par un BERT affiné



(b) Stratégie de masquage des mots à normaliser



(c) BERT



(d) BERTweet

FIGURE 2 – Détection (a) et masquage (b) de mots à normaliser par un détecteur BERT. Prédictions des mots masqués par les correcteurs BERT (c) et BERTweet (d) pour $\alpha = 0{,}0$ et $0{,}3$, triées par score. (Normalisation attendue : `see` **`u 2morrow`** → `see` **`you tomorrow`**, *à demain*)

4.2.2 Le correcteur

Après avoir détecté les mots à normaliser, il reste à les corriger. Une approche consiste à masquer ces mots et à utiliser un MLM pré-entraîné pour les prédire selon le contexte environnant. Ainsi, en fonction du vocabulaire auquel le MLM a été exposé pendant son entraînement, les prédictions appartiendraient à un lexique standard et, par conséquent, les mots non standards seraient normalisés.

L'inconvénient de cette approche est le fait que les modèles de langue soient entraînés à prédire *un* token [10] qui convient au sens du contexte. Il peut donc y avoir plusieurs bonnes prédictions possibles d'un point de vue sémantique, mais aucune du point de vue lexical. Nous avons donc utilisé une distance d'édition pour guider le MLM vers une prédiction non seulement sémantiquement appropriée, mais aussi lexicalement proche du mot masqué.

Concrètement, nous avons défini un nouveau score, qui est une combinaison linéaire, de paramètre $\alpha \in [0, 1]$, entre le score du MLM et la distance de Damerau-Levenshtein (Damerau, 1964) [11]

9. Les tokens spéciaux `[CLS]` et `[SEP]` sont automatiquement considérés comme standards.
10. selon la tokénisation en sous-mots du MLM.
11. choisie car elle considère l'opération de transposition, contrairement à la distance de Levenshtein (Levenshtein, 1966).

normalisée sur une échelle de $[0, 1]$.

Soient $\mathbf{x} = (x_1 \dots x_\ell)$ la séquence source, x_i le token à normaliser, et $\bar{\mathbf{x}} = (x_1 \dots x_{i-1}, [\text{MASK}], x_{i+1} \dots x_\ell)$ la séquence source masquée au token x_i. Soient $\text{MLM}(\bar{\mathbf{x}})$ le vecteur de scores prédits par le modèle de langue pour tous les tokens de son vocabulaire $\mathcal{V}$, et $\text{Lev}_{\text{norm}}(x_i, \mathcal{V})$ le vecteur des distances de Damerau-Levenshtein normalisées entre x_i et tous les tokens du vocabulaire. Alors, la prédiction $\hat{x}_i$ du correcteur est :

$$\hat{x}_i = \arg\max_{\mathcal{V}} \quad [(1-\alpha)\text{MLM}(\bar{\mathbf{x}}) + \alpha(1 - \text{Lev}_{\text{norm}}(x_i, \mathcal{V}))] \tag{2}$$

Les MLM pré-entraînés que nous avons comparés pour la correction sont :

1. **BERT** : entraîné sur le corpus de livres BookCorpus et Wikipedia, tous des textes édités, donc « standards » ;
2. **RoBERTa** [12] (Liu *et al.*, 2019) : entraîné sur les mêmes données que BERT, en plus d'autres ressources éditées (CC-News, OpenWebText, Stories). Il a la même architecture que BERT mais avec des optimisations, et a été entraîné plus longtemps ;
3. **ELECTRA** [13] (Clark *et al.*, 2020) : ayant la même architecture que BERT et entraîné sur les mêmes données, mais avec un objectif d'entraînement différent : c'est un modèle discriminant entraîné à distinguer les tokens remplacés des tokens d'origine masqués ;
4. **BERTweet** [14] (Nguyen *et al.*, 2020) : identique à BERT, mais entraîné entièrement sur des données Twitter ;
5. **Twitter RoBERTa** [15] (Barbieri *et al.*, 2020) : le modèle RoBERTa pour lequel l'entraînement a été poursuivi sur des données Twitter.

Tous ces modèles ont été entraînés sur des données en anglais. Parmi les cinq, seuls BERTweet et Twitter Roberta ont été exposés à des CGU issus de Twitter ; nous les avons choisis sous l'hypothèse qu'ils seraient plus robustes au bruit des CGU. Le tableau 2 résume les propriétés de ces MLM.

Modèle	**Données d'entraînement**	**Taille du vocabulaire**
BERT	*BookCorpus, Wikipedia*	30 522
RoBERTa	= BERT + *CC-News, OpenWebText, Stories*	50 265
ELECTRA	= BERT	30 522
BERTweet	***Twitter***	64 000
Twitter RoBERTa	= RoBERTa + ***Twitter***	50 265

TABLE 2 – Propriétés des modèles de langue pour la correction.

Revenons à l'exemple de la phrase `see` **`u 2morrow`**. La figure 2b illustre la stratégie de masquage des deux mots à normaliser détectés : la séquence source est dupliquée autant de fois que de mots à normaliser. Pour chacune de ces copies, un mot à normaliser à la fois est masqué, c'est-à-dire remplacé par le token de masquage du MLM (`[MASK]` pour BERT).

12. https://huggingface.co/roberta-base
13. https://huggingface.co/google/electra-base-generator
14. https://huggingface.co/vinai/bertweet-base
15. https://huggingface.co/cardiffnlp/twitter-roberta-base

Après le masquage, les séquences sont passées au correcteur. Les figures 2c et 2d illustrent les cinq meilleures prédictions des modèles BERT et BERTweet, triées par score. Elles montrent en particulier les sorties des MLM purs ($\alpha = 0{,}0$) et celles des MLM + distance d'édition ($\alpha = 0{,}3$) [16].

Nous remarquons que, sans prendre en compte le mot masqué ($\alpha = 0{,}0$), les deux modèles prédisent des mots lexicalement divers et variés. Par contre, BERTweet prédit des mots plus sémantiquement adaptés. Par exemple, lorsque **`2morrow`** est masqué, BERT ne privilégie que des signes de ponctuation (et il est très confiant pour le point) alors que BERTweet prédit *à plus tard* (**`later`**), *à bientôt* (**`soon`**) et *à demain* (**`tomorrow`**), la cible, en troisième place. En outre, il normalise déjà **`u`** en **`you`**. Cela pourrait être attribué au fait qu'il ait été entraîné sur des données Twitter. Il est donc plus robuste face à une phrase source avec des mots non standards.

En combinant les MLM avec la distance d'édition ($\alpha = 0{,}3$), leurs prédictions se rapprochent lexicalement du mot masqué. À la place de **`2morrow`**, BERT prédit toujours un point en premier, mais avec moins de confiance, et ensuite des mots proches lexicalement (dont la cible en troisième place). D'autre part, BERTweet exhibe un réordonnancement des prédictions, avec la cible en deuxième place. Il prédit en premier le mot même à normaliser : comme il fait déjà partie du lexique de BERTweet, ce dernier le considère comme standard et, par conséquent, ne le normalise pas.

4.3 Expériences

Dans un premier temps, nous avons entraîné le détecteur (cf. la section 4.2.1) sur le jeu d'entraînement de LexNorm2015 et évalué sa performance de détection de mots à normaliser sur le jeu de test. Ensuite, nous avons comparé les cinq MLM pré-entraînés (cf. la section 4.2.2) pour la correction des mots détectés dans le jeu de test. Nous avons choisi comme système de base le modèle « identité » (qui consiste à ne rien changer). Pour chaque correcteur, nous avons généré les prédictions en tenant de plus en plus compte de la distance d'édition par rapport aux mots masqués, c'est-à-dire en augmentant la valeur α (de 0 à 1 par pas de $0{,}1$).

Nous avons utilisé les métriques automatiques (exactitude, précision, rappel et F-mesure) pour évaluer le détecteur. Pour les correcteurs, nous nous aussi sommes servi des métriques automatiques, comme dans la campagne d'évaluation LexNorm2015 (Baldwin *et al.*, 2015), et de l'ERR utilisée dans la campagne d'évaluation MultiLexNorm2021 (van der Goot *et al.*, 2021). En outre, nous avons évalué les sorties normalisées par deux métriques usuelles de traduction : BLEU (Papineni *et al.*, 2002), en particulier l'implémentation SacreBLEU [17] (Post, 2018), et COMET [18] (Rei *et al.*, 2020). Afin de pouvoir comparer les différents modèles, toutes les métriques ont été effectuées à l'échelle des mots.

4.4 Résultats

Nous avons évalué la performance du détecteur sur le jeu de test de LexNorm2015 et nous avons obtenu les scores suivants : $97{,}82\%$ d'exactitude, $90{,}14\%$ de précision, $86{,}41\%$ de rappel et $88{,}24\%$ de F-mesure. Dans les sections suivantes, nous analyserons la performance des différents correcteurs.

16. Nous avons choisi la valeur $0{,}3$ car elle obtient quasiment les meilleurs scores dans les expériences (cf. la section 4.4).
17. https://huggingface.co/spaces/evaluate-metric/sacrebleu
18. https://huggingface.co/spaces/evaluate-metric/comet

4.4.1 Analyse qualitative

Le tableau 3 illustre les normalisations d'un tweet du jeu de test LexNorm2015 par les correcteurs BERT et BERTweet. Nous avons choisi les sorties pour $\alpha = 0$ (MLM seul), $\alpha = 0{,}3$ (MLM + distance d'édition) et $\alpha = 1$ (distance d'édition seule).

Source	rt @tehreelhov : **wen ur** at a restaurant **nd u c ur** food **comin** http://t.co/ducpxt7dry
Cible	rt @tehreelhov : **when you're** at a restaurant **and you see your** food **coming** http://t.co/ducpxt7dry
$\alpha = 0$	rt @tehreelhov : r ##d at a restaurant . : : a food . http://t.co/ducpxt7dry
$\alpha = 0{,}3$	rt @tehreelhov : wen ur at a restaurant **and** u c ur food **coming** http://t.co/ducpxt7dry
$\alpha = 1$	rt @tehreelhov : wen ur at a restaurant **and** u c ur food **coming** http://t.co/ducpxt7dry

(a) BERT

Source	rt @tehreelhov : **wen ur** at a restaurant **nd u c ur** food **comin** http://t.co/ducpxt7dry
Cible	rt @tehreelhov : **when you're** at a restaurant **and you see your** food **coming** http://t.co/ducpxt7dry
$\alpha = 0$	rt @tehreelhov : **when** ur at a restaurant **and** u **see** ur food <@@ http://t.co/ducpxt7dry
$\alpha = 0{,}3$	rt @tehreelhov : **when** ur at a restaurant **and** u **see** ur food comin http://t.co/ducpxt7dry
$\alpha = 1$	rt @tehreelhov : wen ur at a restaurant nd u c ur food comin http://t.co/ducpxt7dry

(b) BERTweet

TABLE 3 – Normalisation d'un tweet par les correcteurs BERT et BERTweet pour $\alpha = 0$, 0,3 et 1.
(Traduction : *Retweet [Utilisateur] : quand tu es dans un restaurant et tu vois ta nourriture arriver [Lien URL]*)

BERT (cf. le tableau 3a) Lorsque le correcteur comprend le MLM uniquement ($\alpha = 0$), les mots à normaliser sont remplacés par des tokens fréquents (des lettres ou de la ponctuation) qui ne leur sont ni lexicalement ni sémantiquement proches. D'ailleurs, nous pouvons avancer que le sens du tweet d'origine se dégrade, voire se perd complètement.

En combinant le MLM avec la distance d'édition ($\alpha = 0{,}3$), seulement deux mots sont correctement normalisés. Même si les autres mots à normaliser restent inchangés (puisqu'ils appartiennent au vocabulaire de BERT), cette sortie est une amélioration car le modèle a corrigé quelques mots sans y introduire plus de bruit.

Enfin, lorsque nous prenons la distance d'édition seule ($\alpha = 1$), le correcteur prédit les mêmes sorties obtenues pour $\alpha = 0{,}3$. Comme BERT a été entraîné sur des données standards, nous conjecturons qu'il n'a pas réussi à se représenter correctement la phrase source (assez bruitée) et n'a donc pas su prédire de tokens pertinents. Par conséquent, il semble que toute la normalisation avait été effectuée par le calcul de la distance d'édition.

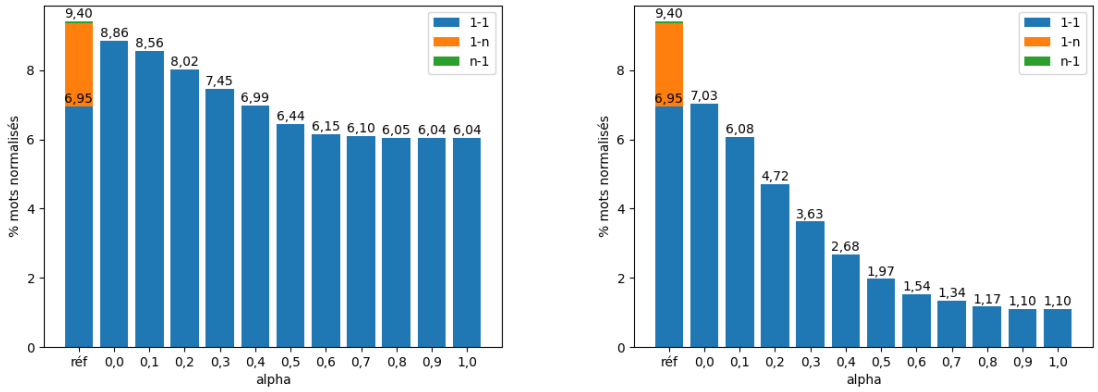
BERTweet (cf. le tableau 3b) Ayant été entraîné sur des données Twitter, BERTweet est plus apte à modéliser la phrase source et à prédire des tokens sémantiquement pertinents à la place des mots masqués. À lui tout seul ($\alpha = 0$), il normalise déjà correctement trois mots, et certains mots restent inchangés car ils appartiennent à son vocabulaire. Par ailleurs, il introduit un seul token erroné **<@@** [19].

Similairement à BERT, lorsque BERTweet est combiné avec la distance d'édition ($\alpha = 0{,}3$), il n'introduit plus de bruit : il remplace tous les tokens précédemment mal prédits par les mots masqués.

Lorsque $\alpha = 1$, la sortie du correcteur est aussi bruitée que la phrase source : tous les mots non

19. qui se colle au lien URL pour former `<http://t.co/ducpxt7dry`.

standards du tweet font partie du vocabulaire. Nous perdons aussi toutes les normalisations que ce dernier avait réussi à réaliser. Nous pouvons donc conclure que la distance d'édition ne suffit pas pour corriger la phrase. Par contre, nous voyons aussi les limites de cette distance : la normalisation attendue pour **c** est **see**, qui ne lui est pas lexicalement proche [20] alors qu'elle l'est phonétiquement.



(a) BERT

(b) BERTweet

FIGURE 3 – Statistiques des mots normalisés par les correcteurs BERT et BERTweet.

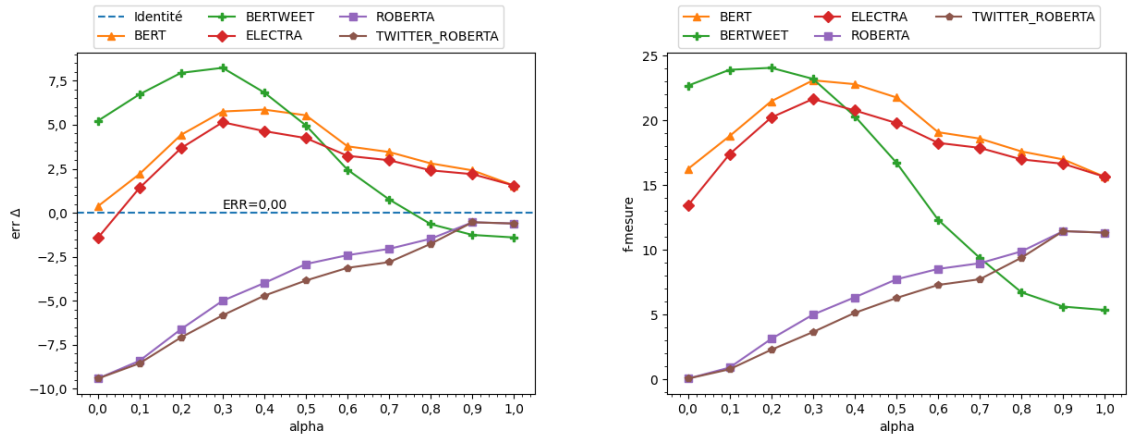
Ces phénomènes sont aussi observables sur la figure 3. D'une part, BERT tout seul « normalise » plus de mots que nécessaire : il introduit du bruit dans la phrase. Or, plus le poids de la distance d'édition augmente, plus le pourcentage de mots normalisés se rapproche de la référence. D'autre part, BERTweet tout seul modifie presque autant de mots que la référence. Cependant, le pourcentage de mots normalisés diminue exponentiellement lorsque α augmente.

Remarque. Les correcteurs utilisés sont à base de MLM entraînés pour prédire un seul token pour chaque masque. Les normalisations de type 1-N et N-1 ne sont donc pas réalisables. De plus, même les normalisations 1-1 réalisées sont parfois suboptimales car les modèles ne peuvent prédire que des mots composés d'un seul token. Ils prédisent même parfois des tokens isolés comme **##d** et **<@@** dans les tableaux 3a et 3b respectivement.

4.4.2 Analyse quantitative avec métriques automatiques

La figure 4 illustre les scores d'ERR (4a) et de F-mesure (4b) des correcteurs considérés pour différentes valeurs de α. Nous remarquons trois comportements différents. D'abord, BERTweet a déjà des scores au dessus du système de base pour $\alpha = 0$. Cela rejoint l'hypothèse qu'il est plus robuste au bruit et qu'il arrive à extraire des informations sémantiques des tweets malgré leurs mots non standards. En le combinant avec la distance d'édition, les scores augmentent pour les premières valeurs de $\alpha \leq 0{,}3$ avant de dégrader progressivement jusqu'en dessous du système de base. Ensuite, BERT et ELECTRA (qui lui reste légèrement en dessous), sont proches du système de base pour $\alpha = 0$. Ils augmentent progressivement lorsque $\alpha \leq 0{,}5$ et se dégradent tout en restant au dessus du système de base. Finalement, RoBERTa et Twitter RoBERTa restent toujours en dessous du système de base, même si leurs scores augmentent avec α. Lorsque $\alpha = 1$, nous observons la convergence des modèles qui ont été entraînés sur les mêmes données et la même taille de vocabulaire.

20. La distance de Damerau-Levenshtein entre **c** et **see** est 3 (une substitution et 2 insertions).



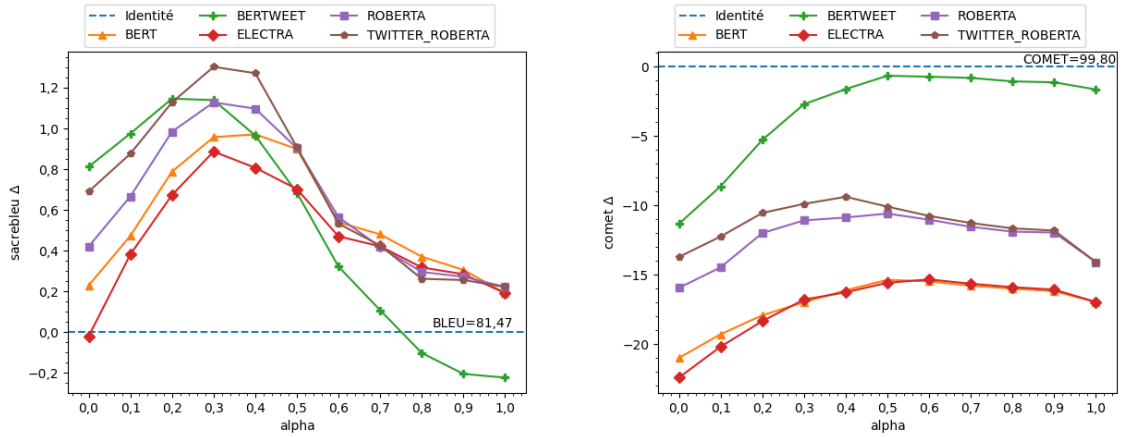
(a) ERR (b) F-mesure

FIGURE 4 – Évaluation des correcteurs par métriques automatiques.

Les courbes de scores d'exactitude suivent les mêmes tendances que celles de l'ERR. De même, celles de précision et de rappel suivent les mêmes tendances que celles de la F-mesure.

Remarque. Comme le système de base ne change rien aux phrases sources, il n'y a pas de mots normalisés (pas de prédictions positives). Nous ne pouvons donc rien conclure sur la précision, ni calculer la F-mesure (d'où l'absence du modèle identité sur la figure 4b).

4.4.3 Analyse quantitative avec métriques de traduction



(a) BLEU (b) COMET

FIGURE 5 – Évaluation des correcteurs par métriques de traduction.

La figure 5 illustre les scores BLEU (5a) et COMET (5b) des correcteurs pour différentes valeurs de α. Nous observons des comportements très différents sur ces deux métriques. D'une part, tous les modèles montrent une amélioration de score BLEU par rapport au modèle identité (sauf BERTweet lorsque α est grand). Nous observons des pics autour des valeurs de $\alpha = 0{,}3_{\pm 0{,}1}$. En plus, les modèles qui ont été exposés à des données Twitter lors de leur entraînement (BERTweet et Twitter RoBERTa)

ont les meilleurs scores, suivis de RoBERTa, puis BERT et ELECTRA. D'autre part, nous remarquons que tous les scores COMET augmentent progressivement avec α mais restent toujours en dessous du modèle de base. En particulier, BERTweet est le meilleur, suivi de RoBERTa et Twitter RoBERTa, et enfin de BERT et ELECTRA.

Les différences observées entre les deux métriques sont frappantes (pour BLEU, la plupart des scores sont supérieurs à la ligne de base, alors que pour COMET, ils sont tous inférieurs), mais elles s'expliquent simplement : alors que BLEU est une métrique de *surface* qui repose sur le comptage de n-grammes partagés entre une référence et une prédiction, COMET est une métrique neuronale dont l'objectif est de dépasser la variation de surface pour comparer le *sens* des deux textes. Selon BLEU, une prédiction erronée qui est plus éloignée de la forme normalisée de référence ne sera pas plus fortement pénalisée. D'autre part, la fonction identité obtient un score COMET élevé parce que celui-ci est assez robuste pour juger que le texte de départ est sémantiquement très similaire à la référence normalisée. À l'inverse, toute erreur introduite par un modèle de normalisation sera sanctionnée à des scores COMET plus faibles. Ainsi, pour la tâche de normalisation, aucune de ces mesures n'est totalement adéquate à elle seule ; il serait donc intéressant de les examiner ensemble.

5 Discussion et Conclusion

Cet article a pour vocation d'illustrer l'utilité mais aussi les limites de la tâche de normalisation lexicale des contenus générés par les utilisateurs (CGU), et d'ouvrir la porte à plusieurs perspectives de travaux de recherche. Dans un premier temps, nous avons présenté un état de l'art de la normalisation lexicale des CGU sur les réseaux sociaux. Nous avons montré qu'ils sont un fléau pour les modèles de TAL entraînés sur des données standards à cause de leur multitude de phénomènes de langage non standard, et que la normalisation lexicale est l'une des approches pratiques pour pallier ce problème.

Dans un second temps, nous avons fait une étude expérimentale pour montrer que, malgré tous ses avantages, la normalisation lexicale de CGU reste une tâche difficile à réaliser et à évaluer. Premièrement, nous observons que, même avec des modèles de langue pré-entraînés, il est difficile de faire mieux qu'un système de base qui ne fait rien. Certes, l'intégration d'une distance d'édition et l'entraînement sur des données CGU améliorent la performance des modèles. Ensuite, nous observons l'inadéquation des métriques classiques pour ce genre de tâche. Par exemple, le pire modèle pour les métriques automatiques (Twitter RoBERTa) est l'un des meilleurs pour la métrique de traduction BLEU, mais n'a pas de bon score pour une autre métrique de traduction, COMET.

Nos expériences étant préliminaires, plusieurs pistes d'amélioration peuvent être envisagées, notamment : filtrer les sorties de BERTweet par un lexique standard, apprendre la valeur optimale de α pendant l'entraînement, intégrer un module de similarité phonétique, utiliser des MLM qui masquent tous les tokens d'un mot (*whole-word masking*) ou plusieurs tokens adjacents (*span masking*), affiner ELECTRA pour la détection de mots à normaliser (Yuan *et al.* (2021) ont fait l'hypothèse que c'est un meilleur détecteur car il a un objectif d'entraînement discriminant), normaliser la séquence autorégressivement (Sun & Jiang, 2019), garder le mot d'origine si la normalisation prédite est pire (van der Goot, 2019b), etc.

Remerciements

Un grand merci à mes encadrants Rachel Bawden et Benoît Sagot pour leur soutien, aux relecteurs RECITAL pour leurs commentaires précieux, et à Menel Mahamdi pour sa relecture judicieuse. Ce travail a été financé par la chaire de R. Bawden à l'institut PRAIRIE financé par l'agence nationale française ANR dans le cadre du programme "Investissements d'avenir" sous la référence ANR-19-P3IA-0001 et également par le projet Emergence, DadaNMT, financé par Sorbonne Université.

Références

AHMED B. (2015). Lexical normalisation of Twitter Data. In *Proceedings of the 2015 Science and Information Conference*, p. 326–328, London, UK : IEEE. DOI : 10.1109/SAI.2015.7237164.

ALAM M. M. I. & ANASTASOPOULOS A. (2020). Fine-Tuning MT systems for Robustness to Second-Language Speaker Variations. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, p. 149–158, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.wnut-1.20.

ALEGRIA I., ARANBERRI N., FRESNO-FERNÁNDEZ V., GAMALLO P., PADRÓ L., VICENTE I. S., TURMO J. & ZUBIAGA A. (2013). Introducción a la Tarea Compartida Tweet-Norm 2013 : Normalización Léxica de Tuits en Español. In *Proceedings of the XXIX Congreso de la Sociedad Española de Procesamiento de Lenguaje Natural*, p. 38–46, Madrid, Spain.

AMINIAN M., AVONTUUR T., BALEMANS I., ELSHOF L., NEWELL R., NOORD N. V., NTAVELOS A., VAN ZAANEN M. & AZAR E. Z. (2012). Assigning part-of-speech to Dutch tweets.

AW A., ZHANG M., XIAO J. & SU J. (2006). A phrase-based statistical model for SMS text normalization. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, p. 33–40, Sydney, Australia : Association for Computational Linguistics.

BALDWIN T., COOK P., LUI M., MACKINLAY A. & WANG L. (2013). How noisy social media text, how diffrnt social media sources ? In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, p. 356–364, Nagoya, Japan : Asian Federation of Natural Language Processing.

BALDWIN T., DE MARNEFFE M. C., HAN B., KIM Y.-B., RITTER A. & XU W. (2015). Shared tasks of the 2015 workshop on noisy user-generated text : Twitter lexical normalization and named entity recognition. In *Proceedings of the Workshop on Noisy User-generated Text*, p. 126–135, Beijing, China : Association for Computational Linguistics. DOI : 10.18653/v1/W15-4319.

BARANES M. & SAGOT B. (2014). Analogy-based text normalization : the case of unknowns words (normalisation de textes par analogie : le cas des mots inconnus) [in French]. In *Proceedings of TALN 2014 (Volume 1 : Long Papers)*, p. 137–148, Marseille, France : Association pour le Traitement Automatique des Langues.

BARBIERI F., CAMACHO-COLLADOS J., ESPINOSA ANKE L. & NEVES L. (2020). TweetEval : Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics : EMNLP 2020*, p. 1644–1650, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.findings-emnlp.148.

BARIK A. M., MAHENDRA R. & ADRIANI M. (2019). Normalization of Indonesian-English Code-Mixed Twitter Data. In *Proceedings of the 5th Workshop on Noisy User-generated Text*

(W-NUT 2019), p. 417–424, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-5554.

BAWDEN R., POINHOS J., KOGKITSIDOU E., GAMBETTE P., SAGOT B. & GABAY S. (2022). Automatic normalisation of early Modern French. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 3354–3366, Marseille, France : European Language Resources Association.

BECKLEY R. (2015). Bekli :A Simple Approach to Twitter Text Normalization. In *Proceedings of the Workshop on Noisy User-generated Text*, p. 82–86, Beijing, China : Association for Computational Linguistics. DOI : 10.18653/v1/W15-4312.

BELINKOV Y. & BISK Y. (2017). Synthetic and Natural Noise Both Break Neural Machine Translation. *ICLR*.

BIKAUN T., FRENCH T., HODKIEWICZ M., STEWART M. & LIU W. (2021). LexiClean : An annotation tool for rapid multi-task lexical normalisation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, p. 212–219, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-demo.25.

BUCUR A.-M., COSMA A. & DINU L. P. (2021). Sequence-to-sequence lexical normalization with multilingual transformers. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, p. 473–482, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.wnut-1.53.

CERÓN-GUZMÁN J. A. & LEÓN-GUZMÁN E. (2016). Lexical normalization of spanish tweets. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, p. 605–610, Montréal, Québec, Canada : International World Wide Web Conferences Steering Committee. DOI : 10.1145/2872518.2890558.

CHOUDHURY M., SARAF R., JAIN V., MUKHERJEE A., SARKAR S. & BASU A. (2007). Investigation and modeling of the structure of texting language. *International Journal of Document Analysis and Recognition (IJDAR)*, **10**(3-4), 157–174. DOI : 10.1007/s10032-007-0054-0.

CLARK E. & ARAKI K. (2011). Text Normalization in Social Media : Progress, Problems and Applications for a Pre-Processing System of Casual English. *Procedia - Social and Behavioral Sciences*, **27**, 2–11. DOI : 10.1016/j.sbspro.2011.10.577.

CLARK K., LUONG M.-T., LE Q. V. & MANNING C. D. (2020). ELECTRA : Pre-training Text Encoders as Discriminators Rather Than Generators. In *Proceedings of the Eighth International Conference on Learning Representations*, Online.

COTELO J., CRUZ F., TROYANO J. & ORTEGA F. (2015). A modular approach for lexical normalization applied to Spanish tweets. *Expert Systems with Applications*, **42**(10), 4743–4754. DOI : 10.1016/j.eswa.2015.02.003.

DAMERAU F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM*, **7**(3), 171–176. DOI : 10.1145/363958.363994.

DE CLERCQ O., SCHULZ S., DESMET B. & HOSTE V. (2014). Towards Shared Datasets for Normalization Research. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, p. 1218–1223, Reykjavik, Iceland : European Language Resources Association (ELRA).

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the*

North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers), p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

EHARA Y. (2021). To What Extent Does Lexical Normalization Help English-as-a-Second Language Learners to Read Noisy English Texts ? In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, p. 451–456, Online : Association for Computational Linguistics.

EISENSTEIN J. (2013). What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 359–369, Atlanta, Georgia : Association for Computational Linguistics.

ERJAVEC T., FIŠER D., ČIBEJ J., ARHAR HOLDT Š., LJUBEŠIĆ N. & ZUPAN K. (2017). CMC training corpus janes-tag 2.0. Slovenian language resource repository CLARIN.SI.

FORMIGA L. & FONOLLOSA J. A. R. (2012). Dealing with input noise in statistical machine translation. In *Proceedings of COLING 2012 : Posters*, p. 319–328, Mumbai, India : The COLING 2012 Organizing Committee.

FOSTER J. (2010). "cba to check the spelling" : Investigating Parser Performance on Discussion Forum Posts. In *Human Language Technologies : The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, p. 381–384, Los Angeles, California : Association for Computational Linguistics.

HAN B. & BALDWIN T. (2011). Lexical Normalisation of Short Text Messages : Makn Sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*, p. 368–378, Portland, Oregon, USA : Association for Computational Linguistics.

HIGASHIYAMA S., UTIYAMA M., WATANABE T. & SUMITA E. (2021). User-generated text corpus for evaluating Japanese morphological analysis and lexical normalization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 5532–5541, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-main.438.

JIANG N., LUO C., LAKSHMAN V., DATTATREYA Y. & XUE Y. (2022). Massive Text Normalization via an Efficient Randomized Algorithm. In *Proceedings of the ACM Web Conference 2022*, p. 2946–2956, Virtual Event, Lyon France : ACM. DOI : 10.1145/3485447.3512015.

KARPUKHIN V., LEVY O., EISENSTEIN J. & GHAZVININEJAD M. (2019). Training on Synthetic Noise Improves Robustness to Natural Noise in Machine Translation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, p. 42–47, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-5506.

KOBUS C., YVON F. & DAMNATI G. (2008). Normalizing SMS : are Two Metaphors Better than One ? In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, p. 441–448, Manchester, UK : Coling 2008 Organizing Committee.

KOGKITSIDOU E. & ANTONIADIS G. (2016). L'architecture d'un modèle hybride pour la normalisation de SMS (a hybrid model architecture for SMS normalization). In *Actes de la conférence conjointe JEP-TALN-RECITAL 2016. volume 2 : TALN (Posters)*, p. 355–363, Paris, France : AFCP - ATALA.

KUBAL D. & NAGVENKAR A. (2021). Multilingual Sequence Labeling Approach to solve Lexical Normalization. In *Proceedings of the Seventh Workshop on Noisy User-generated*

Text (W-NUT 2021), p. 457–464, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.wnut-1.51.

KUMAR A., MAKHIJA P. & GUPTA A. (2020). Noisy Text Data : Achilles' Heel of BERT. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, p. 16–21, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.wnut-1.3.

LEEMAN-MUNK S., LESTER J. & COX J. (2015). NCSU_sas_sam : Deep Encoding and Reconstruction for Normalization of Noisy Text. In *Proceedings of the Workshop on Noisy User-generated Text*, p. 154–161, Beijing, China : Association for Computational Linguistics. DOI : 10.18653/v1/W15-4323.

LEVENSHTEIN V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. In *Soviet Physics Doklady*, volume 10 de *8*, p. 707–710.

LI C. & LIU Y. (2012). Improving text normalization using character-blocks based models and system combination. In *Proceedings of COLING 2012*, p. 1587–1602, Mumbai, India : The COLING 2012 Organizing Committee.

LIU F., WENG F. & JIANG X. (2012). A broad-coverage normalization system for social media language. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1035–1044, Jeju Island, Korea : Association for Computational Linguistics.

LIU Y., GU J., GOYAL N., LI X., EDUNOV S., GHAZVININEJAD M., LEWIS M. & ZETTLEMOYER L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, **8**, 726–742. DOI : 10.1162/tacl_a_00343.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). Roberta : A robustly optimized BERT pretraining approach. *CoRR*, **abs/1907.11692**.

LIU Z., NI S., AW A. T. & CHEN N. F. (2022). Singlish message paraphrasing : A joint task of creole translation and text normalization. In *Proceedings of the 29th International Conference on Computational Linguistics*, p. 3924–3936, Gyeongju, Republic of Korea : International Committee on Computational Linguistics.

LJUBEŠIĆ N., ERJAVEC T., MILIČEVIĆ M. & SAMARDŽIĆ T. (2017a). Croatian twitter training corpus ReLDI-NormTagNER-hr 2.0. Slovenian language resource repository CLARIN.SI.

LJUBEŠIĆ N., ERJAVEC T., MILIČEVIĆ M. & SAMARDŽIĆ T. (2017b). Serbian twitter training corpus ReLDI-NormTagNER-sr 2.0. Slovenian language resource repository CLARIN.SI.

LOURENTZOU I., MANGHNANI K. & ZHAI C. (2019). Adapting Sequence to Sequence models for Text Normalization in Social Media. In *Proceedings of the Thirteenth International AAAI Conference on Web and Social Media (ICWSM 2019)*, p. 335–345, Münich, Germany.

MATOS VELIZ C., DE CLERCQ O. & HOSTE V. (2019). Comparing MT Approaches for Text Normalization. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, p. 740–749, Varna, Bulgaria : INCOMA Ltd. DOI : 10.26615/978-954-452-056-4_086.

MELERO M., COSTA-JUSSÀ M. R., LAMBERT P. & QUIXAL M. (2016). Selection of correction candidates for the normalization of Spanish user-generated content. *Natural Language Engineering*, **22**(1), 135–161. Publisher : Cambridge University Press, DOI : 10.1017/S1351324914000011.

MICHEL P. & NEUBIG G. (2018). MTNT : A testbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 543–553, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/D18-1050.

MOON S., NEVES L. & CARVALHO V. (2018). Multimodal Named Entity Recognition for Short Social Media Posts. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, p. 852–860, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/N18-1078.

MUÑOZ-GARCÍA Ó., NAVARRO C., AVONTUUR T., AZAR Z., BALEMANS I., ELSHOF L., NEWELL R., NOORD N. V., NTAVELOS A., MAYNARD D., BONTCHEVA K., ROUT D., STRASSEL S., ISMAEL S., SONG Z. & LEE H. (2012). Comparing user generated content published in different social media sources.

MULLER B., SAGOT B. & SEDDAH D. (2019). Enhancing BERT for Lexical Normalization. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, p. 297–306, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-5539.

NGUYEN D. Q., VU T. & TUAN NGUYEN A. (2020). BERTweet : A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, p. 9–14, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-demos.2.

NGUYEN V. H., NGUYEN H. T. & SNASEL V. (2016). Text normalization for named entity recognition in Vietnamese tweets. *Computational Social Networks*, **3**(1), 10. DOI : 10.1186/s40649-016-0032-0.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). BLEU : a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, p. 311–318, Philadelphia, Pennsylvania, USA : Association for Computational Linguistics. DOI : 10.3115/1073083.1073135.

PARTANEN N., HÄMÄLÄINEN M. & ALNAJJAR K. (2019). Dialect Text Normalization to Normative Standard Finnish. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, p. 141–146, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-5519.

PENNELL D. & LIU Y. (2011). A character-level machine translation approach for normalization of SMS abbreviations. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, p. 974–982, Chiang Mai, Thailand : Asian Federation of Natural Language Processing.

PLANK B., JENSEN K. N. & VAN DER GOOT R. (2020). DaN+ : Danish nested named entities and lexical normalization. In *Proceedings of the 28th International Conference on Computational Linguistics*, p. 6649–6662, Barcelona, Spain (Online) : International Committee on Computational Linguistics. DOI : 10.18653/v1/2020.coling-main.583.

POST M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation : Research Papers*, p. 186–191, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/W18-6319.

QIN W., LI X., SUN Y., XIONG D., CUI J. & WANG B. (2021). Modeling Homophone Noise for Robust Neural Machine Translation. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, p. 7533–7537, Toronto, ON, Canada. DOI : 10.1109/ICASSP39728.2021.9413586.

REI R., STEWART C., FARINHA A. C. & LAVIE A. (2020). COMET : A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 2685–2702, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.213.

RIABI A., SAGOT B. & SEDDAH D. (2021). Can Character-based Language Models Improve Downstream Task Performances In Low-Resource And Noisy Language Scenarios ? In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, p. 423–436, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.wnut-1.47.

ROSALES NÚÑEZ J. C., SEDDAH D. & WISNIEWSKI G. (2019a). Comparison between NMT and PBSMT Performance for Translating Noisy User-Generated Content. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, p. 2–14, Turku, Finland : Linköping University Electronic Press.

ROSALES NÚÑEZ J. C., SEDDAH D. & WISNIEWSKI G. (2019b). Phonetic Normalization for Machine Translation of User Generated Content. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, p. 407–416, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-5553.

ROSALES NÚÑEZ J. C., SEDDAH D. & WISNIEWSKI G. (2021a). Understanding the Impact of UGC Specificities on Translation Quality. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, p. 189–198, Online : Association for Computational Linguistics.

ROSALES NÚÑEZ J. C., WISNIEWSKI G. & SEDDAH D. (2021b). Noisy UGC Translation at the Character Level : Revisiting Open-Vocabulary Capabilities and Robustness of Char-Based Models. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, p. 199–211, Online : Association for Computational Linguistics.

RUIZ P., CUADROS M. & ETCHEGOYHEN T. (2014). Lexical Normalization of Spanish Tweets with Rule-Based Components and Language Models. *Procesamiento del Lenguaje Natural*, **52**, 45–52.

SANGUINETTI M., BOSCO C., CASSIDY L., ÇETINOĞLU Ö., CIGNARELLA A. T., LYNN T., REHBEIN I., RUPPENHOFER J., SEDDAH D. & ZELDES A. (2020). Treebanking User-Generated Content : A Proposal for a Unified Representation in Universal Dependencies. In *Proceedings of the 12th Language Resources and Evaluation Conference*, p. 5240–5250, Marseille, France : European Language Resources Association.

SARKAR R., MAHINDER S. & KHUDABUKHSH A. (2020). The Non-native Speaker Aspect : Indian English in Social Media. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, p. 61–70, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.wnut-1.9.

SCHERRER Y. & LJUBEŠIĆ N. (2021). Sesame Street to Mount Sinai : BERT-constrained character-level Moses models for multilingual lexical normalization. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, p. 465–472, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.wnut-1.52.

SCHUUR Y. (2020). Normalization for Dutch for improved POS tagging. Mémoire de master, University of Groningen.

SEDDAH D., SAGOT B., CANDITO M., MOUILLERON V. & COMBET V. (2012). The French Social Media Bank : a Treebank of Noisy User Generated Content. In *Proceedings of COLING 2012*, p. 2441–2458, Mumbai, India : The COLING 2012 Organizing Committee.

SIDARENKA U., SCHEFFLER T. & STEDE M. (2013). Rule-Based Normalization of German Twitter Messages. In *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology*, Darmstadt, Germany.

SPROAT R., BLACK A. W., CHEN S., KUMAR S., OSTENDORF M. & RICHARDS C. D. (2001). Normalization of non-standard words. *Computer Speech & Language*, **15**(3), 287–333. DOI : 10.1006/csla.2001.0169.

STEWART M., LIU W. & CARDELL-OLIVER R. (2019). Word-level lexical normalisation using context-dependent embeddings. *CoRR*, **abs/1911.06172**.

SUN Y. & JIANG H. (2019). Contextual Text Denoising with Masked Language Model. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, p. 286–290, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-5537.

SUPRANOVICH D. & PATSEPNIA V. (2015). IHS_rd : Lexical Normalization for English Tweets. In *Proceedings of the Workshop on Noisy User-generated Text*, p. 78–81, Beijing, China : Association for Computational Linguistics. DOI : 10.18653/v1/W15-4311.

TIAN T., TELLIER I., DINARELLI M. & CARDOSO P. (2017). Détection des mots non-standards dans les tweets avec des réseaux de neurones (detecting non-standard words in tweets with neural networks). In *Actes des 24ème Conférence sur le Traitement Automatique des Langues Naturelles. Volume 2 - Articles courts*, p. 174–182, Orléans, France : ATALA.

TIWARI A. S. & NASKAR S. K. (2017). Normalization of social media text using deep neural networks. In *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)*, p. 312–321, Kolkata, India : NLP Association of India.

VAN DER GOOT R. (2019a). An In-depth Analysis of the Effect of Lexical Normalization on the Dependency Parsing of Social Media. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, p. 115–120, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-5515.

VAN DER GOOT R. (2019b). MoNoise : A Multi-lingual and Easy-to-use Lexical Normalization Tool. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, p. 201–206, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-3032.

VAN DER GOOT R. (2019c). *Normalization and parsing algorithms for uncertain input*. Thèse de doctorat, University of Groningen.

VAN DER GOOT R. (2021). CL-MoNoise : Cross-lingual lexical normalization. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, p. 510–514, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.wnut-1.56.

VAN DER GOOT R., PLANK B. & NISSIM M. (2017). To normalize, or not to normalize : The impact of normalization on Part-of-Speech tagging. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, p. 31–39, Copenhagen, Denmark : Association for Computational Linguistics. DOI : 10.18653/v1/W17-4404.

VAN DER GOOT R., RAMPONI A., CASELLI T., CAFAGNA M. & DE MATTEI L. (2020). Norm it ! lexical normalization for Italian and its downstream effects for dependency parsing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 6272–6278, Marseille, France : European Language Resources Association.

VAN DER GOOT R., RAMPONI A., ZUBIAGA A., PLANK B., MULLER B., SAN VICENTE RONCAL I., LJUBEŠIĆ N., ÇETINOĞLU Ö., MAHENDRA R., ÇOLAKOĞLU T., BALDWIN T., CASELLI T. & SIDORENKO W. (2021). MultiLexNorm : A Shared Task on Multilingual Lexical Normalization. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, p. 493–509, Online : Association for Computational Linguistics.

VAN DER GOOT R. & VAN NOORD G. (2017). MoNoise : Modeling Noise Using a Modular Normalization System. *Computational Linguistics in the Netherlands Journal*, **7**, 129–144.

VAN DER GOOT R. & VAN NOORD G. (2018). Modeling Input Uncertainty in Neural Network Dependency Parsing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 4984–4991, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/D18-1542.

VAN DER GOOT R., VAN NOORD R. & VAN NOORD G. (2018). A taxonomy for in-depth evaluation of normalization for user generated content. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, p. 684–688, Miyazaki, Japan : European Language Resources Association (ELRA).

VAN DER GOOT R. & ÇETINOĞLU Ö. (2021). Lexical Normalization for Code-switched Data and its Effect on POS Tagging. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : Main Volume*, p. 2352–2365, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.eacl-main.200.

VITERBI A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, **13**(2), 260–269. DOI : 10.1109/TIT.1967.1054010.

XU K., XIA Y. & LEE C.-H. (2015). Tweet Normalization with Syllables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 920–928, Beijing, China : Association for Computational Linguistics. DOI : 10.3115/v1/P15-1089.

YANG F., BAGHERI GARAKANI A., TENG Y., GAO Y., LIU J., DENG J. & SUN Y. (2022). Spelling Correction using Phonetics in E-commerce Search. In *Proceedings of The Fifth Workshop on e-Commerce and NLP (ECNLP 5)*, p. 63–67, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.ecnlp-1.9.

YUAN Z., TASLIMIPOOR S., DAVIS C. & BRYANT C. (2021). Multi-Class Grammatical Error Detection for Correction : A Tale of Two Systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 8722–8736, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.687.

ÇOLAKOĞLU T., SULUBACAK U. & TANTUĞ A. C. (2019). Normalizing Non-canonical Turkish Texts Using Machine Translation Approaches. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics : Student Research Workshop*, p. 267–272, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-2037.