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

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.45̃ : 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.45̃ : 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.