Derrière les plongements de relations

Hugo Thomas[1] Guillaume Gravier[1] Pascale Sébillot[2]
(1) Univ Rennes, CNRS, Inria - IRISA, Campus de Beaulieu 35042 Rennes, France
(2) Univ Rennes, CNRS, Inria, INSA Rennes - IRISA, Campus de Beaulieu 35042 Rennes, France
hugo.thomas@irisa.fr, guillaume.gravier@irisa.fr,
pascale.sebillot@irisa.fr

RÉSUMÉ

Dans cet article, plutôt que nous arrêter aux scores de performance habituellement fournis (par ex. mesure F1), nous proposons une analyse approfondie, selon différents critères, des modélisations de relations employées par plusieurs architectures de modèles de typage de relations. Cette analyse vise à mieux comprendre l'organisation de l'espace latent des modélisations et ses propriétés, enjeu important pour les modèles se fondant sur les distances dans cet espace. Dans cet objectif d'analyse des plongements, nous étudions l'influence, sur ces modélisations, du vocabulaire, de la syntaxe, de la sémantique des relations, de la représentation des entités nommées liées, ainsi que la géométrie de leur espace latent. Il en ressort que les modélisations de relations sont apprises de manière inégale d'un modèle à un autre entraînés de la même manière ; dans ce cas, les indicateurs que nous proposons sont de nouveaux éléments de compréhension de l'espace latent d'un modèle afin de mieux exploiter ses propriétés.

ABSTRACT

What hides behind relation embeddings ?

In this paper, rather than focusing on the performance scores usually provided (e.g., the F1 measure), we propose a more in-depth analysis, according to several criteria, of the relation modeling of different model architectures for relation typing. This analysis aims to better understand the organization and properties of the latent modeling space, an important issue for models exploiting distances in this vector space. In order to study these modelings, we evaluate the influence on these models of the lexicon, the syntax, and the semantics of relations, the representation of the entities, as well as the geometry of their latent spaces. It appears that the relation modelings are learned unevenly from one model to another trained in the same way ; in this case, the indicators we proposed are additional knowledge about the latent space to better exploit its properties.

MOTS-CLÉS : Traitement automatique des langues, extraction de relations, classification de relations, modélisation de relations.

KEYWORDS: Natural language processing, relation extraction, relation classification, relation modeling.

1 Introduction

L'extraction de relations entre entités – en particulier entre entités nommées (noms de personnes, lieux, entreprises...) – est un champ de recherche très actif du traitement automatique des langues (TAL)

(*cf.* par exemple (Nasar *et al.*, 2022) pour un résumé récent de nombreux travaux et des techniques utilisées). Pouvoir acquérir automatiquement et exploiter les relations décrites au sein de textes est un enjeu important, en particulier pour peupler des bases de connaissance et pour comprendre les interactions entre entités d'intérêt, et ce, dans de nombreux domaines (médecine (Karaa *et al.*, 2021), finance (Jabbari *et al.*, 2020), justice (Hong *et al.*, 2021), journalisme (Riedel *et al.*, 2010), publications scientifiques (Bhattacharya & Getoor, 2007)...).

Les relations considérées sont fréquemment binaires, exprimées au sein d'une phrase – ce sera aussi le focus de cet article – et sont modélisées par un triplet constitué d'une **entité tête** (ou sujet), d'une **entité queue** (ou objet) et d'une **relation** ou prédicat qui les lie. Par exemple, la phrase *Joe Biden est président des États-Unis depuis 2021* exprime le fait que l'entité *Joe Biden* et l'entité *États-Unis* sont liées par une relation modélisable par le triplet *(Joe Biden, est_président_de, États-Unis)*.

La littérature actuelle du TAL concernant l'extraction de relations se focalise sur deux tâches de classification principales : la première est la *détection* de relations qui consiste à déterminer si deux entités dans un texte (ou une phrase) donné sont ou non liées par une relation quelconque ; la seconde est le *typage* de relations, reposant sur la détermination du type de la relation entre deux entités supposées liées par une relation dans un texte donné.

Pour la suite de l'article, notre attention se porte sur la tâche de typage de relations : comme expliqué ensuite, nous nous focalisons sur les modélisations des relations, et nous supposerons donc que les relations dans le texte ont déjà été détectées afin de pouvoir directement les étudier en étant certains de leur présence. Cette tâche de typage se traduit alors dans le cas général par la résolution de

$$\arg\max_{r} \; p(r \mid s, e_{tête}, e_{queue}) \tag{1}$$

où s est la phrase étudiée, r est un type de relation, $\mathcal{R}$ est l'ensemble des types de relations, $(e_{tête}, e_{queue})$ sont les entités et $\mathcal{E}$ est l'ensemble des entités ($r \in \mathcal{R}$ et $(e_{tête}, e_{queue}) \in \mathcal{E}^2$).

Une étude extensive des méthodes de détection et de typage de relations (Nasar *et al.*, 2022) dégage les grandes familles de modèles et architectures adaptées à ces tâches. Elle n'aborde toutefois pas les modèles Transformers qui désormais font l'état de l'art dans un grand nombre de tâches de TAL, probablement par manque de recul sur ces modèles encore récents. La comparaison des modèles se limite à celle de leurs performances évaluées à l'aide de métriques classiques telles que la précision, le rappel et la mesure F1, qui ne suffisent toutefois pas à comprendre la nature des différences entre les modélisations. Du côté des modèles Transformers, des avancées notables sont faites en termes de scores de classification, mais encore une fois, sans qu'on ne dispose d'une analyse fine de la modélisation des relations sous-jacente. C'est par exemple le cas de modèles tels que LUKE (Yamada *et al.*, 2020) ou le modèle proposé par (Zhou & Chen, 2022), qui figurent parmi l'état de l'art du typage de relations et proposent une modélisation novatrice des entités sans approfondir la modélisation résultante des relations. Enfin, les modèles récents d'extraction de relations non supervisée ou sans exemple (*zero shot*) s'appuient sur des encodeurs Transformers et notamment sur la mesure de similarité dans leur espace latent (par exemple (Chen & Li, 2021) ou plusieurs modèles de l'état de l'art décrit par Simon (2022)). Il est dans ce cas primordial de comprendre les attributs et informations de la relation reflétés par leur représentation, afin de saisir sur quoi s'appuie un modèle pour définir la similarité entre deux relations. Sans cela, il est impossible de comprendre les erreurs du modèle, ou au contraire ce sur quoi repose son efficacité.

Il apparaît donc qu'une meilleure appréhension des modélisations des relations apprises par les modèles de typage de relations manque dans ce domaine du TAL. S'arrêter à des scores de classification

est une manière objective et efficace de comparer la performance de modèles, mais, dans une optique académique, est insuffisant pour comprendre la nature des différences de représentations entre ces classifieurs. Il est donc nécessaire d'enrichir la comparaison des modèles avec des indicateurs plus fins. Par exemple, bien que des modèles aient la capacité de tirer parti de la syntaxe des relations, la littérature récente ne vérifie pas l'influence de celle-ci sur la modélisation résultante. L'impact de la représentation des entités liées n'est également pas éprouvé au-delà des scores de classification. L'espace latent des modélisations de relations de chaque modèle est le siège de l'information sur ces relations, mais il n'est pas sondé au-delà des visualisations de ses projections en deux dimensions par t-SNE.

Pour améliorer les comparaisons des modèles, l'initiative de Alt *et al.* (2020) d'étude critique du jeu de données TACRED procède à une analyse approfondie des erreurs de quatre modèles (taux d'erreurs sur les longues relations, les relations avec le même type d'entités tête et queue, les relations aux entités éloignées...). Nous nous positionnons dans la lignée de cette démarche, mais souhaitons aller plus loin et comparer des modèles représentatifs d'architectures plus hétérogènes sur davantage de critères afin d'aboutir à une analyse plus fine.

Dans un premier temps, nous présentons les jeux de données utilisés et les modèles comparés, afin d'exposer leurs différences pour contextualiser les résultats de nos expériences. Nous examinons dans un deuxième temps plusieurs critères : l'influence de la prise en compte de la syntaxe sur le typage des relations ; la représentation des entités et la façon dont elle affecte les représentations de relations ; l'intensité du lien entre la similarité des plongements de deux phrases supports de relations et la similarité lexicale, syntaxique et sémantique de ces deux phrases, ainsi que les distances et dispersions inter-relations dans l'espace latent de représentation des relations. Nous explorons ces critères pour déceler des différences de plusieurs natures entre les modélisations étudiées sur les jeux de données choisis. Nous concluons avec un résumé des résultats des expériences et leurs implications grâce au recul apporté.

2 Contexte et cadre expérimental

Notre analyse porte sur une variété de modèles de typage et extraction de relations représentatifs de différentes avancées du domaine. L'objectif étant de proposer des éléments de comparaison fiables et fins entre ces modèles, nous choisissons d'homogénéiser l'apprentissage de ceux-ci. Ainsi, tous les classifieurs ont pour point commun qu'une relation est modélisée par un vecteur de dimension fixe $\mathcal{D}$. Le deuxième point commun est l'architecture de la tête de classification utilisée pour déduire de chacune de ces modélisations une prédiction de type de relation par l'équation 1. Cette tête de classification effectue une prédiction sous la forme du vecteur $x_{pred} \in \mathbb{R}^n$ (n est le nombre de types de relation égal à $card(\mathcal{R})$) pour la phrase support d'une relation modélisée par x_{rel}, tel que $x_{pred} = W_2 ReLU(W_1 x_{rel} + b_1) + b_2$. W_1 et W_2 sont les matrices des poids des deux couches de neurones, et b_1 et b_2 les biais de ces mêmes neurones. $x_{rel} \in \mathbb{R}^{\mathcal{D}}$, $W_1 \in \mathbb{R}^{\mathcal{D}\times\mathcal{D}}$, $W_2 \in \mathbb{R}^{\mathcal{D}\times n}$, $b_1 \in \mathbb{R}^{\mathcal{D}}$, $b_2 \in \mathbb{R}^n$, et nous choisissons $\mathcal{D} = 768$.

L'hypothèse de domaine clos est admise, c'est-à-dire que la tâche de typage de relations est restreinte à identifier une relation parmi $card(\mathcal{R})$ types et aucun autre type de relation n'est supposé possible. Ce contexte permet de se focaliser sur des relations précises et nommées afin de faciliter l'étude de celles-ci.

Une première modélisation que nous examinons s'appuie sur la similarité lexicale (de l'expression) des relations, des (portions de) phrases supports d'une même relation utilisant supposément un vocabulaire voisin. Ce modèle n'est pas neuronal, mais on applique sur sa modélisation des couches de neurones de prédiction du type de la relation, comme décrit au premier paragraphe. Une phrase est donc traitée comme un vecteur sac-de-mots ayant une composante par mot du dictionnaire, pondéré par TF-IDF (*term frequency-inverse document frequency*). Nous concaténons à la fin de ce vecteur l'encodage *one hot* du type des deux entités de la relation. Ce vecteur concaténé est ramené par analyse en composantes principales à la dimension désirée $\mathcal{D}$ pour obtenir la modélisation finale. On remarquera que la notion de séquence est perdue dans ce modèle, ce qui rend, pour lui, impossible l'exploitation de la syntaxe des phrases.

Un deuxième modèle, popularisé par le domaine de la vision par ordinateur, est le réseau de neurones convolutif (Kim, 2014) : différentes tailles de noyaux de convolution sont appliquées sur les plongements des *tokens* en entrée pour extraire de l'information à plusieurs échelles. La valeur maximale en sortie de chaque strate de traitement est extraite pour obtenir un vecteur par taille de noyau de convolution. Ces vecteurs sont ensuite concaténés, puis leur dimension est réduite par des couches de neurones artificiels. En sortie, la modélisation des relations pour ce modèle est obtenue. Cette méthode repose en grande partie sur des schémas répétés dans les phrases d'entraînement et peut donc souffrir en cas de grande variabilité des formes de surface exprimant un même type de relation.

La troisième modélisation concerne les modèles de réseaux de neurones récurrents, qui traitent de façon séquentielle l'information. On leur reproche de noyer l'information en début de séquence, ainsi que de souffrir de l'évanescence du gradient, mais ce sont des problèmes que les LSTM bidirectionnels ou Bi-LSTM atténuent. Ainsi, nous considérons dans notre étude la concaténation des vecteurs cachés des deux LSTM (avant et arrière) comme modélisation de relation.

Les Transformers, encodeurs-décodeurs tirant parti des mécanismes d'attention, offrent depuis 2017 (Vaswani *et al.*, 2017) des performances élevées sur les données textuelles dans des tâches requérant une compréhension fine du langage, par exemple la traduction ou l'analyse des sentiments. On étudie l'état du vecteur associé au token [CLS] en sortie du modèle comme représentation du texte en entrée. Nous examinons les modèles RoBERTa (Liu *et al.*, 2019), ainsi que LUKE (Yamada *et al.*, 2020) qui possède des plongements et des mécanismes d'attention spécifiques aux entités pour améliorer la classification de relations. Pour nos entraînements, les Transformers ne sont ni appris à partir de poids aléatoires, ni ajustés entièrement à partir de modèles pré-entraînés : nous adaptons les modèles en entraînant des couches de réduction et augmentation de dimensionnalité dans chaque bloc Transformer comme décrit dans (Houlsby *et al.*, 2019). Cette méthode permet l'apprentissage d'une fraction des paramètres originaux du modèle pour des performances similaires. Le modèle RoBERTa base est adapté pour typer les relations, et le modèle RoBERTa+entités est adapté pour typer les relations et les entités tête et queue. Nous considérons LUKE uniquement comme référence pour ses scores, mais l'analyse de sa modélisation n'est pas effectuée, n'étant pas compatible à l'adaptation et nos tentatives de *fine-tuning* ayant échoué.

Enfin, nous étudions la modélisation des relations d'un Transformer appris de manière supervisée pour une tâche proxy, c'est-à-dire la prédiction du type des entités de la relation (entreprise, personne, lieu...), dont les poids sont ensuite gelés pour apprendre la tête de typage de relation : ce modèle est analogue à celui de Huang & Wong (2020), avec pour encodeur un Transformer et non des LSTM. Puisque les types d'entités déterminent en partie le type de la relation, il est réaliste d'estimer que cette tâche proxy de prédiction d'entités apprend des attributs pertinents pour la classification de relations.

Les fonctions utilisées pour entraîner ces différents modèles sont disponibles sur le dépôt de code de l'article [1].

Les réseaux de neurones sont entraînés par descente de gradient optimisée par Adam avec un taux d'apprentissage de 0.001 par *batch* de 8 relations pendant 128 époques avec interruption précoce (conditionnée sur la mesure F1 du modèle sur le jeu de validation). Chaque modèle est entraîné cinq fois avec une graine aléatoire différente. Les plongements de mots de chaque modèle (sauf TF-IDF+ACP qui n'en nécessite pas) sont ceux du modèle préentraîné RoBERTa base.

Afin d'effectuer selon plusieurs critères l'analyse des modélisations sous-jacentes de modèles, nous employons deux jeux de données conçus pour la tâche de typage de relation, dont les caractéristiques sont déclinées dans le tableau 1.

TACRED (Zhang *et al.*, 2017) (*TAC Relation Extraction Dataset*) présente 41 types de relations dans 106 264 exemples en anglais. 79,5% de ces relations sont du type « no_relation » (pas de relation entre les deux entités courantes dans une phrase), et les fréquences des types de relations sont peu équilibrées, présentant ainsi des classes très majoritaires et minoritaires (cadre de classification frugale). Nous ignorons la classe « no_relation » car nous étudions la tâche de typage de relations et pas de détection de relations.

FewRel (Han *et al.*, 2018), également en langue anglaise, contient 100 types de relations dans 56 000 exemples, et sa spécificité est son évaluation : il est conçu pour être évalué sur la classification frugale, avec peu d'exemples. Le script d'évaluation échantillonne le jeu de validation selon le schéma *N-way K-shot*, c'est-à-dire la classification parmi N classes avec seulement K exemples par classe. Souhaitant nous focaliser sur la modélisation des relations des modèles à leur plein potentiel, l'aspect de classification frugale n'est pas pris en compte dans cet article ; ainsi, nous découpons le jeu d'entraînement en deux parties (jeu d'entraînement raccourci et jeu de test) afin d'obtenir un jeu de test dédié en plus des jeux d'entraînement et de validation.

jeu de données	jeu d'entraînement	jeu de validation	jeu de test	total
TACRED	13 012	5436	3325	21 773
FewRel 1.0	33 600	11 200	11 200	56 000

TABLE 1 – Résumé du contenu des jeux de données.

3 Expériences

Nous procédons par étape aux différentes analyses des modélisations de relations en commençant par étudier l'impact de la syntaxe et de sa prise en compte sur les modélisations obtenues (et donc sur la qualité des représentations pour la tâche de typage de relations). Nous nous penchons ensuite sur l'influence de la représentation choisie des entités de la relation sur ces modélisations. Enfin, pour aller au-delà d'une analyse s'étayant à l'aune de la seule qualité plus ou moins forte en classification, nous proposons une étude comparative des modélisations de relations obtenues d'un point de vue langue (deux relations proches en termes de représentations vectorielles ont-elles des « propriétés langagières » proches ?) et d'un point de vue géométrie des espaces vectoriels appris.

1. Dépôt disponible sur ce lien : https://gitlab.inria.fr/huthomas/taln_experiments

3.1 Impact de la syntaxe

L'identification des relations peut reposer en partie sur la syntaxe des phrases (ou parties de phrases) supports de celles-ci. Étant donné que le typage de la relation dépend de la modélisation, nous souhaitons vérifier l'importance de la syntaxe sur cette modélisation de relation ; nous procédons à une comparaison des différents modèles en conservant d'une part la phrase entière en entrée, d'autre part uniquement les mots du plus court chemin de dépendance syntaxique entre les deux entités de la relation, dans l'ordre de ce chemin, de l'entité tête à l'entité queue.

modèle	partie de phrase considérée	F1 sur TACRED		F1 sur FewRel	
		moyenne micro	moyenne macro	moyenne micro	moyenne macro
TF-IDF+ACP	plus court chemin	**83.3±0.62%**	**61.19±0.75%**	62.32±0.51%	61.92±0.61%
	phrase entière	**84.14±0.64%**	**60.83±1.31%**	61.79±0.59%	61.16±0.63%
CNN	plus court chemin	74.07±0.73%	53.85±0.91%	**71.04±0.22%**	**70.98±0.23%**
	phrase entière	42.1±0.93%	27.39±0.64%	54.24±0.41%	52.91±0.67%
Bi-LSTM	plus court chemin	72.67±1.13%	53.71±1.79%	67.42±2.05%	66.64±2.41%
	phrase entière	37.91±0.68%	23.48±1.32%	49.74±1.07%	47.34±1.48%
RoBERTa base	plus court chemin	77.89±0.37%	58.96±1.53%	**71.43±0.37%**	**70.55±0.9%**
	phrase entière	43.17±0.54%	28.66±0.91%	58.43±0.95%	56.06±1.15%
RoBERTa+entités	plus court chemin	76.05±1.13%	56.36±0.98%	**73.75±0.81%**	**73.14±0.91%**
	phrase entière	42.17±1.09%	29.06±0.77%	56.14±0.54%	54±0.63%
RoBERTa proxy	plus court chemin	72.49±1.64%	50.73±2.2%	55±0.96%	53.54±1.15%
	phrase entière	37.38±1.69%	23.63±1.28%	48.97±0.47%	47.16±0.48%

TABLE 2 – Comparaison des résultats de typage de relations en considérant la phrase entière ou seulement le plus court chemin de dépendance syntaxique entre les deux entités de la phrase. Les meilleurs scores au risque de 5% sont indiqués en gras.

Les résultats sont présentés dans le tableau 2. En étudiant d'abord le modèle TF-IDF+ACP sur les premières lignes, nous observons une faible différence entre les deux configurations sur les deux jeux de données : ce modèle ne prend pas en compte la syntaxe, et cette différence est seulement due au nombre de mots inférieur dans le cas du plus court chemin de dépendance. Les scores de ce modèle sont les meilleurs sur TACRED – signifiant la forte dépendance des types de relations au vocabulaire dans ce jeu de données – et sont bons sur FewRel. Le vocabulaire influe donc en partie sur le typage de relations, mais la syntaxe est nécessaire pour de meilleurs scores, notamment sur FewRel. En se penchant sur les autres lignes du tableau, nous constatons que les autres modèles prenant la syntaxe en compte sont systématiquement meilleurs lorsque les entrées sont les plus courts chemins de dépendance syntaxique : cette normalisation grammaticale des relations est bénéfique, démontrant l'importance de la prise en compte de la syntaxe pour la modélisation des relations dans l'objectif de leur typage.

3.2 Influence des représentations des entités

Nous examinons ici l'impact de la représentation des entités sur la modélisation résultante des relations. En effet, les entités sont les sujets et objets mêmes des relations étudiées, et y jouent donc très vraisemblablement un rôle-clé. Les entités sont, dans un premier temps, conservées telles qu'elles apparaissent dans la forme de surface où s'exprime la relation. Ensuite, elles sont entourées de balises afin de signifier au modèle où elles commencent et finissent dans la phrase. Enfin, les entités sont remplacées par leur type (lieu, personne, entreprise...) pour vérifier s'il n'est pas suffisant pour prédire la relation. Le modèle TF-IDF+ACP est exclu, car ajouter les balises d'entités dans son vocabulaire

ne changera rien puisqu'elles apparaissent identiquement dans chaque phrase ; de plus, le modèle bénéficie déjà du type des entités parmi ses attributs.

modèle	représentation des entités	F1 sur TACRED		F1 sur FewRel	
		moyenne micro	moyenne macro	moyenne micro	moyenne macro
CNN	entité non modifiée	42.1±0.93%	27.39±0.64%	54.24±0.41%	52.91±0.67%
	type de l'entité	79.19±0.34%	61.78±0.92%	56.67±0.35%	55.9±0.32%
	entité balisée	69.71±0.7%	52.67±1.13%	66.37±0.35%	65.55±0.34%
Bi-LSTM	entité non modifiée	37.91±0.68%	23.48±1.32%	49.74±1.07%	47.34±1.48%
	type de l'entité	**80.87±0.35%**	**64.39±0.7%**	53.7±1.81%	52.52±1.61%
	entité balisée	69.95±0.94%	54.48±1.29%	63.34±0.35%	62.17±0.62%
RoBERTa base	entité non modifiée	43.17±0.54%	28.66±0.91%	58.43±0.95%	56.06±1.15%
	type de l'entité	**80.86±0.27%**	**65.62±1.01%**	63.35±0.5%	61.89±0.78%
	entité balisée	77.43±0.91%	61.15±1.34%	**73.79±1.24%**	**72.65±1.47%**
RoBERTa+entités	entité non modifiée	42.17±1.09%	29.06±0.77%	56.14±0.54%	54±0.63%
	type de l'entité	**79.75±0.36%**	**64.39±1.79%**	64.47±0.78%	63.04±1.12%
	entité balisée	74.45±0.59%	57.89±2.88%	**75.16±0.13%**	**74.09±0.12%**
RoBERTa proxy	entité non modifiée	33.12±8.29%	19.47±8.52%	48.97±0.47%	47.16±0.48%
	type de l'entité	71.96±1.87%	46.28±1.71%	41.67±1.51%	40.03±1.64%
	entité balisée	69.08±0.89%	51.05±2.25%	59.28±0.88%	58.03±0.61%

TABLE 3 – Comparaison des résultats en fonction de la représentation des entités dans la phrase entière. Les meilleurs scores au risque de 5% sont indiqués en gras.

Le tableau 3 résume les scores obtenus. Sur la colonne de TACRED, les lignes liées aux entités non modifiées conduisent systématiquement à des scores plus faibles que les autres configurations : les modèles n'ont pas de repères pour la position des entités et peuvent confondre celles-ci avec d'autres entités de la phrase. Baliser les entités améliore les scores en levant cette confusion. Les meilleurs scores sont obtenus en remplaçant les entités par leur type, qui apporte une information riche et condensée sur elles, au prix de la perte de leurs mentions exactes dans la phrase.

Sur FewRel, le constat reste le même pour les entités non modifiées menant à des scores faibles. Remplacer les entités par leur type améliore encore une fois les scores sauf pour RoBERTa proxy, qui focalise vraisemblablement trop son apprentissage sur le typage des entités et dégrade ses attributs pour le typage de relations. Les balises autour des entités conduisent aux meilleurs scores sur ce jeu de données, conservant la mention d'entité intacte et indiquant sa position.

D'autres représentations des entités existent et peuvent améliorer encore les scores ; par exemple, LUKE obtient sur TACRED une mesure F1 micro de 88.91% et macro de 59.82%, mais n'entre pas dans notre comparaison à cause de la différence de largeur du modèle LUKE préentrainé sur TACRED (dimension supérieure des plongements). Ces scores vont tout de même dans le sens de l'existence d'un impact des représentations des entités sur le typage de relations. Il découle de ce constat et des expériences que les entités des relations et leur représentation ont une influence importante pour le typage de celles-ci, impactant directement le score des modèles prenant en compte ces entités.

3.3 Analyse des plongements des relations

Les deux expériences précédentes se penchant uniquement sur les scores de classification (mesure F1) pour tirer des conclusions, nous complétons notre analyse avec des métriques décrivant plus finement les différences entre modélisations de relations. Nous choisissons de nous attacher tout d'abord à des mesures directement liées au langage, avant d'effectuer des mesures géométriques.

Dans un premier temps, nous voulons étudier, pour chaque modèle, à quel point la similarité entre deux plongements de relations est proche de la similarité lexicale, syntaxique ou sémantique des formes de surface exprimant ces relations. Ceci permet de nous renseigner sur le fait qu'un modèle s'appuie plus ou moins sur ces trois types d'information. Pour ce faire, nous calculons la corrélation de Spearman entre la similarité cosinus des deux plongements de relations et chacune des trois autres similarités. La similarité lexicale est calculée par un Jaccard (taille de l'intersection des mots divisée par la taille de l'union des mots) entre les (portions de) phrases exprimant les relations. La similarité syntaxique est estimée en calculant la distance entre les arbres syntaxiques des deux phrases selon le noyau décrit dans (Culotta & Sorensen, 2004). La similarité sémantique est obtenue par similarité cosinus entre les moyennes des vecteurs de sortie d'un modèle SentenceBERT (Reimers & Gurevych, 2019) (moyenne des vecteurs de sortie de l'encodeur de chaque token) de détection de paraphrases. Nous échantillonnons 3 000 paires aléatoires de phrases toutes relations confondues dans chacun des jeux de données, mesurons chaque type de similarité sur ces paires, puis calculons la corrélation de Spearman entre la similarité des plongements de relations et chacune des trois autres similarités. Pour les corrélations données au tableau 4, la forme de surface considérée pour tous les modèles est la phrase entière sans modification des entités afin de garantir des similarités fiables ; les mêmes calculs ont été effectués avec les meilleures configurations par modèle, mais donnant des résultats comparables, ne sont pas répertoriés ici par manque de place.

Dans un second temps, nous souhaitons examiner l'organisation géométrique des espaces latents de représentations des relations des différents modèles. En effet, les différents encodeurs étudiés plongent des phrases support de relation dans un espace vectoriel de dimension élevée (768), il est donc important de s'intéresser à l'organisation spatiale de ces vecteurs de relations au sein de l'espace vectoriel pour différencier ces encodeurs. Après avoir constitué des *clusters* de plongements de relations par type, nous calculons les distances entre leurs centroïdes (vecteur moyen de tous les plongements des phrases exprimant une même relation) , afin d'observer la densité relative de chaque espace vectoriel : à cette fin, nous calculons la distance moyenne de chaque centroïde à ses cinq plus proches centroïdes voisins, et établissons la moyenne de ces distances. Nous cherchons, d'autre part, à mesurer le recouvrement des *clusters*, qui traduit la confusion d'un encodeur entre plusieurs types de relations. Pour cela, pour chaque plongement de relation, ses cinq plus proches voisins sont considérés, et la moyenne, pour tous les plongements, de leurs plus proches voisins (parmi les cinq) n'étant pas du même type de relation qu'eux est calculée. À nouveau, dans le tableau 5 recensant les distances et les recouvrements, les plongements manipulés concernent les phrases entières exprimant les relations (même remarque que précédemment pour ce qui est des expériences faites avec les meilleures configurations des modèles).

Malgré les valeurs assez faibles du tableau 4, explicables par la métrique (corrélation de Spearman) qui détecte les tendances monotones mais décroît rapidement avec la dispersion des valeurs, les différences marquées entre les valeurs de corrélation permettent cependant de tirer des conclusions. Sur la colonne de TACRED, la corrélation avec la similarité sémantique est assez élevée, surtout pour les modèles CNN, RoBERTa base et RoBERTa+entités : leur modélisation s'appuie en partie sur la sémantique des relations pour les représenter. Ces trois modèles ont également des corrélations marquées avec la similarité syntaxique, alors que TF-IDF+ACP a une corrélation extrêmement faible due à son incapacité de prise en compte de la syntaxe. Quant à la similarité lexicale, les modèles RoBERTa base et RoBERTA+entités se démarquent avec des corrélations plus fortes : ces deux modèles prennent le mieux en compte les trois aspects langagiers étudiés. La colonne correspondant à FewRel donne des résultats différents. La similarité des plongements du modèle CNN est la plus fortement corrélée à la similarité sémantique, suivi des autres modèles. Seul TF-IDF+ACP

Modèles	TACRED corrélation (*100) avec similarité			FewRel corrélation (*100) avec similarité		
	lexicale	syntaxique	sémantique	lexicale	syntaxique	sémantique
TF-IDF+ACP	12.58	2.18	21.15	6.76	4.84	7.02
CNN	14.23	25.67	29.13	14.87	19.64	30.68
Bi-LSTM	16.07	11.75	22.37	6.26	10.21	23.49
RoBERTa base	25.76	34.01	31.39	7.40	14.26	18.95
RoBERTa+entités	20.86	31.42	33.43	8.43	11.42	21.30
RoBERTa proxy	16.78	15.61	24.70	5.52	12.81	18.75

TABLE 4 – Corrélation de la similarité cosinus des modélisations de relations avec différentes natures de similarité (lexicale, syntaxique et sémantique). Les corrélations sont multipliées par 100 pour améliorer la lisibilité.

modèles	TACRED		FewRel	
	distance inter-classes des centroïdes de relations	recouvrement des *clusters* de types de relations	distance inter-classes des centroïdes de relations	recouvrement des *clusters* de types de relations
TF-IDF+ACP	0.6916±0.5255	17.84%	0.5447±0.3361	64.12%
CNN	6.9164±3.8321	59.54%	7.6531±4.6236	60.97%
Bi-LSTM	0.5577±0.3038	62.92%	0.8039±0.4639	54.94%
RoBERTa base	0.7980±0.4573	57.71%	1.3204±0.7979	50.26%
RoBERTa+entités	1.1839±0.6462	56.00%	0.7464±0.4305	50.17%
RoBERTa proxy	0.5315±0.3745	65.39%	0.6276±0.3581	63.03%

TABLE 5 – Comparaison de la géométrie des espaces latents des différents modèles.

possède une très faible valeur de corrélation, qui était pourtant élevée sur TACRED ; cette chute peut être due à la complexité relative de FewRel, possédant notamment 100 types de relations contre 41 dans TACRED. La corrélation syntaxique pour le modèle CNN est la plus élevée, explicable par le fonctionnement de celui-ci, s'attachant aux motifs dans la phrase et donc possiblement à la construction grammaticale. Cette corrélation reste faible pour TF-IDF+ACP pour la même raison que sur TACRED. La corrélation avec la similarité lexicale est faible sur ce jeu de données, mais le modèle CNN se démarque encore une fois.

De manière générale sur les deux jeux de données (scores les plus élevés, restant consistants sur TACRED et FewRel), les meilleurs modèles dans leur meilleure configuration selon les tableaux 2 et 3 – soit RoBERTa base et RoBERTa+entités, et CNN dans une moindre mesure – sont parmi les plus corrélés avec les similarités sémantique, syntaxique et lexicale dans ce tableau. Ce résultat suggère des capacités langagières élevées de ces modèles, en particulier les Transformers, et leur faculté à exploiter ces informations, justifiant en partie leurs bonnes performances. Ces corrélations restent toutefois faibles dans l'absolu, n'expliquant donc pas tout, ce qui signifie que les modèles ont probablement appris une autre forme de similarité plus pertinente pour la tâche, que l'on peut nommer similarité de typage de relation.

Ce constat est renforcé par les résultats du tableau 5, révélant des recouvrements des *clusters* relativement plus faibles pour ces modèles aux scores de typage de relations élevés et aux corrélations précédentes les plus fortes. Sur TACRED, le modèle CNN possède la plus grande distance moyenne inter-classes, suivi de loin par RoBERTa+entités puis RoBERTa base, et ces modèles ont un recouvrement des *clusters* conséquent malgré leur éloignement suggéré par la métrique précédente. Les modèles Bi-LSTM et RoBERTa proxy souffrent du plus fort recouvrement de leurs *clusters* et de distances inter-classes faibles, les rendant mal adaptés à l'exploitation des distances dans leur espace

latent. À l'opposé, les types de relations modélisés par TF-IDF+ACP sont bien séparés comme le suggèrent le très faible recouvrement de leurs *clusters*. Dans la colonne de FewRel, nous constatons que les *clusters* modélisés par TF-IDF+ACP et RoBERTa proxy se recouvrent fortement, ces modèles étant donc de mauvais candidats à la recherche de relations similaires dans leur espace latent comme expliqué précédemment. Les modèles RoBERTa avec et sans typage d'entités sont au contraire de bons candidats avec les plus faibles recouvrements.

4 Conclusion

Grâce aux expériences effectuées, nous observons que la syntaxe des relations et la représentation des entités ont une influence importante sur le typage des relations : avoir un modèle capable de prendre en compte la syntaxe et adopter une représentation des entités pertinente pour la tâche visée sont donc indispensables pour obtenir la modélisation de relations la plus propice à des meilleurs scores de classification.

La comparaison des espaces latents des différentes architectures de modèles dévoile des modélisations de relations sous-jacentes inégales à plusieurs égards : la corrélation de la similarité des plongements de relations dans ces modélisations avec la similarité lexicale, syntaxique ou sémantique varie significativement d'un modèle à l'autre, révélant leurs affinités respectives avec chacun de ces trois aspects langagiers. La répartition des *clusters* de types de relations nous renseigne également sur ces espaces latents : nous constatons pour certains modèles un recouvrement important des *clusters* et des distances inter-*clusters* faibles. Dans le cas où les distances dans l'espace latent sont à exploiter – par exemple pour la classification frugale à l'aide d'un modèle préentraîné –, il est utile de prendre en compte ces affinités lexicale, syntaxique et sémantique, ainsi que les informations sur la distribution des *clusters* afin d'obtenir les natures de similarité voulues. Il peut, par exemple, être souhaitable d'éviter une similarité des plongements de relations corrélée à la similarité lexicale lorsque les phrases considérées ont un vocabulaire très semblable mais de fines différences syntaxiques ou sémantiques déterminant la classe à prédire ; il peut également être souhaitable de chercher un espace avec des *clusters* distincts lorsque l'objectif est d'utiliser la proximité des voisins d'une relation dans l'espace latent comme notion de similarité du type de relation, afin d'éviter de mauvaises prédictions causées par le recouvrement de plusieurs *clusters*.

Notre étude sur deux jeux de données met donc en lumière l'influence de la syntaxe et de la représentation des entités sur les résultats de la tâche de typage de relations, ainsi que la variété des espaces latents appris par les différentes architectures de modèles pour cette tâche, aux propriétés langagières et spatiales diverses. Elle donne des indicateurs pour la compréhension et l'exploitation d'une modélisation de relations appropriée au traitement désiré, puisque que ceux-ci sont révélateurs de ces hétérogénéités dans les représentations des relations.

Références

ALT C., GABRYSZAK A. & HENNIG L. (2020). TACRED revisited : A thorough evaluation of the TACRED relation extraction task. In *58th Annual Meeting of the Association for Computational Linguistics*, p. 1558–1569, Online.

BHATTACHARYA I. & GETOOR L. (2007). Collective entity resolution in relational data. *ACM Transactions on Knowledge Discovery from Data*, **1**(1).

CHEN C.-Y. & LI C.-T. (2021). ZS-BERT : Towards zero-shot relation extraction with attribute representation learning. In *2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 3470–3479, Online.

CULOTTA A. & SORENSEN J. (2004). Dependency tree kernels for relation extraction. In *42nd Annual Meeting of the Association for Computational Linguistics*, p. 423–429, Barcelona, Spain.

HAN X., ZHU H., YU P., WANG Z., YAO Y., LIU Z. & SUN M. (2018). FewRel : A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *2018 Conference on Empirical Methods in Natural Language Processing*, p. 4803–4809, Brussels, Belgium.

HONG J., VOSS C. & MANNING C. (2021). Challenges for information extraction from dialogue in criminal law. In *1st Workshop on NLP for Positive Impact*, p. 71–81, Online.

HOULSBY N., GIURGIU A., JASTRZEBSKI S., MORRONE B., DE LAROUSSILHE Q., GESMUNDO A., ATTARIYAN M. & GELLY S. (2019). Parameter-efficient transfer learning for NLP. In *36th International Conference on Machine Learning*, p. 2790–2799, Long Beach, California, USA.

HUANG H. & WONG R. (2020). Deep embedding for relation extraction on insufficient labelled data. In *2020 International Joint Conference on Neural Networks*, p. 1–8, Glasgow, United Kingdom.

JABBARI A., SAUVAGE O., ZEINE H. & CHERGUI H. (2020). A French corpus and annotation schema for named entity recognition and relation extraction of financial news. In *12th Language Resources and Evaluation Conference*, p. 2293–2299, Marseille, France.

KARAA W. B. A., ALKHAMMASH E. H. & AIDA B. (2021). Drug disease relation extraction from biomedical literature using NLP and machine learning. *Mobile Information Systems, special issue Intelligent Data Analytics for Internet of Things-Based Applications*, **2021**.

KIM Y. (2014). Convolutional neural networks for sentence classification. In *2014 Conference on Empirical Methods in Natural Language Processing*, p. 1746–1751, Doha, Qatar.

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. In *20th China National Conference on Computational Linguistics*, p. 1218–1227, Hohhot, China.

NASAR Z., JAFFRY S. W. & MALIK M. K. (2022). Named entity recognition and relation extraction : State-of-the-art. *ACM Computing Surveys*, **54**(1).

REIMERS N. & GUREVYCH I. (2019). Sentence-BERT : Sentence embeddings using siamese BERT-networks. In *2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing*, p. 3982–3992, Hong Kong, China.

RIEDEL S., YAO L. & MCCALLUM A. (2010). Modeling relations and their mentions without labeled text. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, p. 148–163. Barcelona, Spain : Lecture Notes in Computer Science, vol. 6323, Springer.

SIMON E. (2022). *Deep Learning for Unsupervised Relation Extraction*. Thèse de doctorat, Sorbonne Université.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. & POLOSUKHIN I. (2017). Attention is all you need. In *31st Conference on Neural Information Processing Systems*, p. 6000–6010, Long Beach, California, USA.

YAMADA I., ASAI A., SHINDO H., TAKEDA H. & MATSUMOTO Y. (2020). LUKE : Deep contextualized entity representations with entity-aware self-attention. In *2020 Conference on Empirical Methods in Natural Language Processing*, p. 6442–6454, Online.

ZHANG Y., ZHONG V., CHEN D., ANGELI G. & MANNING C. D. (2017). Position-aware attention and supervised data improve slot filling. In *2017 Conference on Empirical Methods in Natural Language Processing*, p. 35–45, Copenhagen, Denmark.

ZHOU W. & CHEN M. (2022). An improved baseline for sentence-level relation extraction. In *2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and 12th International Joint Conference on Natural Language Processing (Volume 2 : Short Papers)*, p. 161–168, Online.