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

Extraction de relations sémantiques et modèles de langue : pour une relation à double sens

Olivier Ferret
Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
olivier.ferret@cea.fr

RÉSUMÉ

Les modèles de langue contextuels se sont rapidement imposés comme des outils essentiels du Traitement Automatique des Langues. Néanmoins, certains travaux ont montré que leurs capacités en termes de sémantique lexicale ne les distinguent pas vraiment sur ce plan de modèles plus anciens, comme les modèles statiques ou les modèles à base de comptes. Une des façons d'améliorer ces capacités est d'injecter dans les modèles contextuels des connaissances sémantiques. Dans cet article, nous proposons une méthode pour réaliser cette injection en nous appuyant sur des connaissances extraites automatiquement. Par ailleurs, nous proposons d'extraire de telles connaissances par deux voies différentes, l'une s'appuyant sur un modèle de langue statique, l'autre sur un modèle contextuel. Des évaluations réalisées pour l'anglais et focalisées sur la similarité sémantique ont montré l'intérêt de cette démarche, permettant d'enrichir sémantiquement un modèle de type BERT sans utilisation de ressources sémantiques externes.

ABSTRACT

Extraction of semantic relations and language models : for a two-way relationship

Contextual language models have rapidly become essential tools in Natural Language Processing. Nevertheless, some works have shown that their capabilities in terms of lexical semantics do not really distinguish them from older models, such as static models or count-based models. One way to improve these capabilities is to inject semantic knowledge into contextual models. In this paper, we propose a method to perform this injection based on automatically extracted knowledge. Moreover, we propose to extract such knowledge in two different ways, one based on a static language model, the other on a contextual model. Evaluations performed for English and focused on semantic similarity have shown the interest of this approach, allowing to semantically enrich a BERT model without using external semantic resources.

MOTS-CLÉS : Extraction de relations sémantiques lexicales, modèles de langue, injection de connaissances dans les modèles de langue.

KEYWORDS: Extraction of lexical semantic relations, language models, knowledge injection in language models.

1 Introduction

Les modèles de langue, qu'ils soient à base de comptes ou prédictifs (Baroni *et al.*, 2014), et parmi ces derniers, statiques ou contextuels (Naseem *et al.*, 2021), entretiennent une relation double vis-à-vis des connaissances sémantiques. D'une part, du fait de leur forte inscription dans l'hypothèse

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : travaux de recherche originaux – articles longs, pages 123–136.

distributionnelle (Harris, 1954), ils constituent un moyen utilisé de longue date pour extraire des relations sémantiques lexicales à partir de corpus (Lenci *et al.*, 2022). D'autre part, beaucoup de travaux se sont attachés au problème de l'injection de connaissances sémantiques dans ces modèles afin de les enrichir (Wang *et al.*, 2023), soit dans une perspective générale d'amélioration de la prise en compte des phénomènes sémantiques au niveau des tâches auxquels ils sont appliqués, soit pour leur adaptation à des domaines spécifiques.

L'utilisation des modèles de langue pour l'extraction de relations sémantiques est étroitement liée à la problématique de la similarité sémantique (Budanitsky & Hirst, 2006) et à celle des thésaurus distributionnels (Grefenstette, 1994; Lin, 1998; Curran & Moens, 2002). La façon la plus commune d'extraire des relations sémantiques à partir d'un modèle de langue est en effet de s'appuyer sur la capacité de ces modèles à évaluer la similarité des mots les uns par rapport aux autres sur une base distributionnelle, capacité utilisée par ailleurs pour l'évaluation intrinsèque de ces modèles (Faruqui *et al.*, 2016). Appliquée au vocabulaire d'un corpus, cette capacité permet de construire un thésaurus distributionnel donnant pour chaque mot cible une liste de voisins distributionnels, ordonnés selon la valeur décroissante de leur similarité, évaluée par un modèle de langue, avec le mot cible. Les premiers voisins sont alors supposés les plus pertinents sur le plan sémantique, avec un biais de principe vers les relations paradigmatiques compte tenu de l'hypothèse distributionnelle sous-jacente aux modèles de langue. Compte tenu de ce principe général, la voie principale d'amélioration de cette extraction concerne la similarité sémantique utilisée pour construire les thésaurus distributionnels (Padró *et al.*, 2014a,b). Néanmoins, quelques travaux se concentrent également sur une amélioration des thésaurus en tant que tels au moyen de méthodes de réordonnancement, soit à un niveau global (Claveau *et al.*, 2014), soit plus localement au niveau de chaque entrée du thésaurus (Ferret, 2013a).

La question de l'injection de connaissances sémantiques dans les modèles de langue a fait quant à elle l'objet d'un grand nombre d'études, d'abord axées sur les modèles neuronaux statiques pour ensuite se focaliser sur les modèles contextuels. Malgré les différences existant entre ces deux grands types de modèles, ils partagent la même distinction entre les méthodes opérant lors de la construction du modèle et celles venant enrichir un modèle après sa construction. Les secondes ont clairement l'avantage du nombre dans le cas des modèles statiques, dans le prolongement de Faruqui *et al.* (2015), tandis que la situation est plus contrastée pour les modèles contextuels. En se limitant aux relations sémantiques lexicales [1], on peut ainsi citer le modèle LIBERT de Lauscher *et al.* (2020) pour la première catégorie de méthodes et le modèle LexFit de Vulić *et al.* (2021) pour la seconde.

Le travail présenté dans cet article conjugue les deux dimensions esquissées ci-dessus : il enrichit un modèle neuronal contextuel de type BERT (Devlin *et al.*, 2019) par l'injection de connaissances sémantiques lexicales mais contrairement aux travaux existants, ces connaissances sont elles-mêmes extraites automatiquement par le biais de l'exploitation de modèles de langue neuronaux. Plus précisément, les contributions de ce travail sont :

— la proposition et l'évaluation d'une nouvelle méthode d'extraction de relations sémantiques lexicales entre termes simples en appliquant une tâche de type « mot masqué » à des termes complexes par le biais d'un modèle contextuel ;
— la comparaison et l'association des relations ainsi obtenues avec les relations extraites à partir d'un modèle de langue statique ;
— l'évaluation de l'intérêt de l'utilisation de relations lexicales sémantiques extraites automatiquement pour enrichir un modèle de langue contextuel.

1. Pour les modèles contextuels, les travaux existants sont axés sur des graphes de connaissances représentant des connaissances factuelles plus que sur des relations sémantiques lexicales, la tendance étant inverse pour les modèles statiques.

2 Méthodes

Dans ce qui suit, nous présentons d'abord à la section 2.1 deux méthodes d'extraction de relations sémantiques lexicales à partir de modèles neuronaux de types différents. Dans les deux cas, les relations extraites caractérisent une relation de similarité sémantique entre deux mots mais ne sont pas typées. L'union du produit de chacune de ces deux méthodes sert ensuite de base pour l'injection de connaissances sémantiques dans un modèle de type BERT, objet de la section 2.2.

2.1 Extraction de relations sémantiques

À partir d'un modèle neuronal statique. Pour extraire un premier ensemble de relations de similarité sémantique, nous transposons à un modèle neuronal statique le principe de sélection par réciprocité dans le graphe des k plus proches voisins (*k-NN*) présenté dans (Claveau *et al.*, 2014) pour des modèles à base de comptes. Plus précisément, pour chaque mot cible, ses k plus proches mots voisins sont extraits en s'appuyant sur les similarités données par les plongements du modèle statique considéré, en l'occurrence un modèle Skip-gram (Mikolov *et al.*, 2013a). Cette extraction est réalisée grâce à la bibliothèque Faiss (Johnson *et al.*, 2021) en utilisant classiquement la mesure de similarité *cosinus* [2]. La relation de voisinage distributionnel n'est pas symétrique par nature mais nous utilisons précisément l'observation d'une telle symétrie comme critère de sélection des relations de voisinage les plus représentatives en termes de similarité sémantique. Plus précisément, une telle relation entre les mots x et y est sélectionnée si y se trouve parmi les k premiers voisins distributionnels de x et réciproquement, si x se trouve parmi les k premiers voisins distributionnels de y.

À partir d'un modèle neuronal contextuel. La transposition de l'approche précédente des modèles neuronaux statiques aux modèles neuronaux contextuels est bien moins directe que celle des modèles à base de comptes aux modèles neuronaux statiques, en particulier parce qu'un modèle contextuel produit par définition des représentations de mots en contexte et non des représentations génériques. Le problème plus généralement posé pour réaliser cette transposition est de pouvoir construire à partir d'un modèle de langue un graphe de voisinage entre mots, le voisinage étant fondé sur la notion de similarité sémantique. Pour un modèle contextuel, deux stratégies principales sont envisageables :

— la construction de plongements statiques de mots, ce qui permet de se ramener à la configuration évoquée au point précédent ;
— l'exploitation des capacités d'un tel modèle pour la tâche de modélisation du langage à partir de laquelle il a été entraîné.

La première stratégie a déjà fait l'objet d'un certain nombre de travaux (Ethayarajh, 2019; Bommasani *et al.*, 2020; Vulić *et al.*, 2020; Ferret, 2022), avec deux variantes principales : l'une considère pour un mot cible un ensemble de phrases contenant ce mot et agrège, généralement par une moyenne, les représentations contextuelles produites par le modèle de langue pour ce mot dans chacune de ces phrases [3]. La seconde variante consiste à construire une représentation à partir d'une seule occurrence du mot cible en isolation, sans le contexte d'une phrase. Néanmoins, Ferret (2022) montre que du point de vue de la constitution d'un voisinage sémantique des mots, cette première stratégie ne donne pas de résultats notablement plus intéressants que des plongements statiques, avec tout de même un

2. Concrètement, nous utilisons l'index IndexFlatIP, conçu pour les recherches exactes fondées sur le produit scalaire.
3. Les modèles de langue contextuels existants étant constitués de plusieurs couches, la représentation d'une occurrence de mot admet elle-même différentes variantes.

avantage à la première variante par rapport à la seconde.

Nous avons donc opté pour la seconde stratégie. Nous nous concentrons ici sur les modèles de type BERT, qui reposent sur une tâche de modélisation du langage par mot masqué (*Masked Language Modeling*). Néanmoins, l'approche n'exclut pas pour autant l'utilisation de modèles auto-régressifs de type GPT (Radford *et al.*, 2018). Le principe général s'inspire de l'utilisation de modèles de langue de type BERT pour la substitution lexicale sans utilisation de substituts de référence (Zhou *et al.*, 2019). Néanmoins, au lieu de considérer des occurrences de mots dans le contexte de phrases, nous nous limitons à des occurrences de mots au sein de termes complexes. L'application de la substitution lexicale aux termes complexes se retrouve par ailleurs dans (Wang, 2022) mais dans le cadre de phrases et avec l'objectif différent de valider des relations sémantiques entre termes complexes à partir de relations connues entre termes simples. Dans notre cas, la restriction aux termes complexes, plus spécifiquement de nature nominale, se justifie en premier lieu par des raisons de coût de calcul, le traitement de termes par un modèle de type BERT étant nettement moins coûteux que celui de phrases [4]. Par ailleurs, les expériences concernant la similarité distributionnelle avec les modèles à base de compte ou les modèles neuronaux statiques montrent de façon récurrente que la similarité sémantique, par opposition à la proximité sémantique, est mieux capturée par un contexte étroit que par un contexte large, ce qui justifie de se limiter à des termes complexes. L'analyse des schémas d'attention dans les modèles BERT (Clark *et al.*, 2019) montre en outre que certaines de leurs têtes prennent en compte spécifiquement ces interactions à courte portée, laissant à penser que ce choix n'est pas trop limitant. Enfin, cette méthode permet également, dans le cadre de domaines de spécialité, d'exploiter non seulement des termes extraits de corpus mais également des termes issus de terminologies de référence pour ces domaines.

Concrètement, l'approche consiste à soumettre à un modèle BERT en mode prédiction de mot masqué un ensemble de termes dont l'un des constituants a été masqué et de recueillir les k premières prédictions du modèle, avec leur score, en excluant le constituant à prédire. Chaque terme ainsi soumis constitue une séquence autonome. L'hypothèse est que les prédictions obtenues correspondent à des voisins sémantiques du constituant cible. Appliqué à un grand nombre de termes complexes, cette méthode conduit à recueillir des voisins sémantiques pour un ensemble conséquent de mots simples, ce qui permet ensuite d'appliquer le principe de sélection par réciprocité dans le graphe des *k-NN* vu ci-dessus. Un point important de l'approche est le fait qu'un même mot peut se voir associer autant de listes de voisins que le nombre de fois où il apparaît comme mot masqué dans un terme complexe. Pour constituer une liste de voisins unique pour chaque mot, nous appliquons une méthode de fusion de listes, en l'occurrence la méthode CombSum (Fox & Shaw, 1994), en exploitant les scores de prédiction normalisés avec la méthode Zero-one (Lee, 1997; Wu *et al.*, 2006). Outre le fait d'assurer l'unicité de la liste des voisins d'un mot, cette fusion a l'intérêt de mettre en avant les substituts prédits régulièrement avec les meilleurs scores et donc de placer aux premiers rangs les voisins supposés les plus proches du mot cible.

La prédiction d'un modèle BERT concernant un mot à substituer est clairement dépendante du contexte linguistique de ce mot. Dans notre cas, ce contexte est déterminé par plusieurs facteurs : la forme des termes complexes dans lesquels ces mots apparaissent, le rôle que les mots cibles y jouent et enfin, le contexte plus général dans lequel les termes complexes sont placés. Pour ce qui est du premier facteur, le travail présenté ayant été réalisé pour l'anglais avec des noms pour cibles, nous avons observé, en prenant comme base la version anglaise de Wikipédia, les termes composés

4. La complexité du mécanisme d'attention des transformeurs est quadratique en fonction de la longueur de la séquence considérée.

incluant deux mots pleins, obtenant ainsi les trois structures de termes suivantes, la première étant environ deux fois plus fréquente que la deuxième, qui est elle-même environ vingt fois plus fréquente que la troisième [5] :

ADJ	NOM		rough estimate, wearable device, motherless child
NOM	NOM		prison guard, science academy, college student
NOM	PREP	NOM	lack of food, degree in education, return on investment

Concernant le dernier facteur, nous avons repris le schéma général proposé par Qiang *et al.* (2020) dans un contexte de simplification lexical et consistant à conditionner la structure contenant une unité à prédire par cette même structure sous sa forme complète. Dans notre cas de figure, si TERM désigne le terme utilisé comme contexte immédiat et TERM_MSK, ce même terme avec le « trou » correspondant au mot cible, la séquence, appelée amorce, soumise à un modèle de type BERT en mode prédiction de mot masqué est ainsi de la forme :

	TERM . [SEP] TERM_MSK .
soit par exemple ici :	ADJ NOM . [SEP] ADJ __ .
qui peut s'instancier en :	civil defence . [SEP] civil __ .
	black magic . [SEP] black __ .

où __ correspond à l'emplacement du mot cible masqué et [SEP] au marqueur de changement de séquence [6]. Cette amorce est appelée P0 dans ce qui suit.

Nous avons également testé les variantes suivantes, destinées en particulier à donner un contexte immédiat un peu plus large :

P1 this is a/an TERM . [SEP] this is a/an TERM_MSK .
P2 TERM . [SEP] this is a/an TERM_MSK .
P3 a/an TERM . [SEP] a/an TERM_MSK .
P4 TERM . [SEP] a/an TERM_MSK is a kind of TERM .
P5 TERM . [SEP] a/an TERM_MSK is a/an TERM .
P6 TERM . [SEP] a/an TERM is a/an TERM_MSK .
P7 TERM . [SEP] a/an TERM_MSK and a/an TERM .
P8 TERM . [SEP] a/an TERM_MSK or a/an TERM .

2.2 Injection de relations sémantiques dans un modèle contextuel

Pour l'injection des relations sémantiques extraites, nous nous sommes appuyés sur une approche contrastive relevant de l'apprentissage de métriques. Ce type d'approches a déjà été exploré pour réaliser cette tâche d'injection aussi bien pour les plongements statiques (Shah *et al.*, 2020) que pour les modèles de langue contextuels (Vulić *et al.*, 2021). Dans le cas présent, nous nous situons dans le cadre défini par Vulić *et al.* (2021), qui ont eux-mêmes réutilisé le cadre défini par Sentence-BERT pour la similarité de phrases (Reimers & Gurevych, 2019). Plus précisément, l'architecture de base

5. ADJ : adjectif ; PREP : préposition.
6. Nous avons repris l'utilisation de [SEP] de (Qiang *et al.*, 2020) mais la nécessité de sa présence reste à tester, en particulier dans le cas de l'utilisation d'un modèle de type RoBERTa (Liu *et al.*, 2019), qui n'est pas entraîné pour la tâche de prédiction de la phrase suivante.

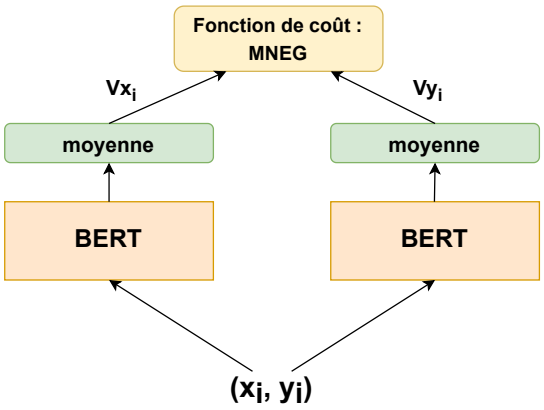

FIGURE 1: Architecture du modèle d'injection de relations sémantiques lexicales

de Sentence-BERT est celle d'un réseau siamois exploitant un encodeur double (*dual encoder*) : deux phrases dont on connaît la similarité sont encodées séparément par le même modèle de type BERT, une représentation de chacune des deux phrases est construite via un processus de regroupement (*pooling*) et les deux représentations ainsi obtenues sont prises en compte par une fonction de coût visant à rapprocher, au travers du mécanisme de rétropropagation, les représentations des phrases connues comme similaires tandis qu'elles tendent à écarter les représentations des phrases connues comme dissemblables. Le modèle LexFit de Vulić *et al.* (2021) réutilise directement cette architecture en donnant en entrée des couples de mots plutôt que des couples de phrases et en utilisant des relations lexicales sémantiques comme référence en termes de similarité, comme illustré par la figure 1. De nombreuses fonctions de coût sont possibles pour mettre en œuvre le principe général ci-dessus, et donc indirectement injecter ces relations lexicales dans un modèle de langue de type BERT ; mais au vu des expérimentations faites avec LexFit, nous avons choisi la Multiple Negatives Ranking (MNEG) loss (Henderson *et al.*, 2019), qui se définit ainsi pour un lot (*batch*) de B paires de mots $(x_1, y_1), \ldots, (x_B, y_B)$ telles que chaque paire (x_i, y_i) est sous-tendue par une relation sémantique :

$$\mathcal{L} = -\sum_{i=1}^{B} S(x_i, y_i) + \sum_{i=1}^{B} \log \sum_{j=1, j\neq i}^{B} e^{S(x_i, y_j)} \quad (1)$$

Cette fonction de coût permet d'adapter le modèle de langue de l'encodeur de façon à maximiser la similarité de chaque paire de mots (x_i, y_i) du lot (1^{er} terme de l'équation 1) tout en minimisant la similarité des $B - 1$ paires (x_i, y_j) (2^{nd} terme de l'équation 1) formées chacune d'un x_i et de l'y_j de toutes les autres paires du lot, ces (x_i, y_j) étant considérées comme des exemples négatifs de similarité. $S(x_i, y_i)$ correspond à la fonction utilisée pour évaluer la similarité de la paire (x_i, y_i).

3 Expérimentations

3.1 Cadre d'évaluation

Pour évaluer les résultats de notre processus d'injection de relations sémantiques lexicales dans un modèle de langue, nous avons choisi de tester les capacités du modèle cible en termes de similarité sémantique par le biais de plongements construits à partir de ce modèle. À la suite de Budanitsky &

Hirst (2006), nous distinguons la similarité sémantique, incarnée par les relations paradigmatiques (synonymie, hyperonymie...), de la proximité sémantique, que l'on retrouve dans les relations sémantiques de nature plus syntagmatique (comme les relations de prédication par exemple). Notre modèle cible est un modèle de type BERT, dans sa version `base-uncased`, et à l'instar de (Vulić *et al.*, 2021), nous construisons le plongement d'un mot à partir de l'encodage par ce modèle d'une seule occurrence de ce mot hors contexte en sélectionnant la représentation de cette occurrence produite au niveau de l'une des 13 couches du modèle (12 couches internes plus la couche d'entrée). Par ailleurs, comme (Bommasani *et al.*, 2020), lorsqu'un mot se décompose en plusieurs sous-mots (*wordpieces*), nous construisons sa représentation en moyennant les représentations de ses sous-mots.

L'évaluation en elle-même s'appuie sur la similarité entre les représentations ainsi produites pour les mots : pour chaque mot cible w_i, l'ensemble de ses k plus proches voisins est sélectionné en calculant la similarité de w_i avec tous les autres mots cibles w_j, en appliquant la mesure *cosinus* à leurs représentations et en ordonnant ces mots selon la valeur décroissante de leur similarité avec w_i. En pratique, $k = 10$ et le calcul des similarités est réalisée là encore grâce à la bibliothèque Faiss. Nous évaluons l'exactitude de ce classement comme en recherche d'information grâce à la R-précision ($R_{préc.}$), la MAP (Mean Average Precision) et aux précisions à différents rangs (P@r). Les évaluations ont été menées pour 10 305 noms cibles déjà utilisés dans (Ferret, 2022) et couvrant un large éventail de fréquences.

Tout en nous concentrant globalement sur les relations de nature paradigmatique, nous considérons deux références, toutes deux issues de WordNet (Miller, 1990) puisque nos mots cibles sont des noms communs en anglais : *para*, qui rassemble les relations de synonymie, d'hyponymie, d'hyperonymie et de cohyponymie et *syn*, qui se limite au sous-ensemble des synonymes. Nous définissons plus précisément les mots liés à un mot cible par ces différentes relations de la façon suivante :

— synonymes : tous les mots faisant partie d'un synset S_i du mot cible considéré ;
— hyperonymes : tous les mots des synsets S_{hype} ayant un lien direct d'hyperonymie avec un S_i ;
— hyponymes : tous les mots des synsets ayant une relation directe d'hyponymie avec S_i ;
— cohyponymes : tous les mots des synsets, à l'exception des S_i, ayant une relation directe d'hyponymie avec les synsets S_{hype}.

3.2 Mise en œuvre des méthodes proposées

La mise en œuvre de la première méthode d'extraction de relations lexicales sémantiques nécessite de disposer de plongements statiques tandis que la seconde demande principalement un ensemble de termes complexes. Pour ces deux ressources, nous nous sommes appuyés sur un même corpus de base, en l'occurrence un dump de Wikipédia en anglais du 1/10/2018 comprenant 2,16 milliards de tokens [7], étiquetés et lemmatisés grâce à l'outil CoreNLP (Manning *et al.*, 2014), dans sa version 3.9.2. Les termes complexes, limités ici à des bigrammes de mots pleins, ont été extraits par la méthode définie dans (Mikolov *et al.*, 2013b) [8], avec une fréquence minimale des termes extraits égale à 5 et un seuil d'information mutuelle minimale égal à 0. Nous avons ainsi obtenu 394 024 termes de structure ADJ NOUN pour constituer ensuite nos amorces. Pour chaque amorce, nous ne retenons que les 10 premières propositions du modèle, hors le terme masqué si celui-ci apparaît dans les premières propositions. Les plongements statiques ont été appris suivant le modèle Skip-gram avec l'outil *word2vec* [9] à partir de la forme lemmatisée des mots.

7. https://www.dropbox.com/s/cnrhd11zdtc1pic/enwiki-20181001-corpus.xml.bz2?dl=0
8. Suivant l'implémentation de Gensim : https://radimrehurek.com/gensim/models/phrases.html
9. Avec les paramètres : -size 300 -window 5 -negative 10 -hs 0 -sample 1e-5 -min-count 5

modèle	réf.	$R_{préc}$	MAP	P@1	P@2	P@5	P@10
fastText-wiki	para	9,9	6,0	36,5	29,9	21,3	15,9
	syn	15,5	18,4	21,9	15,7	9,2	5,8
BERT ctxt	para	9,5	5,7	36,5	30,4	22,4	17,0
	syn	15,6	17,9	21,8	16,0	9,5	6,1
BERT iso	para	7,4	4,4	30,9	26,2	19,6	14,6
	syn	14,0	15,8	19,2	14,6	8,7	5,5
BERT réfsyn	para	17,2	12,3	55,7	48,5	37,4	29,0
	syn	27,0	31,9	35,9	27,8	17,4	11,4

TABLE 1: Points de référence pour l'évaluation (valeurs x100)

Pour la mise en œuvre de la méthode d'injection, nous avons eu recours à la bibliothèque Sentence-Transformers [10]. Pour chaque relation (m_1, m_2), nous avons aussi considéré la relation (m_2, m_1), ce qui peut être vu comme une forme très simple d'augmentation de données. Comme le laisse apparaître l'équation 1, l'ensemble des relations à injecter est traité par lots, d'une taille de 512 relations chacun. L'apprentissage se fait en 10 époques, avec un taux d'apprentissage de $2e-5$, l'utilisation de l'optimiseur AdamW (Loshchilov & Hutter, 2019) et un nombre d'étapes d'échauffement égal à 10 % des relations à injecter, avec un schéma d'échauffement linéaire.

3.3 Évaluation des méthodes proposées

Points de référence. Le tableau 1 donne un certain nombre de références concernant notre évaluation du résultat de l'injection de relations sémantiques. La première de ces références, *fastText-wiki*, donne les performances en termes de similarité du modèle Skip-gram utilisé par Vulić *et al.* (2021) comme référence, appris à partir de la version anglaise de Wikipédia avec l'outil *fastText* (Bojanowski *et al.*, 2017). La deuxième, *BERT ctxt*, donne ces mêmes performances pour des plongements construits selon la méthode de Bommasani *et al.* (2020), c'est-à-dire en moyennant les représentations des occurrences des mots cibles apparaissant dans un ensemble de phrases. À l'instar de Ferret (2022), nous prenons 10 phrases par mot cible et les meilleurs résultats sont obtenus avec la couche L5. Comme on peut le constater, ces deux premières références sont très proches, indiquant au passage que concernant la similarité sémantique, modèles statiques et modèles contextuels sont très proches comme cela a déjà été observé par d'autres travaux (Lenci *et al.*, 2022). *BERT iso* correspond quant à elle au point départ de notre processus d'injection de relations (cf. section 2.1), les résultats du tableau 1 pour ce modèle étant issus de la couche L0. Le fait d'utiliser une seule occurrence sans contexte pour les mots cibles a clairement une incidence négative sur la performance comme le montre la comparaison avec *BERT ctxt* mais mobilise nettement moins de ressources. Notre dernière référence, *BERT réfsyn*, peut être considérée comme notre référence haute puisqu'elle correspond à l'injection faite dans un modèle BERT par Vulić *et al.* (2021) de 1 023 082 relations sémantiques issues de ressources constituées manuellement, en l'occurrence WordNet et le thésaurus Roget [11]. Les mesures sont données pour la couche 12 et le nombre d'époques dans ce cas est réduit à 2 compte tenu du nombre de relations.

10. https://www.sbert.net/
11. Il s'agit plus précisément de notre reproduction du travail de Vulić *et al.* (2021), dont le code n'est pas disponible.

	para	syn	# relations
BERT	13,9	5,2	17 007
CBERT	22,9	10,9	17 023
CBERT – têtes	24,2	11,8	13 465
CBERT – modifieurs	16,0	6,7	7 792
CBERT – ADJ NOM	26,2	12,4	10 511

TABLE 2: Exactitude (x100) des relations extraites par un modèle contextuel suivant le type de modèle et la structure des termes amorces

	P0	P1	P2	P3	P4	P5	P6	P7
para	32,0	30,3	24,9	31,1	30,9	31,7	26,5	31,7
syn	16,0	15,8	12,5	15,6	16,0	16,8	13,0	15,7

TABLE 3: Exactitude (x100) des relations extraites par les types formes d'amorces

Extraction de relations à partir d'un modèle contextuel. La méthode que nous avons proposée à la section 2.1 pour extraire des relations sémantiques à partir d'un modèle contextuel pose un certain nombre de questions auxquelles nous essayons de répondre ici en commençant par aborder, au travers des résultats du tableau 2, le problème du type de modèle et de la structure syntaxique des termes servant d'amorce. Les résultats sont donnés en termes d'exactitude des relations extraites par rapport à nos deux références. Il est à noter qu'ils ont été obtenus pour une forme générale d'amorce correspondant à P0 mais avec TERM et TERM_MSK faisant partie d'une même séquence. Le seuil d'information mutuelle minimale pour les termes extraits (cf. méthode de Mikolov *et al.* (2013b) évoquée ci-dessus) était par ailleurs égal à 10. Les deux premières lignes comparent le modèle BERT *base-uncased* avec le modèle CharacterBERT (El Boukkouri *et al.*, 2020), équivalent au modèle BERT base en termes de structure mais présentant la caractéristique de ne pas découper les mots en sous-mots. Cette comparaison permet de constater l'impact très notable de ce découpage des mots pour notre tâche d'extraction de relations, avec un très net avantage pour le modèle CharacterBERT, qui sera utilisé dans ce qui suit.

Les trois lignes suivantes ont trait quant à elles à la structure syntaxique des termes amorces et à la place qu'y occupe le mot cible. Nous constatons tout d'abord grâce aux deux première lignes que le mot cible en position de modifieur sur le plan syntaxique produit moins de relations qu'en position de tête et surtout, que ces relations sont beaucoup plus bruitées. Cette observation va dans le sens des travaux de Ferret (2013b), qui a montré qu'au sein de deux termes complexes entretenant un lien de similarité sémantique, il est plus probable d'avoir une relation de similarité sémantique entre les têtes syntaxiques des termes pour un même modifieur que le contraire, ce qui conduit à privilégier les termes ayant le mot cible pour tête syntaxique. La dernière ligne permet enfin de constater que la très grande majorité des relations sont obtenues à partir de la structure de terme ADJ NOM (NOM étant le mot cible), avec là encore des relations moins bruitées que pour les autres structures. Nous ne retiendrons donc pour TERM et TERM_MSK que des termes de type ADJ NOM.

Le tableau 3 permet pour sa part de juger des performances des différents types d'amorces présentés à la section 2.1, toujours avec un seuil d'information mutuelle minimale égal à 10 pour l'extraction des termes. Si la plupart de ces amorces donnent des résultats voisins, il faut remarquer que P2 et

relations	modèle	para	syn	# relations	# mots
extraites	statique	30,0	19,4	35 246	35 246
	contextuel	30,6	15,6	15 473	17 019
sélectionnées	statique	44,1	34,0	11 298	11 298
	contextuel	42,6	21,3	8 558	5 507
	fusion	41,1	24,2	18 430	14 199

TABLE 4: Exactitude (x100) et volumétrie des relations extraites puis sélectionnées

P6 obtiennent des résultats nettement inférieurs aux autres, sans qu'une raison très évidente puisse expliquer ce constat. La conclusion quant à la sensibilité par rapport à la forme des amorces est donc incertaine : si cette sensibilité n'est globalement pas très forte, elle peut être ponctuellement marquée, sans explication très claire. Dans ce qui suit, nous retiendrons l'amorce P0, la plus simple et une des deux meilleures.

Extraction et sélection de relations : synthèse. L'extraction des relations à partir d'un modèle statique telle que décrite à la section 2.1 ne demande de fixer que la taille du voisinage k et les mots cibles considérés. Dans le cas présent, nous avons retenu comme cibles les mots ayant une fréquence supérieure à 200 dans Wikipédia et une valeur $k = 1$ pour le voisinage. Il faut souligner à cet égard une différence importante entre les deux types de modèles : alors que le voisinage se limite au premier voisin pour le modèle statique, la qualité décroissant rapidement au-delà, nous l'étendons aux cinq premiers voisins pour le modèle contextuel, la dégradation de la qualité des relations étant beaucoup plus limitée à mesure de l'augmentation du rang.

Le tableau 4 évalue la qualité des relations extraites par nos deux types de modèles puis sélectionnées par réciprocité dans le graphe des *k-NN*, en regard avec leur volumétrie et le nombre de mots impliqués dans ces relations. Cette évaluation est faite sur la base de la présence de ces relations dans nos deux références, *syn* et *para*. Le modèle statique produit beaucoup plus de relations mais les deux types de modèles sont plus proches en termes de qualité, avec une équivalence sur l'ensemble des relations mais un avantage pour le modèle statique concernant les relations de synonymie. La sélection par réciprocité reproduit, et accentue même, ce biais initial pour ce qui est de la qualité des relations mais l'atténue fortement pour la volumétrie. Finalement, la fusion des deux ensembles de relations permet de constater leur complémentarité, avec un recoupement assez faible entre les deux et un niveau de performance plus proche du modèle contextuel que du modèle statique.

Enrichissement d'un modèle de langue. La dernière partie de notre évaluation concerne les résultats de l'injection des relations extraites dans un modèle de type BERT, illustrés par le tableau 5. Ce dernier rappelle notre point de départ, *BERT iso*, et donne les résultats pour chaque ensemble de relations extraites : celles issues du modèle statique, celles issues du modèle contextuel et la fusion de ces deux ensembles. Le premier constat est que l'injection des relations permet d'obtenir un gain de performance très significatif [12] par rapport à *BERT iso*, en particulier pour toutes les mesures P@r. Ce gain est globalement assez comparable pour les relations issues du modèle statique et celles issues du modèle contextuel. Il suit logiquement le même biais que les relations injectées. Il est ainsi plus important pour la synonymie dans le cas des relations du modèle statique et plus marqué pour

12. La significativité statistique des différences entre *BERT iso* et les autres modèles a été évaluée grâce à un test de Wilcoxon pour échantillons appariés avec $p < 0.01$.

modèle (couche)	réf.	$R_{préc}$	MAP	P@1	P@2	P@5	P@10
BERT iso (L0)	para	7,4	4,4	30,9	26,2	19,6	14,6
	syn	14,0	15,8	19,2	14,6	8,7	5,5
modèle statique (L11)	para	11,7	7,4	42,2	35,4	26,1	19,7
	syn	18,8	21,7	25,9	19,0	11,2	7,0
modèle contextuel (L11)	para	11,9	7,6	42,8	36,2	26,9	20,3
	syn	18,4	21,5	25,4	18,9	11,1	7,1
fusion (L12)	para	12,1	7,8	44,2	36,9	27,0	20,4
	syn	19,2	22,2	26,7	19,3	11,3	7,1

TABLE 5: Résultats de l'injection des relations sémantiques extraites (valeurs x100)

l'ensemble de nos relations sémantiques de référence en considérant les relations issues du modèle contextuel. La fusion des deux ensembles de relations permet d'obtenir un gain de performance supplémentaire pour nos deux références. Le niveau final atteint reste bien entendu en deçà du niveau constaté après l'injection de relations issues de ressources construites manuellement (cf. tableau 1) mais il est tout de même intéressant de constater qu'en l'absence de telles ressources, surtout dans les quantités utilisées par Vulić *et al.* (2021), il est tout de même possible d'enrichir sémantiquement un modèle de type BERT avec des relations sémantiques acquises automatiquement.

4 Conclusion et perspectives

Dans cet article, nous avons présenté à la fois deux méthodes pour extraire des relations sémantiques lexicales, l'une à partir d'un modèle de langue statique, l'autre à partir d'un modèle contextuel, et une méthode pour injecter ces relations afin d'enrichir sémantiquement un modèle contextuel. Les évaluations menées en termes de similarité sémantique ont montré le caractère effectif de la démarche proposée, qui permet de s'affranchir, au moins en partie, de relations sémantiques définies manuellement.

Outre le fait d'étendre les évaluations menées à d'autres tâches, un prolongement naturel de ce travail est d'étudier plus avant différentes formes d'amorces pour l'extraction de relations à partir de termes complexes et de comprendre plus précisément, notamment par l'observation des mécanismes d'attention, pourquoi certaines formes sont plus intéressantes que d'autres. Cette étude va par ailleurs de pair avec la prise en compte d'un ensemble plus vaste de structures de termes que la structure ADJ NOM principalement considérée ici. Le test d'autres modèles contextuels, tels que les modèles auto-régressifs, fait aussi partie des extensions très directes du travail présenté. Au-delà, nous souhaiterions également tester si un modèle contextuel enrichi sémantiquement comme nous l'avons réalisé pourrait conduire à une meilleure extraction de relations sémantiques, pouvant à son tour conduire à un meilleur enrichissement selon une procédure d'amorçage.

Remerciements

Nous remercions les relecteurs pour leur retour constructif. Ces travaux ont été réalisés grâce au supercalculateur Factory-IA financé par le Conseil Régional d'Île-de-France.

Références

BARONI M., DINU G. & KRUSZEWSKI G. (2014). Don't count, predict ! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, p. 238–247, Baltimore, Maryland.

BOJANOWSKI P., GRAVE E., JOULIN A. & MIKOLOV T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, **5**, 135–146. DOI : 10.1162/tacl_a_00051.

BOMMASANI R., DAVIS K. & CARDIE C. (2020). Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In *58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, p. 4758–4781, Online.

BUDANITSKY A. & HIRST G. (2006). Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*, **32**(1), 13–47.

CLARK K., KHANDELWAL U., LEVY O. & MANNING C. D. (2019). What does BERT look at ? an analysis of BERT's attention. In *2019 ACL Workshop BlackboxNLP : Analyzing and Interpreting Neural Networks for NLP*, p. 276–286, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/W19-4828.

CLAVEAU V., KIJAK E. & FERRET O. (2014). Improving distributional thesauri by exploring the graph of neighbors. In *25th International Conference on Computational Linguistics (COLING 2014)*, p. 709–720, Dublin, Ireland.

CURRAN J. R. & MOENS M. (2002). Improvements in automatic thesaurus extraction. In *Workshop of the ACL Special Interest Group on the Lexicon (SIGLEX)*, p. 59–66, Philadelphia, USA.

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

EL BOUKKOURI H., FERRET O., LAVERGNE T., NOJI H., ZWEIGENBAUM P. & TSUJII J. (2020). CharacterBERT : Reconciling ELMo and BERT for word-level open-vocabulary representations from characters. In *28th International Conference on Computational Linguistics (COLING 2020)*, p. 6903–6915, Barcelona, Spain (Online : International Committee on Computational Linguistics.

ETHAYARAJH K. (2019). How Contextual are Contextualized Word Representations ? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. In *2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019)*, p. 55–65, Hong Kong, China. DOI : 10.18653/v1/D19-1006.

FARUQUI M., DODGE J., JAUHAR S. K., DYER C., HOVY E. & SMITH N. A. (2015). Retrofitting Word Vectors to Semantic Lexicons. In *2015 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (NAACL HLT 2015)*, p. 1606–1615, Denver, Colorado.

FARUQUI M., TSVETKOV Y., RASTOGI P. & DYER C. (2016). Problems with evaluation of word embeddings using word similarity tasks. In *Workshop on Evaluating Vector-Space Representations for NLP (RepEval 2016)*, p. 30–35, Berlin, Germany : Association for Computational Linguistics. DOI : 10.18653/v1/W16-2506.

FERRET O. (2013a). Identifying bad semantic neighbors for improving distributional thesauri. In *51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, p. 561–571.

FERRET O. (2013b). Sélection non supervisée de relations sémantiques pour améliorer un thésaurus distributionnel. In *20ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2013)*, p. 48–61, Les Sables d'Olonne, France.

FERRET O. (2022). Building static embeddings from contextual ones : Is it useful for building distributional thesauri ? In *13th Language Resources and Evaluation Conference (LREC 2022)*, p. 2583–2590, Marseille, France.

FOX E. A. & SHAW J. A. (1994). Combination of multiple searches. In *2nd Text REtrieval Conference (TREC-2)*, volume 243 : NIST.

GREFENSTETTE G. (1994). *Explorations in automatic thesaurus discovery*. Kluwer Academic Publishers.

HARRIS Z. S. (1954). Distributional Structure. *Word*, **10**(2-3), 146–162.

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 *57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, p. 5392–5404, Florence, Italy. DOI : 10.18653/v1/P19-1536.

JOHNSON J., DOUZE M. & JÉGOU H. (2021). Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, **7**(3), 535–547.

LAUSCHER A., VULIĆ I., PONTI E. M., KORHONEN A. & GLAVAŠ G. (2020). Specializing unsupervised pretraining models for word-level semantic similarity. In *28th International Conference on Computational Linguistics (COLING 2020)*, p. 1371–1383, Barcelona, Spain (Online). DOI : 10.18653/v1/2020.coling-main.118.

LEE J. H. (1997). Analyses of multiple evidence combination. In *20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'97)*, p. 267—-276, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/258525.258587.

LENCI A., SAHLGREN M., JEUNIAUX P., CUBA GYLLENSTEN A. & MILIANI M. (2022). A comparative evaluation and analysis of three generations of Distributional Semantic Models. *Language Resources and Evaluation*. DOI : 10.1007/s10579-021-09575-z.

LIN D. (1998). Automatic retrieval and clustering of similar words. In *17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics (ACL-COLING'98)*, p. 768–774, Montréal, Canada.

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. *arXiv preprint arXiv :1907.11692*.

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

MANNING C. D., SURDEANU M., BAUER J., FINKEL J., BETHARD S. J. & MCCLOSKY D. (2014). The Stanford CoreNLP natural language processing toolkit. In *52nd Annual Meeting of the Association for Computational Linguistics : System Demonstrations (ACL 2014)*, p. 55–60.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013a). Efficient estimation of word representations in vector space. In *ICLR 2013, workshop track*.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. S. & DEAN J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, p. 3111–3119.

MILLER G. A. (1990). WordNet : An On-Line Lexical Database. *International Journal of Lexicography*, **3**(4).

NASEEM U., RAZZAK I., KHAN S. K. & PRASAD M. (2021). A Comprehensive Survey on Word Representation Models : From Classical to State-of-the-Art Word Representation Language Models. *ACM Transactions on Asian and Low-Resource Language Information Processing*, **20**(5), 74 :1–74 :35. DOI : 10.1145/3434237.

PADRÓ M., IDIART M., VILLAVICENCIO A. & RAMISCH C. (2014a). Comparing similarity measures for distributional thesauri. In *Ninth International Conference on Language Resources and Evaluation (LREC'14)*, p. 2964–2971, Reykjavik, Iceland : European Language Resources Association (ELRA).

PADRÓ M., IDIART M., VILLAVICENCIO A. & RAMISCH C. (2014b). Nothing like good old frequency : Studying context filters for distributional thesauri. In *2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, p. 419–424, Doha, Qatar : Association for Computational Linguistics. DOI : 10.3115/v1/D14-1047.

QIANG J., LI Y., ZHU Y., YUAN Y. & WU X. (2020). *Lexical Simplification with Pretrained Encoders*. Rapport interne, arXiv preprint arXiv :1907.06226.

RADFORD A., NARASIMHAN K., SALIMANS T. & SUTSKEVER I. (2018). *Improving language understanding by generative pre-training*. Rapport interne, OpenAI.

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. DOI : 10.18653/v1/D19-1410.

SHAH S., REDDY S. & BHATTACHARYYA P. (2020). A retrofitting model for incorporating semantic relations into word embeddings. In *Proceedings of the 28th International Conference on Computational Linguistics*, p. 1292–1298, Barcelona, Spain (Online). DOI : 10.18653/v1/2020.coling-main.111.

VULIĆ I., PONTI E. M., KORHONEN A. & GLAVAŠ G. (2021). LexFit : Lexical fine-tuning of pretrained language models. In *59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJNLP 2021)*, p. 5269–5283, Online. DOI : 10.18653/v1/2021.acl-long.410.

VULIĆ I., PONTI E. M., LITSCHKO R., GLAVAŠ G. & KORHONEN A. (2020). Probing Pretrained Language Models for Lexical Semantics. In *2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, p. 7222–7240, Online. DOI : 10.18653/v1/2020.emnlp-main.586.

WANG Y. (2022). *Exploration des relations terminologiques entre les termes multi-mots dans les modèles de sémantique distributionnelle*. Thèse de doctorat, Université Toulouse-Jean Jaurès.

WANG Y., WANG W., CHEN Q., HUANG K., NGUYEN A., DE S. & HUSSAIN A. (2023). Fusing external knowledge resources for natural language understanding techniques : A survey. *Information Fusion*, **92**, 190–204. DOI : https ://doi.org/10.1016/j.inffus.2022.11.025.

WU S., CRESTANI F. & BI Y. (2006). Evaluating score normalization methods in data fusion. In *Third Asia Conference on Information Retrieval Technology (AIRS'06)*, p. 642–648 : Springer-Verlag.

ZHOU W., GE T., XU K., WEI F. & ZHOU M. (2019). BERT-based lexical substitution. In *57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, p. 3368–3373, Florence, Italy. DOI : 10.18653/v1/P19-1328.