Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles
Lille, France, 28 juin au 2 juillet 2021

Exploration des relations sémantiques sous-jacentes aux plongements contextuels de mots

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

RÉSUMÉ

De nombreuses études ont récemment été réalisées pour étudier les propriétés des modèles de langue contextuels mais, de manière surprenante, seules quelques-unes d'entre elles se concentrent sur les propriétés de ces modèles en termes de similarité sémantique. Dans cet article, nous proposons d'abord, en nous appuyant sur le principe distributionnel de substituabilité, une méthode permettant d'utiliser ces modèles pour ordonner un ensemble de mots cibles en fonction de leur similarité avec un mot source. Nous appliquons d'abord cette méthode pour l'anglais comme mécanisme de sondage pour explorer les propriétés sémantiques des modèles ELMo et BERT du point de vue des relations paradigmatiques de WordNet et dans le contexte contrôlé du corpus SemCor. Dans un second temps, nous la transposons à l'étude des différences entre ces modèles contextuels et un modèle de plongement statique.

ABSTRACT

Exploring semantic relations underlying contextual word embeddings.

Many studies were recently done for investigating the properties of contextual language models but surprisingly, only a few of them focus on the properties of these models in terms of semantic similarity. In this article, we first propose a method that exploits, by relying on the distributional principle of substitutability, these models for ranking a set of target words according to their similarity with a source word. Then, we apply this method for English as a probing mechanism for investigating the semantic properties of ELMo and BERT models from the viewpoint of WordNet's paradigmatic relations and in the controlled context of SemCor. Finally, we adapt and apply this method to the study of differences between these contextual models and static embeddings.

MOTS-CLÉS : Modèles de langue contextuels, sémantique distributionnelle, relations sémantiques.

KEYWORDS: Contextual language models, distributional semantics, semantic relations.

1 Introduction

L'introduction de plongements contextuels de mots tels qu'ELMo (Peters *et al.*, 2018) ou BERT (Devlin *et al.*, 2019) est considérée comme une avancée majeure dans le traitement automatique des langues, en particulier pour les tâches de classification ou d'étiquetage de séquences traitées par des approches d'apprentissage supervisé. Cependant, ce type de plongements représente également un changement significatif du point de vue de la sémantique distributionnelle. Alors que les approches

Volume 1 : conférence principale, pages 26–36.

antérieures, y compris les plongements statiques de mots tels que les modèles Skip-gram ou CBOW (Mikolov *et al.*, 2013), construisaient la représentation distributionnelle d'un mot en cumulant ses contextes d'occurrence, les modèles de langage neuronaux tels qu'ELMo ou BERT produisent une représentation pour chaque occurrence des mots, ce qui, à la fois, pose des problèmes pour l'étude des propriétés sémantiques de ces représentations mais ouvrent aussi certaines opportunités.

Comme l'illustrent les travaux de Rogers *et al.* (2020), les propriétés des plongements contextuels de mots sont au centre de nombreuses études récentes, notamment dans le cas de BERT. Rogers *et al.* (2020) signalent principalement les travaux portant sur la sémantique au niveau phrastique, soit dans le contexte de l'étiquetage en rôles sémantiques (Tenney *et al.*, 2019), soit dans des tâches relevant de la substitution lexicale (Ettinger, 2020; Garí Soler *et al.*, 2019). Mais les travaux récents de Vulić *et al.* (2020) présentent clairement le plus grand ensemble d'expériences concernant les propriétés sémantiques des modèles BERT et RoBERTa, y compris concernant leur corrélation avec des jugements humains en termes de similarité sémantique. D'autres études ont également examiné les plongements contextuels d'un point de vue sémantique mais dans un contexte plus spécifique : l'impact des fonctions objectifs utilisées pour l'entraînement de ces modèles (Mickus *et al.*, 2020), leur niveau de contextualisation (Ethayarajh, 2019), leurs biais possibles (Bommasani *et al.*, 2020), leur capacité à représenter les sens des mots (Coenen *et al.*, 2019; Chronis & Erk, 2020) ou à construire des représentations pour des mots rares (Schick & Schütze, 2020).

Bien que notre travail ait des liens avec certains de ces travaux, son étude des plongements contextuels repose sur une méthode spécifique, fondée sur des principes distributionnels, qui vise à caractériser les plongements contextuels en termes de relations paradigmatiques à la fois au niveau des occurrences de mots et des mots. Plus précisément, nous présentons les contributions suivantes :

— nous montrons que l'hypothèse distributionnelle peut être appliquée au niveau des occurrences de mots avec des plongements contextuels et que plus spécifiquement, ELMo et BERT présentent des propriétés sémantiques propres relevant davantage de la similarité sémantique que de la proximité sémantique ;

— nous montrons comment la transposition de cette méthode au niveau des mots peut servir à étudier les différences entre les plongements contextuels et les plongements statiques de mots.

2 Méthode de test de plongements contextuels

2.1 Principes

Avec les modèles de plongements contextuels, la similarité entre les plongements de mots ne peut être calculée que pour les mots en contexte, c'est-à-dire les occurrences de mots, ce qui rend l'application des principes distributionnels au niveau des mots plus difficile mais offre la possibilité de les appliquer au niveau de ces occurrences. Plus précisément, selon Harris (1954), l'hypothèse distributionnelle se définit par :

« [...] difference of meaning correlates with difference of distribution. »

ce qui, exprimé de façon plus développée, donne :

« If A and B have almost identical environments except chiefly for sentences which contain both, we say they are synonyms : *oculist* and *eye-doctor*. If A and B have some environments in common and some not (e.g.

oculist and *lawyer*) we say that they have different meanings, the amount of meaning difference corresponding roughly to the amount of difference in their environments. »

Ce principe conduit à l'idée de *substituabilité*, que l'on retrouve dans :

« We group A and B into a substitution set whenever A and B each have the same (or partially same) environments X [...] »

Classiquement, les environnements mentionnés par Harris (1954) font référence aux ensembles de cooccurrents de *A* et *B*. Dans le cas des approches prédictives (Baroni *et al.*, 2014), ces environnements sont condensés dans des représentations distribuées appelées plongements de mots. Ainsi, selon les principes ci-dessus, plus deux plongements sont proches l'un de l'autre, plus les mots qui leur sont associés sont substituables et donc, plus ces mots sont susceptibles d'entretenir une relation sémantique proche de la synonymie. Dans le cas des modèles contextuels, le plongement construit pour une occurrence d'un mot avec ce type de modèles intègre deux dimensions : l'une, résultant de l'entraînement sur la tâche de modélisation linguistique, prend en compte l'agrégation de tous les contextes des occurrences du mot, comme pour tous les modèles distributionnels ; la seconde caractérise le contexte local de l'occurrence considérée, c'est-à-dire les mots qui l'entourent.

Dans notre cas, nous nous concentrons sur la première dimension pour évaluer la similarité entre les mots, ce qui nécessite de pouvoir contrôler la seconde. Nous proposons d'effectuer ce contrôle très simplement en nous appuyant sur le principe de base de l'approche distributionnelle : pour évaluer la similarité de deux mots, nous plaçons les deux mots dans le même contexte, c'est-à-dire à la même position dans la même phrase, et calculons un plongement de mot pour chacun d'eux en appliquant le modèle de langage contextuel considéré. En pratique, ce contrôle est mis en œuvre par une substitution : pour deux mots w_1 et w_2, une phrase S_i contenant une occurrence de w_1 est sélectionnée et une représentation contextualisée de w_1 est construite ; w_1 est ensuite remplacé par w_2 dans la phrase S_i et une représentation contextualisée de w_2 est construite de la même manière que pour w_1. En calculant la similarité entre les représentations de w_1 et w_2, nous pouvons évaluer dans quelle mesure w_1 peut être remplacé par w_2 et, selon le principe de substituabilité mentionné ci-dessus, obtenir ainsi une évaluation de leur similarité sémantique.

Cette façon de faire peut d'une certaine façon être vue comme un renversement de la substitution lexicale (McCarthy & Navigli, 2009) : au lieu de choisir un substitut à partir de sa similarité avec un contexte et un mot de référence, un substitut donné permet d'évaluer la similarité avec un mot de référence. Cette évaluation est bien sûr limitée à la phrase sélectionnée mais l'application de cette stratégie à un ensemble de phrases donne une image plus globale de la relation sémantique entre les deux mots. Il est à noter que w_1 et w_2 n'ont pas de rôles symétriques : nous évaluons la similarité de w_2 par rapport à w_1 puisque la représentation de w_2 est construite dans le contexte d'une phrase dans laquelle w_1 apparaît initialement. D'un point de vue pratique, les représentations contextuelles de w_1 et w_2 sont obtenues en encodant la phrase qui les abrite avec le modèle visé et en les extrayant des couches internes de ce modèle, avec donc une représentation par couche. Dans la suite de l'article, les termes *mot source*, *mot cible* et *phrase test* feront référence respectivement à w_1, w_2 et S_i.

2.2 Étude sémantique des plongements contextuels

La section précédente donne un principe pour évaluer la similarité d'un mot cible vis-à-vis d'une occurrence d'un mot source dans le contexte d'une phrase. Nous montrons maintenant comment ce principe peut être appliqué pour explorer les relations sémantiques sous-jacentes à des modèles

contextuels tels qu'ELMo ou BERT. Plus précisément, l'idée est, pour chaque mot d'un ensemble de mots sources, de rassembler un ensemble de mots cibles de telle sorte que chaque paire (mot source, mot cible) soit liée par une relation sémantique de type connu. À partir d'une phrase test contenant une occurrence du mot source, ses mots cibles peuvent être classés en fonction de leur valeur de similarité avec le mot source selon le principe de substituabilité décrit précédemment. Le classement qui en résulte ordonne indirectement les types de relations sémantiques associées aux mots cibles et donne ainsi des indications sur les propriétés sémantiques d'ELMo ou de BERT.

Dans ce travail, nous nous intéressons plus particulièrement aux relations paradigmatiques, plus précisément la synonymie [SYN], l'hyperonymie [HYPE], l'hyponymie [HYPO] et la cohyponymie [COHYP], et nous adoptons WordNet 3.0 comme ressource de référence pour ces types de relations. Cependant, WordNet (Miller, 1990) est fondé sur la notion de synset alors que les phrases contiennent des mots et non des sens de mots. Pour contourner cette difficulté, nous utilisons comme phrases test des phrases du corpus SemCor (Miller *et al.*, 1993), un sous-ensemble du Corpus Brown dont les mots de classe ouverte sont étiquetés avec des synsets de WordNet. Ainsi, le sens d'une occurrence d'un mot source est connu et la relation avec le mot cible dépend de ce sens, ce qui rend l'utilisation de la substitution particulièrement précise. Par exemple, le deuxième sens du mot source *disaster* (*an event resulting in great loss and misfortune*) a, parmi d'autres, la phrase test suivante dans le corpus SemCor :

[1] Since the 1946 **disaster** there have been 15 tsunami in the Pacific, but only one was of any consequence.

Ce sens du mot *disaster* a des mots tels que *cataclysm* ou *catastrophe* comme synonymes, *misfortune* comme hyperonyme, *tsunami* or *meltdown* comme hyponymes et *adversity* or *misadventure* comme cohyponymes. Pour évaluer dans quelle mesure ELMo, par exemple, est plus orienté vers la synonymie que l'hyperonymie, la phrase test [1] est transformée pour donner les phrases [2] et [3] en remplaçant le mot source par un synonyme ou un hyperonyme.

[2] Since the 1946 **catastrophe** there have been 15 tsunami in the Pacific, but only one was of any consequence.
[3] Since the 1946 **misfortune** there have been 15 tsunami in the Pacific, but only one was of any consequence.

Les trois phrases sont encodées à l'aide d'ELMo et la représentation du mot source dans la phrase [1] et des deux mots cibles dans les phrases [2] et [3] sont extraites des couches internes d'ELMo. Comme ELMo comporte trois couches – une couche non contextuelle d'entrée, la couche 0, et deux couches contextuelles, les couches 1 et 2 – nous obtenons trois représentations pour chaque mot. Dans cet exemple et pour la couche 1, la similarité, évaluée classiquement par la mesure *cosinus* appliquée entre la représentation du mot source et celle de la cible synonyme *catastrophe* est égale à 0,89 alors que la similarité du mot source et de la cible hyperonyme *misfortune* n'est égale qu'à 0,55, ce qui donne dans ce cas un net avantage à la synonymie par rapport l'hyperonymie. Le processus est le même pour BERT, sauf que nous avons 12 couches (de la couche 1 à la couche 12) dans ce cas. Une image globale des propriétés sémantiques d'ELMo ou de BERT est obtenue en considérant un nombre significatif de mots sources et de mots cibles dans le contexte d'un grand nombre de phrases test du corpus SemCor. De plus, nous enrichissons cette image en considérant les relations entre les plongements d'ELMo et de BERT et les plongements de mots statiques plus classiques en ajoutant comme cibles DIST_NGH, les voisins les plus similaires aux mots cibles de notre étude obtenus grâce à la mesure *cosinus* sur la base d'un modèle Skip-gram entraîné à partir d'un grand corpus.

Cette étude des propriétés sémantiques d'ELMo et de BERT peut être considérée comme une sorte de sondage sémantique et être liée en tant que telle aux travaux de Schick & Schütze (2020) et à

leur méthode *WordNet Language Model Probing*. Néanmoins, leur objectif global est très différent du nôtre et leur tâche de sondage, fondée sur une notion de patron, est plus adaptée aux relations syntagmatiques qu'aux relations paradigmatiques.

2.3 Plongements contextuels versus plongements statiques

Outre l'étude des propriétés sémantiques des plongements contextuels, la méthode que nous avons présentée peut également être adaptée pour étudier les différences entre ces plongements et les plongements statiques du point de vue de ces mêmes propriétés. Plus précisément, l'idée est de définir les cibles en remplaçant les relations sémantiques de WordNet par des relations caractérisant les plongements statiques. En raison de la nature distributionnelle de ces plongements, nous avons choisi d'associer comme cibles à chaque clé, correspondant à un mot, ses voisins distributionnels les plus similaires selon les plongements statiques considérés. Comme précédemment, les cibles sont ordonnées en fonction d'un ensemble de phrases test par les plongements contextuels étudiés et nous utilisons les relations paradigmatiques de WordNet pour déterminer a posteriori les types de relations que les plongements contextuels favorisent par rapport aux plongements statiques. Puisque les clés sont des mots et non des sens de mots dans ce protocole, nous agrégeons les représentations contextuelles des clés et des cibles construites à partir des phrases test en les moyennant, comme recommandé par Bommasani *et al.* (2020), pour obtenir des représentations au niveau des mots.

3 Expérimentations

3.1 Étude sémantique des plongements contextuels

3.1.1 Cadre d'expérimentation

La mise en œuvre des principes présentés précédemment est associée à certains choix concernant les phrases test et les relations sémantiques entre les mots sources et les mots cibles. Notre premier choix a été de nous concentrer sur les noms. Le nombre de mots cibles pour chaque mot source a été fixé à 40 pour avoir des résultats comparables pour tous les mots sources. De plus, nous n'avons retenu que les mots sources ayant des mots cibles pour les cinq types de relations que nous considérons, une fois encore pour l'homogénéité des résultats entre mots sources. Nous avons également limité le nombre de mots cibles pour chaque type de relations à 10, avec une limite supplémentaire à 30 pour l'ensemble des mots cibles issus de relations de WordNet. En pratique, cette limite ne concerne que les relations d'hyponymie et de cohyponymie, dont le nombre tend à être élevé. La première colonne du tableau 1 donne le nombre moyen de mots cibles de chaque mot source en fonction de leur type de relations sémantiques.

Par ailleurs, nous avons adopté les définitions suivantes pour chaque type de cibles issues de relations de WordNet :

- synonymes [SYN] : tous les mots de $Synset_{srce}$, le synset correspondant au sens du mot source présent dans la phrase test considérée ;
- hyperonymes [HYPE] : tous les mots des synsets $Synset_{hype}$ ayant un lien direct d'hyperonymie avec $Synset_{srce}$;

	#cibles	aléatoire	ELMo			BERT					
			L_0	L_1	L_2	L_1	L_3	L_5	L_8	L_{10}	L_{12}
SYN	2,1	7,4	30,9	29,3	26,9	33,5	34,5	36,1	**36,7**	36,0	35,1
HYPE	1,9	6,7	4,3	6,1	6,2	7,8	9,3	9,9	10,9	11,1	**11,2**
HYPO	5,9	20,9	11,4	13,8	**14,8**	10,5	11,1	11,7	12,3	11,9	12,3
COHYP	8,3	29,4	6,7	7,9	**9,6**	6,4	6,7	7,5	7,7	7,8	7,9
DIST_NGH	10	35,5	**46,6**	42,9	42,5	41,7	38,4	34,7	32,3	33,1	33,5

TABLE 1 – P@1 ($\times$100) pour l'ordonnancement, fondé sur les phrases du corpus SemCor, des mots cibles associés à un ensemble de mots sources

— hyponymes [HYPO] : tous les mots des synsets ayant une relation directe d'hyponymie avec $Synset_{srce}$;
— cohyponymes [COHYP] : tous les mots des synsets, à l'exception de $Synset_{srce}$, ayant une relation directe d'hyponymie avec les synsets $Synset_{hype}$.

Comme mentionné précédemment, les mots cibles DIST_NGH sont obtenus par un modèle Skip-gram, entraîné sur un sous-ensemble d'1 milliard de mots du corpus anglais annoté Gigaword (Napoles *et al.*, 2012) avec les meilleures valeurs d'hyperparamètres de (Baroni *et al.*, 2014) [1]. Plus précisément, nous utilisons ce modèle pour sélectionner les 10 premiers voisins distributionnels de chaque mot source, parmi un vocabulaire de 20 813 noms, ne correspondant pas à un mot cible issu d'une relation de WordNet. Dans ce cas, cette sélection n'est pas faite selon le sens du mot cible dans une phrase test puisque nous n'avons pas accès aux sens des mots.

En ce qui concerne les phrases test, une limite supérieure de 20 phrases a également été fixée pour chaque mot source, toujours pour avoir des résultats comparables entre mots sources. Finalement, notre évaluation s'appuie sur 41 079 phrases du corpus SemCor, ce qui représente environ 4,5 phrases par sens en moyenne et 7,9 phrases pour chacun des 5 241 mots sources. Ces derniers couvrent un large spectre de fréquences puisque si nous nous référons à la sous-partie que nous utilisons du corpus Gigaword, le mot source le plus fréquent, *year*, a 2 991 899 occurrences alors que les mots sources les moins fréquents, comme *inadvertence*, n'ont que 22 occurrences.

3.1.2 Évaluation

Le résultat du classement des mots cibles sélectionnés pour tous nos mots sources et toutes nos phrases test est présenté dans le tableau 1 pour différentes couches d'ELMo et de BERT (L_x) et chaque type de mots cibles. De plus, la deuxième colonne fournit les valeurs de précision au rang 1 (P@1) d'un classement aléatoire selon la distribution de la première colonne. P@1 dans ce contexte correspond à la proportion de phrases test qui classent en premier un mot cible lié à un mot source avec un type de relations spécifique (un par ligne).

Ce tableau montre en premier lieu que les différentes couches d'ELMo et de BERT n'ont pas exactement les mêmes propriétés sémantiques, ce qui confirme les conclusions d'Ethayarajh (2019) ou de Wu *et al.* (2020). Il montre également certaines similitudes et différences entre ELMo et BERT. La différence la plus évidente est que BERT obtient des résultats beaucoup plus élevés

1. Fréquence minimale=5, taille vecteurs=300, taille fenêtre=5, 10 exemples négatifs et 10^{-5} pour le sous-échantillonnage des mots les plus fréquents.

qu'ELMo pour les synonymes et les hyperonymes, mais des résultats un peu plus faibles pour les hyponymes et les cohyponymes. BERT favorise donc la similarité sémantique plutôt que la proximité sémantique au sens de Budanitsky & Hirst (2006). On peut en effet considérer que les hyponymes et les cohyponymes, malgré leur nature paradigmatique, sont moins représentatifs, en termes de similarité sémantique, que les synonymes et les hyperonymes. La tendance est encore plus évidente pour les relations DIST_NGH provenant des plongements statiques, qui sont davantage susceptibles d'être de nature syntagmatique.

ELMo et BERT présentent également quelques différences quant aux propriétés de leurs couches respectives. Alors que la capacité d'ELMo à classer au rang 1 les synonymes diminue régulièrement à mesure que l'on considère des couches de plus en plus hautes, elle augmente d'abord jusqu'à la couche 8 pour BERT, puis tend à diminuer. Cette observation doit également être mise en relation avec les résultats d'Ethayarajh (2019), qui a observé que les représentations des mots sont plus contextualisées à mesure que le niveau de leurs couches augmente. Du point de vue d'ELMo, cette plus grande contextualisation des représentations tendrait donc à favoriser la proximité sémantique au détriment de la similarité sémantique. Ce n'est pas complètement surprenant dans la mesure où, du point de vue sémantique, la contextualisation est plus susceptible de conduire à des relations syntagmatiques que paradigmatiques. Ce résultat est moins évident pour BERT puisque ses premières couches ont la meilleure précision pour les relations DIST_NGH mais un changement plus compatible avec les résultats d'ELMo se produit après la couche 8 concernant le classement des synonymes et des relations DIST_NGH.

Cependant, ELMo et BERT ont également de fortes convergences. En premier lieu, l'effet le plus significatif du classement des mots cibles par ELMo et BERT est obtenu pour les synonymes. P@1 est jusqu'à quatre fois plus élevée pour la meilleure couche d'ELMO et cinq fois pour la meilleure couche de BERT par rapport au classement aléatoire initial. L'application de l'hypothèse distributionnelle telle que décrite à la section 2.1 est donc une méthode intéressante pour identifier les synonymes d'un mot parmi une liste de ses voisins distributionnels. Plus globalement, si certaines différences peuvent être notées entre les couches de ces modèles en termes d'orientation sémantique, elles ont toutes un fort penchant pour la synonymie. Nous pouvons également observer que tant ELMo que BERT ont une forte corrélation inverse entre leur P@1 pour les synonymes et leur P@1 pour les relations DIST_NGH : lorsqu'une couche tend à favoriser une stricte similarité sémantique, elle obtient logiquement des résultats moins bons pour la proximité sémantique. Cependant, cette corrélation ne conduit pas à des résultats pour les relations DIST_NGH beaucoup plus faibles que les résultats d'un classement aléatoire, en particulier pour ELMo, ce qui suggère qu'ELMo et BERT ne sont pas radicalement différents des plongements statiques du point de vue de la similarité sémantique qu'ils véhiculent.

Un examen plus approfondi des mot cibles DIST_NGH classés en premier montre également qu'ils ont souvent une forte relation sémantique avec le mot source. Dans certains cas, il est l'un de ses synonymes mais pour un sens différent qui est proche du sens de l'occurrence courante. Par exemple, dans la phrase :

Land reform programs need to be supplemented with programs for promoting rural credits [...] in **agriculture**.

le mot classé en premier par la première couche d'ELMo, *farming*, est synonyme du deuxième sens de *agriculture – the practice of cultivating the land or raising stock* – alors que cette occurrence est étiquetée dans le SemCor avec son premier sens – *a large-scale farming enterprise*. Dans d'autres cas, le mot cible de rang 1 est en fait un synonyme du mot source mais n'est pas considéré comme tel dans WordNet. Par exemple, le mot cible *capability* pour le mot source *ability*. Enfin, il est également

	#relations		ELMo			BERT					
	référence	Skip-gram	L_0	L_1	L_2	L_1	L_3	L_5	L_8	L_{10}	L_{12}
SYN	3,5	18,6	22,6	**23,0**	21,4	20,1	20,6	20,9	21,4	21,3	21,9
HYPE	5,0	6,9	5,8	6,6	6,3	8,1	8,3	8,6	8,9	**9,0**	8,5
HYPO	10,8	11,8	13,3	**14,2**	13,7	10,5	10,8	11,3	11,5	10,9	10,2
COHYP	56,3	27,9	33,5	**34,8**	33,0	30,0	30,7	31,2	31,3	31,2	31,6

TABLE 2 – P@1 ($\times 100$) pour l'ordonnancement des voisins distributionnels du modèle Skip-gram par les plongements contextuels

fréquent de trouver comme mot cible de rang 1 un antonyme du mot source, comme *inaction* pour le mot source *action*. Bien que ce ne soit pas un résultat attendu du point de vue de notre référence, les antonymes sont connus pour être très similaires aux synonymes sur le plan distributionnel et leur présence confirme donc la tendance observée à favoriser la similarité sémantique.

3.2 Plongements contextuels versus plongements statiques

3.2.1 Cadre d'expérimentation

La principale différence avec l'expérience précédente concerne les phrases test. Elles sont ici sélectionnées aléatoirement dans une sous-partie du corpus utilisé pour l'entraînement des plongements statiques et les clés n'y sont pas sémantiquement désambiguïsées. Plus précisément, nous avons sélectionné 10 phrases test pour chaque clé, d'une taille comprise entre 10 et 90 mots pour avoir un contexte significatif mais ciblé. Pour les plongements statiques, nous nous appuyons sur le même modèle Skip-gram que celui utilisé pour extraire les relations DIST_NGH. Le processus d'ordonnancement est appliqué aux mêmes 5 241 clés que dans la première expérience et se concentre sur leurs 10 premiers voisins distributionnels, résultant de l'application de la mesure cosinus. Comme dans (Piasecki *et al.*, 2018), nos voisins de référence sont obtenus en extrayant de WordNet les mots liés à une clé par les mêmes types de relations qu'à la section 3.1. Cependant, le nombre de relations est ici plus important puisque nous considérons tous les synsets dans lesquels la clé est présente du fait de l'absence de désambiguïsation sémantique des clés. La mesure utilisée est comme précédemment la précision au premier rang (P@1) des voisins reclassés pour chaque type de relations.

3.2.2 Évaluation

Le tableau 2 compare les voisins distributionnels ordonnés par le modèle Skip-gram avec leur ordonnancement par ELMo et BERT selon la méthodologie que nous proposons. La deuxième colonne de ce tableau donne le nombre moyen de relations du type considéré par clé dans WordNet, c'est-à-dire la richesse de la référence pour l'évaluation.

La première observation est que si ces plongements contextuels favorisent significativement [2] les synonymes et les cohyponymes par rapport à Skip-gram, la situation est plus complexe pour les hyperonymes et les hyponymes : ELMo dégrade les résultats de Skip-gram pour les hyperonymes mais les améliore pour les hyponymes alors que BERT fait exactement le contraire.

2. Les différences sont jugées significatives selon un test de Wilcoxon apparié si $p \leq 0,01$.

On peut également noter que le modèle Skip-gram favorise largement la cohyponymie par rapport aux autres relations lexicales, une tendance qui se trouve globalement accentuée par ELMo et BERT, avec une plus grande amplitude dans le cas d'ELMo. De ce point de vue, il est intéressant de noter que dans le tableau 1, l'ordonnancement des cohyponymes par ELMo et BERT est au contraire très nettement inférieur à celui obtenu par un ordonnancement aléatoire. Cette différence entre les deux expériences illustre la nature contextuelle de ces modèles mais aussi la limite de cette contextualisation. Dans notre première expérience, les clés sont sémantiquement désambiguïsées dans les phrases test et les cibles sont liées au sens de ces clés désambiguïsées. Dans une telle situation, un modèle contextuel peut favoriser les cibles les plus liées sémantiquement à leur clé, comme les synonymes, car il produit des représentations spécifiques au contexte considéré, qui est lié au sens de la clé. Au contraire, dans notre seconde expérience, les cibles couvrent tous les sens de la clé et dans une telle configuration, les plongements contextuels favorisent la proximité sémantique plutôt que la similarité sémantique et sont ainsi plus proches des plongements statiques, même s'ils tendent globalement à les surpasser.

Cet effet a un impact beaucoup plus important pour les synonymes dans le cas de BERT que pour ELMo, ce qui résulte probablement d'une plus grande sensibilité de BERT au contexte qu'ELMo. Cependant, il ne modifie pas la configuration des résultats entre les différentes couches de BERT, avec de meilleurs résultats autour de la couche 8, alors que les meilleurs résultats d'ELMo sont assez étonnamment obtenus par une couche plus contextualisée que précédemment.

4 Conclusion et perspectives

Dans cet article, nous avons étudié comment le principe distributionnel de substitution peut être utilisé comme une forme de sonde pour tester le type de similarité sémantique véhiculé par ELMo et BERT. Des expériences menées avec les relations paradigmatiques de WordNet et le corpus SemCor au niveau des sens des mots ont montré que la nature contextuelle de ces modèles favorise clairement la synonymie en termes de relations lexicales. Dans un second temps, nous avons adapté la même méthode pour étudier les différences entre les plongements contextuels et statiques, avec la conclusion que le biais des plongements contextuels en faveur de la similarité sémantique observé au niveau des occurrences de mots est fortement réduit au niveau des mots hors contexte par rapport aux plongements statiques.

Une des extensions les plus directes de ce travail est l'élargissement de la méthode de comparaison des plongements contextuels et statiques au problème de l'amélioration des thésaurus distributionnels par le biais du réordonnancement de leurs voisins. L'ordonnancement des voisins distributionnels n'est vu dans le travail présenté que comme une façon de caractériser les différences sémantiques entre plongements contextuels et statiques. Mais il peut aussi être envisagé sous l'angle d'une amélioration des thésaurus distributionnels construits à partir de plongements statiques en mettant en avant les voisins distributionnels les plus sémantiquement liés à leurs entrées. Dans cette optique, une étude des meilleures options pour l'agrégation au niveau des mots hors contexte des informations fournies au niveau des occurrences de mots par les plongements contextuels serait d'un grand intérêt.

Remerciements

Le travail présenté dans cet article a été réalisé dans le cadre du projet ADDICTE [3] (Analyse distributionnelle en domaine spécialisé), financé par l'Agence Nationale de la Recherche (ANR-17-CE23-0001).

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.

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 : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.431.

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

CHRONIS G. & ERK K. (2020). When is a bishop not like a rook ? When it's like a rabbi ! Multi-prototype BERT embeddings for estimating semantic relationships. In *24th Conference on Computational Natural Language Learning (CoNLL 2020)*, p. 227–244, Online : Association for Computational Linguistics.

COENEN A., REIF E., YUAN A., KIM B., PEARCE A., VIÉGAS F. & WATTENBERG M. (2019). Visualizing and Measuring the Geometry of BERT. In H. WALLACH, H. LAROCHELLE, A. BEYGELZIMER, F. D'ALCHÉ BUC, E. FOX & R. GARNETT, Éds., *Advances in Neural Information Processing Systems*, volume 32, p. 8594–8603 : Curran Associates, Inc.

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 : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

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 : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1006.

ETTINGER A. (2020). What BERT Is Not : Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. *Transactions of the Association for Computational Linguistics*, **8**, 34–48. DOI : 10.1162/tacl_a_00298.

GARÍ SOLER A., APIDIANAKI M. & ALLAUZEN A. (2019). Word Usage Similarity Estimation with Sentence Representations and Automatic Substitutes. In *Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, p. 9–21, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/S19-1002.

3. https://anr-addicte.ls2n.fr/

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

MCCARTHY D. & NAVIGLI R. (2009). The English lexical substitution task. *Language resources and evaluation*, **43**(2), 139–159.

MICKUS T., CONSTANT M., PAPERNO D. & VAN DEEMTER K. (2020). What do you mean, BERT ? Assessing BERT as a Distributional Semantics Model. *Society for Computation in Linguistics*, **3**. DOI : 10.7275/t778-ja71.

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

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

MILLER G. A., LEACOCK C., TENGI R. & BUNKER R. T. (1993). A Semantic Concordance. In *Human Language Technology*, Plainsboro, USA.

NAPOLES C., GORMLEY M. R. & VAN DURME B. (2012). Annotated Gigaword. In *NAACL Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX)*, p. 95–100, Montréal, Canada.

PETERS M., NEUMANN M., IYYER M., GARDNER M., CLARK C., LEE K. & ZETTLEMOYER L. (2018). Deep Contextualized Word Representations. In *2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (NAACL-HLT 2018)*, p. 2227–2237, New Orleans, Louisiana, USA : Association for Computational Linguistics.

PIASECKI M., CZACHOR G., JANZ A., KASZEWSKI D. & KEDZIA P. (2018). Wordnet-based evaluation of large distributional models for polish. In *9th Global WordNet Conference (GWC 2018)*.

ROGERS A., KOVALEVA O. & RUMSHISKY A. (2020). A Primer in BERTology : What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, **8**, 842–866. DOI : 10.1162/tacl_a_00349.

SCHICK T. & SCHÜTZE H. (2020). Rare Words : A Major Problem for Contextualized Embeddings and How to Fix It by Attentive Mimicking. In *Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, p. 8766–8774, New York, USA.

TENNEY I., XIA P., CHEN B., WANG A., POLIAK A., MCCOY R. T., KIM N., DURME B. V., BOWMAN S., DAS D. & PAVLICK E. (2019). What do you learn from context ? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

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 : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.586.

WU J., BELINKOV Y., SAJJAD H., DURRANI N., DALVI F. & GLASS J. (2020). Similarity Analysis of Contextual Word Representation Models. In *58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, p. 4638–4655, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.422.