Utiliser un modèle neuronal générique pour la substitution lexicale

Olivier Ferret
CEA, LIST, Laboratoire Vision et Ingénierie des Contenus, Gif-sur-Yvette, F-91191 France.
olivier.ferret@cea.fr

Résumé. Dans cet article, nous présentons la participation du laboratoire LVIC du CEA LIST à la tâche de substitution lexicale organisée dans le cadre de l'atelier SemDis 2014. Le travail réalisé s'appuie sur une exploitation très simple du modèle neuronal proposé dans (Mikolov *et al.*, 2013b), qui a montré par ailleurs son intérêt pour différentes tâches ayant trait à la similarité sémantique en anglais. Nous analysons de ce point de vue les capacités de l'instanciation de ce modèle que nous avons construit pour le français. L'article présente également l'impact de l'utilisation de différents types de ressources pour la génération des candidats substituts.

Abstract. In this article, we present the participation of the CEA LIST LVIC laboratory to the lexical substitution task of the SemDis 2014 workshop. This work is based on the neural model proposed by (Mikolov *et al.*, 2013b), which has shown good results on various tasks related to semantic similarity in English. We have used this model in a very simple way for performing lexical substitution in French and more particularly for the contextual selection of a lexical substitute among several candidates. The article also investigates the use of various types of resources for generating the substitution candidates.

Mots-clés : Substitution lexicale, réseau de neurones, représentations lexicales distribuées.

Keywords: Lexical substitution, neural network, distributed lexical representations.

1 Introduction

La problématique de la représentation distributionnelle du sens des mots ou d'unités plus larges telles que des mots composés ou même des phrases fait l'objet actuellement d'un large ensemble de travaux, en particulier du point de vue de la compositionnalité de ces représentations (Grefenstette *et al.*, 2013). Dans ce contexte, la problématique de la similarité sémantique est centrale dans la mesure où déterminer si deux unités linguistiques différentes, quelle que soit leur taille, ont des sens équivalents ou proches est l'opération de base pour tester la validité des représentations élaborées pour représenter leur sens. L'évaluation de cette similarité sémantique a donc elle-même été un objet de préoccupation depuis longtemps, avec une focalisation toute particulière sur les évaluations de nature intrinsèque. Ces dernières se sont pour l'essentiel réparties entre les évaluations prenant comme référence un dictionnaire ou un thésaurus construit manuellement, comme dans (Curran & Moens, 2002) ou (Ferret, 2010), et celles déterminant le degré de corrélation existant entre les similarités obtenues automatiquement et un ensemble de jugements de similarité réalisés par des humains (Gabrilovich, 2007; Bruni *et al.*, 2012).

Les évaluations extrinsèques, c'est-à-dire opérées par le biais d'une tâche exploitant les similarités calculées, sont quant à elles plus rares. Si les thésaurus distributionnels sont utilisés dans un nombre croissant de tâches, allant de l'extraction de relations (Min *et al.*, 2012) à la traduction automatique (Marton, 2013) en passant par l'analyse syntaxique (Henestroza Anguiano & Candito, 2012) ou l'analyse d'opinion (Goyal & Daume, 2011), il est en effet peu fréquent de voir une analyse des différents paramètres propres à une mesure de similarité distributionnelle au travers de leur impact sur une tâche exploitant une telle mesure. Outre que la mise en œuvre de ce type d'évaluation est plus complexe que celle d'une évaluation intrinsèque, une explication possible de cet état de fait est la possible difficulté à montrer des effets significatifs, au niveau de l'évaluation de la tâche cible, de différences existant au niveau de la mesure distributionnelle utilisée. Il faut pour cela que la mesure de similarité y occupe un rôle suffisamment central, à l'instar par exemple d'une tâche comme l'expansion de listes de mots ou d'entités (Pantel *et al.*, 2009).

De ce point de vue, la tâche de substitution lexicale offre un intérêt particulier. En effet, même si elle s'inscrit historique-

ment plutôt dans le contexte de la désambiguïsation sémantique (McCarthy & Navigli, 2009), elle est étroitement liée, de par sa définition intrinsèque, à la notion de similarité sémantique puisque le substitut à trouver doit être sémantiquement similaire au mot cible à remplacer. D'autre part, elle vient introduire une dimension manquante aux évaluations intrinsèques habituellement pratiquées : la dimension du contexte. Dans les évaluations intrinsèques évoquées plus haut, les mots similaires à un mot cible sont trouvés en l'absence de tout contexte et les ressources de référence ne sont elles-mêmes pas liées à un contexte particulier. En reprenant les principes de (Gabrilovich, 2007), (Huang *et al.*, 2012) a proposé un jeu de test introduisant la notion de contexte : au lieu de demander à des sujets de juger du degré de similarité de couples de mots en dehors de tout contexte, ce jugement était demandé pour des mots faisant partie d'une même phrase, les mots étant présentés dans le contexte de cette phrase. Curieusement néanmoins, (Huang *et al.*, 2012) n'a pas proposé de façon parallèle de modèle de similarité tenant compte du contexte[1]. (Dinu & Lapata, 2010) s'est en revanche intéressé plus directement à la prise en compte du contexte dans le calcul des similarités sémantiques et a en particulier utilisé le cadre d'évaluation fourni par la tâche *English Lexical Substitution* de l'évaluation SemEval 2007 pour tester ses propositions.

La tâche de substitution lexicale vient donc apporter une double dimension aux évaluations relatives à la similarité sémantique : d'une part, elle représente une évaluation extrinsèque au sein de laquelle la similarité sémantique à un rôle suffisamment influent pour que des différences la concernant puissent être observées en termes d'impact sur les résultats de la tâche ; d'autre part, elle apporte une dimension contextuelle à ce type d'évaluation. C'est ce double intérêt qui nous a conduit à participer à l'évaluation SemDis 2014 relative à la substitution lexicale en français. Cette participation avait également pour ambition de tester l'intérêt de la mise en œuvre d'un modèle neuronal dans ce contexte et la possibilité de réaliser celle-ci avec un effort limité, estimé à 2 personnes-jours, en se fondant sur la transposition d'une approche de base définie dans (Zweig *et al.*, 2012). Nous décrirons plus en détail cette approche dans la section suivante tandis que la section 3 sera consacrée à une présentation plus fouillée et à une évaluation des ressources utilisées ou construites. La section 4 présentera enfin les résultats de l'évaluation de nos différentes soumissions.

2 Méthode

2.1 Contexte

Les modèles de langage neuronaux font l'objet depuis leur résurgence, il y a quelques temps (Bengio *et al.*, 2003), d'un nombre important de travaux ayant montré leur intérêt dans un large ensemble de tâches allant, en se restreignant aux données langagières, de la reconnaissance de la parole à l'analyse de sentiments en passant par la traduction automatique. Plus spécifiquement, au-delà des tâches reposant de façon importante sur des modèles de langage, l'approche présentée dans (Collobert *et al.*, 2011) a mis en avant la capacité de ces modèles neuronaux à produire des représentations distribuées des mots (*word embeddings*) pouvant être utilisées comme traits dans les classifieurs à la base d'une partie importante des systèmes de traitement automatique des langues.

Ces travaux ont également montré que de telles représentations peuvent être construites en dehors du contexte particulier d'une tâche pour être exploitées de façon assez générique avec profit. En première analyse, une part importante du succès de ce type d'approches repose sur le fait que les représentations construites capturent une forme de proximité lexicale : deux mots dont les représentations distribuées sont proches ont également tendance à entretenir une forme de proximité. La nature de cette proximité est en revanche difficile à établir car elle semble mêler dimensions paradigmatique et syntagmatique, à la fois sur un plan sémantique et syntaxique. Huang *et al.* (2012) ont cherché à caractériser plus précisément cette forme de similarité en confrontant différents types de représentations lexicales distribuées issues de modèles neuronaux à un test classique d'évaluation de la similarité sémantique, en l'occurrence WordSim 353 (Gabrilovich, 2007). Sans dépasser le niveau moyen de l'état de l'art[2], les résultats obtenus suggèrent que cette similarité comporte une composante sémantique significative. Ce constat s'est trouvé renforcé par les travaux présentés dans (Mikolov *et al.*, 2013b), qui reposent sur un autre type de modèle neuronal et s'évaluent sur une tâche, non de similarité sémantique, mais de détection de relations d'analogie. Néanmoins, Mikolov *et al.* (2013a) montrent que l'application du même type de modèle sur les données du *Microsoft Sentence Completion Challenge* (Zweig *et al.*, 2012), qui est proche de la tâche de substitution lexicale même s'il est plus orienté vers les modèles de langage, donne de très bons résultats. Tout récemment, l'évaluation plus systématique de ce même modèle du point de vue de la similarité sémantique semble montrer ses bonnes performances également dans ce domaine (Baroni *et al.*, 2014 to appear), en particulier par rapport à des approches distributionnelles

[1] Le modèle neuronal proposé pour construire la représentation lexicale distribuée (*word embedding*) sur laquelle l'évaluation de la similarité entre mots se fonde intègre deux niveaux de contexte mais l'évaluation de la similarité en tant que telle reste acontextuelle.

[2] Certains modèles se situant très en dessous.

plus classiques.

Malgré le fait que des évaluations menées de façon différente, en l'occurrence dans le contexte de la construction de thésaurus distributionnels, montrent que le modèle présenté dans (Mikolov *et al.*, 2013b) n'obtient pas de meilleurs résultats qu'une approche distributionnelle classique (Ferret, 2014 à paraître)[3], même s'il dépasse celui de (Huang *et al.*, 2012), nous avons choisi de l'appliquer à la tâche de substitution lexicale de SemDis 2014. Il est à noter d'ailleurs que l'application d'un modèle neuronal à ce type de tâche n'est pas complètement inédite. Outre le cas de la tâche de complétion de phrase cité précédemment (Mikolov *et al.*, 2013a), Glickman *et al.* (2006) ont réalisé une telle application dans le contexte spécifique de la génération de sous-titre. Dans ce cas précis, le modèle neuronal était utilisé pour évaluer la vraisemblance d'un substitut possible, compte tenu de son contexte. Le résultat était néanmoins moins intéressant que celui d'approches non contextuelles fondées sur des ressources constituées *a priori*.

2.2 Description de l'approche

Outre l'intérêt de tester une approche encore assez nouvelle pour une tâche connue, l'utilisation d'un modèle neuronal pour une tâche de substitution lexicale se justifie par la nature même de la tâche. Celle-ci peut en effet se décomposer en deux sous-tâches principales, plus ou moins jointes selon la solution adoptée pour résoudre le problème :

– la génération de substituts possibles pour le mot cible à remplacer ;
– le choix d'un des substituts générés en fonction du contexte du mot cible à remplacer.

La première sous-tâche renvoie assez clairement à une dimension paradigmatique. Il s'agit en effet de trouver des synonymes du mot cible. La seconde sous-tâche est en revanche moins univoque quant au type de similarité sémantique qu'elle met en jeu : le substitut doit en principe entretenir avec les mots qui l'environnent le même type de relations que le mot cible originel entretient avec ces mêmes mots. Néanmoins, ces relations ne sont pas des relations d'équivalence et entrent en pratique dans la catégorie des relations dites « non classiques » (Morris & Hirst, 2004). Ces relations ont de plus un caractère local dans la mesure où les mots pris en compte ne dépassent pas l'espace de la phrase. À ce titre, elles portent également une certaine dimension syntaxique, à l'instar d'ailleurs des modèles de ngrammes exploités de façon très majoritaire dans les systèmes existants de substitution lexicale (McCarthy & Navigli, 2009). De par son caractère composite, que nous avons esquissé précédemment, la notion de similarité portée par les représentations lexicales distribuées des modèles neuronaux est donc un candidat intéressant pour capturer ces relations « non classiques ».

En pratique, nous avons donc traité les deux sous-tâches en nous appuyant sur des méthodes et des ressources différentes du point de vue sémantique. La génération des substituts a ainsi été réalisée par une simple recherche dans des dictionnaires existant de synonymes et de mots liés associés au mot cible considéré. De ce point de vue, trois types de dictionnaires ont été testés : deux dictionnaires construits manuellement et un thésaurus distributionnel construit automatiquement. Parmi les premiers, l'un contenait un nombre très limité de synonymes pour chaque entrée tandis que l'autre fournissait un ensemble plus large, à la fois en termes de quantité et de lien sémantique, de synonymes et mots liés.

Pour le choix parmi les substituts générés, nous avons transposé une approche de base proposée dans (Zweig *et al.*, 2012). Le principe de cette approche est simple : une mesure de similarité sémantique est calculée entre chaque candidat substitut et l'ensemble des mots pleins de la phrase contenant le mot cible à remplacer, hors ce dernier. Le substitut retenu est le mot dont la somme des valeurs de similarité ainsi obtenues est la plus élevée. L'approche étant non supervisée, nous n'avons donc pas exploité les données d'entraînement fournies pour l'évaluation SemDis 2014.

Dans le cas de (Zweig *et al.*, 2012), la mesure de la similarité sémantique entre deux mots reposait sur l'Analyse Sémantique Latente (Landauer & Dumais, 1997), qui permet de construire une représentation distribuée des mots à partir d'un corpus en s'appuyant sur une forme de factorisation de matrice. Nous avons repris le principe général de (Zweig *et al.*, 2012) mais en substituant une représentation distribuée issue d'un modèle neuronal à la représentation distribuée issue de l'Analyse Sémantique Latente pour représenter le sens des mots. Il est à noter que cette approche exploite une notion de similarité sémantique de nature assez générique : bien qu'elle ne soit pas focalisée sur la seule notion d'équivalence sémantique, elle ne fait pas apparaître des types de relations distincts dans l'optique de rendre compte d'une relation spécifique unissant un mot de la phrase avec le mot cible à substituer. En pratique, la représentation d'un mot prenant la forme d'un vecteur, la similarité de deux mots est évaluée en calculant une mesure de similarité vectorielle générique entre les vecteurs représentant ces deux mots. La mesure la plus utilisée dans ce contexte est le cosinus mais nous avons aussi testé, du fait du caractère dense des représentations, une transposition de la distance euclidienne en mesure de similarité :

[3]Les raisons de la divergence entre les évaluations menées dans (Baroni *et al.*, 2014 to appear) et dans (Ferret, 2014 à paraître) restent à éclaircir mais ont peut-être à voir avec le nombre et la nature, en termes de fréquence, des candidats voisins sémantiques testés.

$$sim(m_1, m_2) = \frac{1}{1 + l2(m_1, m_2)} \quad (1)$$

avec $l2(m_1, m_2)$, la distance euclidienne entre les représentations des mots m_1 et m_2.

Conformément aux conclusions de la section précédente, nous avons retenu le modèle neuronal proposé dans (Mikolov *et al.*, 2013a). À l'instar des travaux fondateurs de Bengio *et al.* (2003), ce modèle apprend un modèle de langage ; autrement dit, il apprend à prédire la probabilité d'un mot en fonction de la séquence de mots qui le précèdent. Dans le cas précis du modèle que nous avons utilisé, appelé *Skip-gram*, l'objectif est de maximiser en sortie la probabilité d'un mot présent dans la même phrase qu'un mot placé en entrée du réseau. Une contrainte est de plus fixée sur la distance maximale C séparant le mot en entrée et le mot en sortie à prédire dans les phrases dans lesquelles ils cooccurrent. Concrètement, le modèle prend la forme d'un réseau de neurones à trois couches dont la première couche est constituée de la représentation d'un seul mot et la dernière couche, de R représentations de mots, R correspondant au nombre de mots de l'environnement du mot d'entrée à prédire. Chaque représentation de mot est formée de V neurones, V étant la taille du vocabulaire considéré. La construction des représentations distribuées des mots s'effectue en modifiant les pondérations de liaison à la couche cachée (la deuxième couche) de façon à rapprocher progressivement les prédictions faites par le réseau avec les données effectivement observées dans le corpus utilisé pour construire ces représentations. Les exemples sont donc constitués dans le cas présent de couples de mots présents dans une même phrase, séparés d'au plus C mots.

3 Ressources construites et utilisées

3.1 Modèle neuronal

Le modèle *Skip-gram* tel que défini dans (Mikolov *et al.*, 2013a) est caractérisé par un certain nombre d'optimisations rendant son entraînement particulièrement efficace. Il peut donc être appliqué à de larges corpus. Pour l'évaluation SemDis 2014, nous nous sommes limités à un corpus que l'on peut qualifier de taille moyenne mais qui présentait pour nous l'avantage d'avoir déjà été prétraité[4]. Il s'agit plus précisément du corpus utilisé lors de l'évaluation EQueR des systèmes de question-réponse en français (Ayache *et al.*, 2006). Ce corpus, d'une taille de 258 millions de mots environ, est principalement constitué d'articles du journal *Le Monde* (entre 1992 et 2000), auxquels s'ajoutent des articles issus du *Monde Diplomatique*, des dépêches d'agence SDA et quelques rapports issus du Sénat. Le prétraitement du corpus s'est limité à son étiquetage et sa lemmatisation par l'outil TreeTagger (Schmid, 1994) et à l'adaptation au format d'entrée de l'outil *word2vec*[5], qui a réalisé la construction des représentations lexicales distribuées selon le modèle *Skip-gram*. Nous avons en outre éliminé tous les signes de ponctuation mais conservé tous les mots[6], en joignant leur lemme et leur catégorie morpho-syntaxique. En accord avec les expérimentations rapportées dans (Mikolov *et al.*, 2013a), en particulier celles concernant le *Microsoft Sentence Completion Challenge*, nous avons adopté une valeur de 10 pour le paramètre C et un nombre de dimensions pour les représentations distribuées égal à 600.

3.2 Ressources de génération des substituts

Comme nous l'avons précisé à la section 2.2, nous avons fait appel à trois ressources aux caractéristiques complémentaires pour la génération des substituts. Ces trois ressources sont :

– une extraction du dictionnaire de synonymes de Word XP (noté *word*) : ce dictionnaire est composé de 31 007 entrées, dont 7 068 adjectifs, 5 781 verbes et 16 848 noms. Il comporte assez peu de synonymes associés à chaque entrée (cf. $4^{ème}$ colonne du tableau 1) ;

[4] Dans le cadre de l'évaluation, il aurait été judicieux de réaliser la construction du modèle à partir du corpus frWaC dont étaient issues les données d'évaluation.

[5] `http://code.google.com/p/word2vec`

[6] La construction des représentations distribuées s'inscrivant dans la perspective des modèle de langage, il est raisonnable de penser que sélectionner les mots en fonction de leur catégorie morpho-syntaxique ou de leur fréquence peut avoir un impact négatif sur ces représentations dans la mesure les séquences obtenues ne correspondent plus à des séquences réelles de la langue considérées. De fait, nous avons pu constater cet impact négatif dans le cadre d'une tâche de similarité sémantique.

	réf.	#mots éval.	#syn. / mot	rappel	R-préc.	MAP	P@1	P@5	P@10	P@100
noms	word	8 054	3,6	38,9	11,6	–	17,2	9,4	6,5	1,4
#12 154	isc	9 469	14,3	24,7	11,8	9,5	28,1	17,3	12,8	3,5
verbes	word	3 388	3,8	43,3	12,7	14,4	20,2	10,3	7,1	1,6
#4 133	isc	3 417	19,9	26,7	13,5	9,7	37,2	22,6	16,8	5,3
adjectifs	word	2 849	3,6	34,9	10,2	–	15,2	8,1	5,5	1,3
#5 539	isc	2 480	19,4	23,5	12,5	9,4	31,5	19,9	14,8	4,6

TAB. 1 – Évaluation du thésaurus distributionnel FreDist

– le dictionnaire des synonymes Dicosyn (noté *isc*), constitué de 43 202 entrées en excluant les noms composés et les noms propres, dont 7 043 adjectifs, 6 126 verbes et 30 033 noms. À l'inverse du précédent, ce dictionnaire associe beaucoup de mots à chaque entrée, mots qui vont au-delà de la notion de simple synonymie ;
– le thésaurus distributionnel FreDist (Anguiano & Denis, 2011), construit à partir d'articles du journal l'Est Républicain et de pages Wikipédia. Ce thésaurus s'appuie à la fois sur des cooccurrents syntaxiques et des cooccurrents capturés dans une fenêtre glissante de taille très restreinte. Il donne 100 voisins sémantiques pour les entrées de fréquence supérieure à 100 dans le corpus considéré.

Pour évaluer la qualité des substituts générés par FreDist, nous avons procédé à l'évaluation de ce thésaurus en utilisant les deux premiers dictionnaires comme référence. Ses résultats sont donnés par le tableau 1. Cette évaluation a été menée de façon similaire à (Ferret, 2010) en ne se restreignant pas dans un premier temps aux seuls mots cibles de l'évaluation SemDis 2014 afin d'avoir une vue d'ensemble de la qualité de la ressource et donc, de pouvoir mettre en perspective les résultats obtenus pour ces seuls mots cibles.

Ces résultats se déclinent sous la forme de plusieurs mesures, à commencer à la $5^{ème}$ colonne par le taux de rappel par rapport aux ressources considérées. Les voisins étant ordonnés, il est en outre possible de réutiliser les métriques d'évaluation classiquement adoptées en recherche d'information en faisant jouer aux entrées le rôle de requêtes et aux voisins celui des documents. Les dernières colonnes du tableau 1 rendent compte de ces mesures : la R-précision (R-préc.) est la précision obtenue en se limitant aux R premiers voisins, R étant le nombre de synonymes dans la ressource de référence pour l'entrée considérée ; la MAP (Mean Average Precision) est la moyenne des précisions pour chacun des rangs auxquels un synonyme de référence a été identifié ; enfin, sont données les précisions pour différents seuils de nombre de voisins sémantiques examinés (précision après examen des 1, 5, 10 et 100 premiers voisins). Toutes ces valeurs sont données en pourcentage.

	réf.	#mots éval.	#syn. / mot	rappel	R-préc.	MAP	P@1	P@5	P@10	P@100
noms	word	10	3.1	41.9	9.4	8.6	10.0	8.0	6.0	1.3
#10	isc	10	43.6	23.2	15.3	6.8	50.0	28.0	20.0	10.1
verbes	word	10	4.3	37.2	8.1	9.3	20.0	8.0	6.0	1.6
#10	isc	10	44.7	23.7	17.4	9.0	60.0	36.0	29.0	10.6
adjectifs	word	8	5.2	42.9	29.3	30.9	37.5	15.0	8.8	2.2
#8	isc	8	60.2	20.5	17.7	8.9	62.5	42.5	30.0	12.4

TAB. 2 – Évaluation du thésaurus distributionnel FreDist restreinte aux mots cibles utilisés pour l'évaluation

Un rapide examen de ces résultats montre des tendances assez comparables à ce que l'on peut observer pour d'autres thésaurus distributionnels utilisant des cooccurrents syntaxiques (Ferret, 2014 à paraître). Les résultats avec *word* comme référence sont ainsi assez proches de ceux que l'on peut obtenir pour l'anglais avec les synonymes de WordNet. Ceux avec le dictionnaire Dicosyn comme référence sont un peu inférieurs à ceux que l'on obtient en anglais avec une ressource comme Moby mais la différence de richesse des deux ressources explique très probablement cette différence. Une analyse selon la catégorie morpho-syntaxique montre quant à elle que l'approche distributionnelle donne des résultats particulièrement intéressants pour les verbes ; viennent ensuite les noms, puis les adjectifs, les différences entre ces trois catégories étant nettement significatives compte tenu du nombre d'entrées évaluées. Cette évaluation permet en outre d'observer le même phénomène que celui constaté dans (Ferret, 2010) de dépendance des résultats par rapport à la richesse de la

ressource de référence. La plus grande richesse de *isc* par rapport *word* explique ainsi que les résultats avec la première soit supérieurs à ceux obtenus avec la seconde.

Enfin, le tableau 2 restreint cette évaluation aux 30 mots cibles de l'évaluation SemDis 2014, 10 cibles pour chacune des trois catégories morpho-syntaxiques considérées. Ces cibles regroupent des noms comme *capacité*, *vaisseau* ou *don*, des adjectifs comme *aisé*, *hermétique* ou *riche* et des verbes comme *faucher*, *éplucher* ou *arrêter*. Outre le fait de mettre en évidence l'absence de deux adjectifs cibles sur les dix dans FreDist, cette évaluation restreinte montre que ces mots cibles ne sont pas particulièrement « faciles » du point de vue distributionnel pour ce qui est de la synonymie stricte. La situation est en revanche plus favorable pour ce qui est des voisins sémantiquement plus distants. Néanmoins, compte tenu de la dimension très paradigmatique de la tâche de génération des substituts, il n'est pas certain que cette tendance soit très favorable.

4 Évaluation

Malgré le nombre nécessairement limité de soumissions possibles, en l'occurrence 5, nous avons cherché à tester l'influence de trois grands paramètres :

- la nature de la ressource proposant les substituts. Il s'agit des trois ressources évoquées à la section précédente, c'est-à-dire le dictionnaire de synonymes issu de Word XP (*word*), le dictionnaire de synonymes Dicosyn (*isc*) et le thésaurus distributionnel FreDist (*fredist*) ;
- la mesure de similarité appliquée entre les représentations lexicales distribuées afin de juger du degré de similarité de leurs mots associés : transformée de la distance euclidienne (*l2*) ou mesure cosinus (*cos*) ;
- les mots pris en compte pour la sélection du substitut, avec deux possibilités : soit le mot cible seul (*w2*), soit tous les mots pleins de la phrase à l'exception du mot cible (*sent*), ce dernier intervenant déjà au niveau de la génération des candidats substituts. La première possibilité correspond donc à une sélection sans prise en compte du contexte du mot cible.

Plus précisément, nous avons donc fait évaluer les cinq combinaisons suivantes (leur désignation reprenant l'intitulé des soumissions correspondantes) :

cea_list-isc_l2_sent distance euclidienne avec des substituts issus du dictionnaire Dicosyn et une sélection contextuelle ;

cea_list-isc_cos_sent distance cosinus avec des substituts issus du dictionnaire Dicosyn et une sélection contextuelle ;

cea_list-isc_cos_w2 distance cosinus avec des substituts issus du dictionnaire Dicosyn et une sélection non contextuelle ;

cea_list-fredist_cos_sent distance cosinus avec des substituts issus du thésaurus FreDist et une sélection contextuelle ;

cea_list-word_cos_sent distance cosinus avec des substituts issus du dictionnaire Word XP et une sélection contextuelle.

Les résultats globaux de l'évaluation de ces cinq combinaisons sont donnés par le tableau 3, au sein duquel figurent également les soumissions des autres participants (soumission [1-4]) ainsi que les résultats d'une approche de base fondée sur le dictionnaire Dicosyn. Le dictionnaire utilisé pour cette approche de base est *a priori* identique au dictionnaire de même nom que nous avons utilisé[7]. Deux mesures ont été calculées, reprenant celles définies pour l'évaluation SemEval 2007 (McCarthy & Navigli, 2007) : *best* correspond à la proportion de bons substituts en première position tandis que *oot* (*out of ten*) prend en compte les 10 premiers substituts proposés.

Concernant nos soumissions, le tableau 3 fait clairement apparaître l'influence importante de la ressource utilisée pour générer les substituts. Ainsi, le dictionnaire Dicosyn est clairement la moins bonne des solutions pour favoriser les bons substituts au rang 1 : la présence d'un plus grand nombre de choix, certains étant assez éloignés sémantiquement par rapport au mot cible, dégrade les performances à ce niveau et signifie par là même que la méthode proposée ne choisit pas les mots les plus similaires au mot cible sur le plan de la stricte équivalence sémantique. À l'inverse, ce dictionnaire permet d'obtenir de meilleures valeurs pour le score *oot*, ce qui peut s'expliquer par sa plus grande richesse. Le meilleur compromis est obtenu avec le dictionnaire de synonymes de Word XP, qui donne en particulier la meilleure valeur pour la mesure *best*. De façon intéressante, le thésaurus distributionnel FreDist se révèle un meilleur générateur de substitut au premier rang que le dictionnaire Dicosyn tout en rivalisant avec le dictionnaire de Word XP pour les 10 premiers substituts.

[7] Sans certitude néanmoins car il en existe plusieurs versions.

systèmes	best	oot
cea_list-isc_l2_sent	0,99	23,09
cea_list-isc_cos_sent	3,32	28,66
cea_list-isc_cos_w2	3,70	28,41
cea_list-fredist_cos_sent	4,00	23,61
base dicosyn	4,53	32,45
soumission 1	5,11	21,19
soumission 2	6,26	20,48
soumission 3	6,54	35,65
cea_list-word_cos_sent	7,51	23,57
soumission 4	9,70	40,17

TAB. 3 – Résultats globaux

Concernant les deux autres paramètres testés, il apparaît clairement que la mesure de similarité cosinus est supérieure à celle tirée de la distance euclidienne, en dépit du caractère dense des représentations lexicales distribuées qui sont manipulées. Enfin, l'influence de la prise en compte du contexte pour la sélection des substituts n'est pas manifeste. Certes, notre meilleure performance est obtenue avec cette configuration mais il semble que le dictionnaire de génération des substituts en soit principalement responsable : le peu de différence entre **cea_list-isc_cos_sent** et **cea_list-isc_cos_w2** suggère que la prise en compte du contexte est au mieux neutre, voire légèrement pénalisante, rejoignant en cela (Glickman *et al.*, 2006). Ajouté à cela, la performance supérieure de l'approche de base *base dicosyn* par rapport à **cea_list-isc_cos_w2** laisse à penser que l'utilisation des représentations issues du modèle neuronal est inférieure par rapport à un simple critère de fréquence. À ce stade, nous nous garderons néanmoins de conclure sur l'intérêt de ces représentations pour la substitution lexicale. Des expériences préliminaires de construction d'un thésaurus distributionnel s'appuyant sur ces représentations nous ont en effet montré que le thésaurus résultant est d'une qualité très médiocre[8] et très inférieure à ce qui a pu être obtenu pour l'anglais par (Ferret, 2014 à paraître). La possibilité d'un problème au niveau de la construction des représentations distribuées, par exemple lié à un problème d'encodage des caractères accentués, n'est donc pas à exclure et doit être explorée plus avant pour déterminer la cause de ces résultats.

	best			oot		
	nom	adjectif	verbe	nom	adjectif	verbe
cea_list-isc_l2_sent	0,35	1,16	1,47	16,28	22,99	30,00
cea_list-isc_cos_sent	2,53	3,43	4,02	23,29	**28,73**	**33,95**
cea_list-isc_cos_w2	2,95	4,08	4,07	24,27	28,10	32,87
cea_list-fredist_cos_sent	3,18	2,83	5,99	18,12	22,45	30,26
base dicosyn	4,36	4,04	5,20	29,37	33,62	34,35
soumission 1	5,21	4,00	6,11	23,26	16,63	23,70
soumission 2	5,44	7,20	6,13	19,09	21,07	21,30
soumission 3	5,53	5,40	8,71	31,12	39,58	36,26
cea_list-word_cos_sent	7,53	7,39	7,59	19,47	24,47	26,78
soumission 4	11,02	10,58	7,49	39,77	42,85	37,87

TAB. 4 – Résultats par catégorie morpho-syntaxique

Dans ce contexte, il convient donc d'être prudent en examinant les résultats par catégorie morpho-syntaxique présentés dans le tableau 4. On peut néanmoins constater que pour la mesure *oot*, nous obtenons systématiquement l'ordre suivant des catégories : verbe > adjectif > nom. Pour la mesure *best*, la tendance est moins marquée. Le fait que la richesse des ressources utilisées[9], en particulier le dictionnaire Dicosyn, obéisse à ce même ordre n'est sans doute pas étranger à ce constat, d'autant que cette richesse a une influence significative sur les résultats en général et sur les nôtres en particulier

[8] En prenant comme référence le dictionnaire Dicosyn et le dictionnaire de Word XP.
[9] Nombre moyen de synonymes ou de mots liés associés à une entrée.

comme nous l'avons vu précédemment. On peut par ailleurs noter que les performances pour les verbes, dont la plus forte polysémie pourrait représenter *a priori* une difficulté, sont comparables, voire supérieures dans un nombre significatif de cas, à celles des autres catégories.

Si l'on se situe à un niveau plus global, (Van de Cruys *et al.*, 2011) est le travail le plus comparable à ce que nous avons présenté, en particulier parce qu'il s'agit du seul travail à notre connaissance sur la substitution lexicale en français. Il partage également avec notre approche le fait d'exploiter des facteurs latents. Dans le cas de (Van de Cruys *et al.*, 2011), ces facteurs sont induits grâce à une méthode de factorisation en matrices positives (Lee & Seung, 2000) alors que nous nous appuyons sur un modèle neuronal. Néanmoins, l'exploitation de ces facteurs latents s'inscrit dans une approche probabiliste plus élaborée que notre approche de type « sac de mots », ce qui se traduit dans le meilleur des cas par une valeur de 10,64 pour la mesure *best* et de 35,32 pour la mesure *oot* sur un jeu de test en français différent de celui de SemDis 2014 et constitué de seulement 10 noms. Les performances du même modèle sur les données de l'évaluation SemEval 2007 (McCarthy & Navigli, 2007) – 8,81 pour la mesure *best* et de 30,49 pour la mesure *oot* – laissent à penser que ce jeu de test est sans doute un peu plus facile que celui de SemDis 2014 mais la différence des résultats, en particulier pour la mesure *oot* laisse peu d'équivoque sur la supériorité du modèle de (Van de Cruys *et al.*, 2011) par rapport à notre approche. Il est à noter cependant que ces performances restent en deçà des meilleurs résultats obtenus pour l'anglais dans le cadre de l'évaluation SemEval 2007 : 20,33 pour *best* et 68,90 pour *oot*, obtenues par (Giuliano *et al.*, 2007). Comme souligné dans (Van de Cruys *et al.*, 2011), les systèmes de SemEval 2007 exploitent un inventaire préalable de substituts possibles, ce qui n'est pas le cas de (Van de Cruys *et al.*, 2011) mais correspond à l'essentiel de nos soumissions, à l'exception de celui s'appuyant sur FreDist.

5 Conclusion

Dans cet article, nous avons présenté la participation du laboratoire LVIC à l'évaluation SemDis 2014 dédiée à la substitution lexicale. Cette participation s'est centrée sur l'utilisation d'une représentation distribuée des mots construite grâce à un modèle neuronal pour sélectionner les substituts les plus intéressants. Malgré une soumission positionnée en deuxième position pour la mesure *best*, les résultats obtenus montrent que des analyses complémentaires sont nécessaires pour juger de la qualité du modèle neuronal produit et donc, de son impact sur les résultats. Ceux-ci ont montré par ailleurs que l'utilisation d'une ressource assez riche pour générer les substituts peut être problématique et que l'utilisation d'un thésaurus distributionnel pour réaliser cette tâche n'est pas à exclure.

La méthode présentée étant assez simple, les améliorations possibles sont nombreuses. La première d'entre elles consiste bien évidemment à s'assurer que les représentations produites par le modèle neuronal sont adéquates. L'utilisation d'un plus grand corpus, tel que le frWaC, permettrait par ailleurs de juger de l'impact de la taille des corpus sur les résultats. Dans le cas particulier du frWaC, cette utilisation permettrait également de déterminer dans quelle mesure les représentations neuronales sont dépendantes de leur corpus de construction puisque les données d'évaluation ont été constituées à partir de ce corpus. Au-delà, les résultats obtenus par (Mikolov *et al.*, 2013a) sur les données du *Microsoft Sentence Completion Challenge* donnent des indications sur les possibilités d'exploiter ce type de représentations dans une architecture neuronale dédiée à la substitution lexicale. De telles représentations pourraient également être utilisées dans une approche supervisée reposant sur des classifieurs plus traditionnels.

Références

ANGUIANO E. H. & DENIS P. (2011). FreDist : Automatic construction of distributional thesauri for French. In *TALN 2011, session articles courts*, Montpellier, France.

AYACHE C., GRAU B. & VILNAT A. (2006). Equer : the french evaluation campaign of question-answering systems. In *5th International Conference on Language Resources and Evaluation (LREC 2006)*, p. 575–580, Genova, Italy.

BARONI M., DINU G. & KRUSZEWSKI G. (2014, to appear). 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)*.

BENGIO Y., DUCHARME R. & VINCENT P. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, **3**, 1137–1155.

BRUNI E., BOLEDA G., BARONI M. & TRAN N. K. (2012). Distributional semantics in technicolor. In *50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, p. 136–145, Jeju Island, Korea.

COLLOBERT R., WESTON J., BATTOU L., KARLEN M., KAVUKCUOGLU K. & KUKSA P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Approach*, **12**, 2493–2537.

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.

DINU G. & LAPATA M. (2010). Measuring distributional similarity in context. In *2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*, p. 1162–1172, Cambridge, MA.

FERRET O. (2010). Testing semantic similarity measures for extracting synonyms from a corpus. In *7th International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

FERRET O. (2014, à paraître). Typing relations in distributional thesauri. Springer.

GABRILOVICH, EVGENIYAND MARKOVITCH S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI 2007*, p. 6–12.

GIULIANO C., GLIOZZO A. & STRAPPARAVA C. (2007). Fbk-irst : Lexical substitution task exploiting domain and syntagmatic coherence. In *Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, p. 145–148, Prague, Czech Republic.

GLICKMAN O., DAGAN I., DAELEMANS W., KELLER M. & BENGIO S. (2006). Investigating lexical substitution scoring for subtitle generation. In *Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, p. 45–52, New York City.

GOYAL A. & DAUME H. (2011). Generating semantic orientation lexicon using large data and thesaurus. In *2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2011)*, p. 37–43, Portland, Oregon.

GREFENSTETTE E., DINU G., ZHANG Y., SADRZADEH M. & BARONI M. (2013). Multi-step regression learning for compositional distributional semantics. In *10th International Conference on Computational Semantics (IWCS 2013)*, p. 131–142, Potsdam, Germany.

HENESTROZA ANGUIANO E. & CANDITO M. (2012). Probabilistic lexical generalization for french dependency parsing. In *ACL 2012 Joint Workshop on Statistical Parsing and Semantic Processing of Morphologically Rich Languages*, p. 1–11, Jeju, Republic of Korea.

HUANG E. H., SOCHER R., MANNING C. D. & NG A. Y. (2012). Improving word representations via global context and multiple word prototypes. In *50th Annual Meeting of the Association for Computational Linguistics (ACL'12)*, p. 873–882.

LANDAUER T. K. & DUMAIS S. T. (1997). A solution to Plato's problem : the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, **104**(2), 211–240.

LEE D. D. & SEUNG H. S. (2000). Algorithms for non-negative matrix factorization. p. 556–562.

MARTON Y. (2013). Distributional phrasal paraphrase generation for statistical machine translation. *ACM Transactions on Intelligent Systems and Technology*, **4**(3), 1–32.

MCCARTHY D. & NAVIGLI R. (2007). Semeval-2007 task 10 : English lexical substitution task. In *Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, p. 48–53, Prague, Czech Republic.

MCCARTHY D. & NAVIGLI R. (2009). The english lexical substitution task. *Language Resources and Evaluation*, **43**(2), 139–159.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013a). Efficient estimation of word representations in vector space. In *International Conference on Learning Representations 2013 (ICLR 2013), poster session*.

MIKOLOV T., YIH W.-T. & ZWEIG G. (2013b). Linguistic Regularities in Continuous Space Word Representations. In *2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (NAACL HLT 2013)*, p. 746–751, Atlanta, Georgia.

MIN B., SHI S., GRISHMAN R. & LIN C.-Y. (2012). Ensemble semantics for large-scale unsupervised relation extraction. In *2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2012)*, p. 1027–1037, Jeju Island, Korea.

MORRIS J. & HIRST G. (2004). Non-classical lexical semantic relations. In *Workshop on Computational Lexical Semantics of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, p. 46–51, Boston, MA.

PANTEL P., CRESTAN E., BORKOVSKY A., POPESCU A.-M. & VYAS V. (2009). Web-scale distributional similarity and entity set expansion. In *2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*, p. 938–947, Singapore.

SCHMID H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *International Conference on New Methods in Language Processing*.

VAN DE CRUYS T., POIBEAU T. & KORHONEN A. (2011). Latent vector weighting for word meaning in context. In *2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, p. 1012–1022, Edinburgh, Scotland, UK.

ZWEIG G., PLATT J. C., MEEK C., BURGES C. J., YESSENALINA A. & LIU Q. (2012). Computational approaches to sentence completion. In *50^{th} Annual Meeting of the Association for Computational Linguistics (ACL'12)*, p. 601–610, Jeju Island, Korea.