

Combiner des modèles sémantiques distributionnels pour mieux détecter les termes évoquant le même cadre sémantique

Gabriel Bernier-Colborne Patrick Drouin

Observatoire de linguistique Sens-Texte (OLST), Université de Montréal

C.P. 6128, succ. Centre-Ville, Montréal (QC) Canada, H3C 3J7

{gabriel.bernier-colborne|patrick.drouin}@umontreal.ca

RÉSUMÉ

Nous utilisons des modèles sémantiques distributionnels pour détecter des termes qui évoquent le même cadre sémantique. Dans cet article, nous vérifions si une combinaison de différents modèles permet d'obtenir une précision plus élevée qu'un modèle unique. Nous mettons à l'épreuve plusieurs méthodes simples pour combiner les mesures de similarité calculées à partir de chaque modèle. Les résultats indiquent qu'on obtient systématiquement une augmentation de la précision par rapport au meilleur modèle unique en combinant des modèles différents.

ABSTRACT

Combining distributional semantic models to improve the identification of terms that evoke the same semantic frame

We use distributional semantic models to detect terms that evoke the same semantic frame. In this paper, we check whether accuracy can be increased by using a combination of models rather than a single model. We test several simple methods to combine the word similarity scores obtained from each model. Results show that a combination of different models consistently outperforms the best single model.

MOTS-CLÉS : sémantique distributionnelle, sémantique lexicale, sémantique des cadres, évaluation.

KEYWORDS: distributional semantics, lexical semantics, frame semantics, evaluation.

1 Introduction

Dans le cadre d'un projet visant à décrire le vocabulaire du domaine de l'environnement au moyen de la sémantique des cadres (Fillmore, 1982), nous utilisons des modèles sémantiques distributionnels afin d'identifier, pour un terme ou un ensemble de termes donnés, des termes évoquant le même cadre sémantique. Les termes qui évoquent le même cadre sémantique participent entre eux à diverses relations lexicales, notamment la (quasi-)synonymie et la dérivation syntaxique, c'est-à-dire la relation entre deux unités lexicales ayant le même sens, mais appartenant à des parties du discours différentes (Mel'čuk *et al.*, 1995, p. 133). Une évaluation systématique de modèles produits au moyen de l'analyse distributionnelle et de l'outil `word2vec` (Mikolov *et al.*, 2013a,b) a montré que le choix et la paramétrisation du modèle dépendent des relations lexicales ciblées (Bernier-Colborne & Drouin, 2016) ; une différence particulièrement importante a été constatée entre la dérivation et d'autres relations paradigmatiques telles que la synonymie. Puisqu'on doit modéliser à la fois la synonymie et la dérivation pour bien détecter les termes évoquant le même cadre sémantique, les

différences importantes entre les modèles distributionnels qui captent le mieux ces deux types de relations lexicales limitent éventuellement leur capacité à modéliser l'appartenance au même cadre sémantique.

Dans cet article, nous faisons l'hypothèse qu'on peut combiner différents modèles sémantiques distributionnels pour mieux modéliser l'appartenance au même cadre sémantique. Plus précisément, étant donné un ensemble de modèles qui détectent relativement bien cette propriété, nous supposons que les mesures de similarité qu'on obtient à partir de ces modèles sont élevées à la fois pour les quasi-synonymes et les dérivés, mais que les différents modèles représentent différents compromis entre ces deux types de relations, et qu'on peut obtenir une sorte de consensus à partir de ces différents compromis en combinant les mesures de similarité.

2 Ressources

Nous utilisons dans l'expérience décrite ci-dessous le corpus français PANACEA du domaine de l'environnement¹ (ELRA-W0065). Ce corpus, compilé automatiquement (Prokopidis *et al.*, 2012), est constitué de pages Web totalisant environ 50 millions de mots. Le prétraitement appliqué à ce corpus comprend l'extraction du texte des documents XML (contenant plus de 50 mots), la normalisation des caractères, la lemmatisation au moyen de TreeTagger (Schmid, 1994) et la mise en minuscules ; la lemmatisation a été appliquée parce qu'elle offre parfois une légère amélioration des résultats de l'approche distributionnelle (Bullinaria & Levy, 2012).

Dans le cadre de cet article, nous évaluons les modèles distributionnels au moyen de trois jeux de données de référence extraites de dictionnaires spécialisés du domaine de l'environnement² :

- **QSYN** (quasi-synonymes) : 357 termes associés à un ou plusieurs synonymes, quasi-synonymes, variantes, cohyponymes ou autres sens voisins ; p. ex. *conservation* → {*protection*, *préservation*}.
- **DRV** (dérivés) : 259 termes associés à un ou plusieurs³ termes ayant le même sens mais appartenant à une partie du discours différente ; p. ex. *conservation* → {*conserver*}.
- **FRM** (*frames*) : 223 termes associés à un ou plusieurs termes évoquant le même cadre sémantique ; p. ex. *conservation* → {*conserver*, *préserver*, *préservation*}. Dans le cas des termes polysémiques qui évoquent plus d'un cadre, leurs voisins comprennent tous les termes évoquant au moins un de ces cadres.

3 Expérience

L'expérience que nous avons réalisée s'appuie sur les résultats d'une évaluation systématique de deux modèles sémantiques distributionnels, à savoir une technique classique appelée *analyse distributionnelle* (AD) et *word2vec* (W2V). Cette évaluation, dont les détails ont été décrits dans un autre travail (Bernier-Colborne & Drouin, 2016), a été réalisée en évaluant différentes paramétrisations

1. http://catalog.elra.info/product_info.php?products_id=1186&language=fr.

2. Les jeux QSYN et DRV ont été extraits du DiCoEnviro (http://olst.ling.umontreal.ca/cgi-bin/dicoenviro/search_enviro.cgi); FRM a été extrait du Framed DiCoEnviro (<http://olst.ling.umontreal.ca/dicoenviro/framed/index.php>). Ces ressources sont en cours de développement.

3. Un terme peut avoir plusieurs dérivés, mais dans les données que nous avons extraites, tous les termes ont un seul dérivé.

des deux modèles (360 paramétrisations de chaque modèle) au moyen de différents jeux de données de référence (y compris les jeux QSYN, DRV et FRM). La mesure d'évaluation est la moyenne des précisions moyennes (en anglais, *mean average precision* ou *MAP*), une mesure couramment utilisée pour évaluer les modèles distributionnels. La MAP^4 est la moyenne, pour chaque requête dans le jeu de données utilisé pour l'évaluation, de la précision moyenne de la liste triée de ses voisins distributionnels⁵ à chacun des rangs où se trouve un de ses voisins véritables selon les données de référence (ses quasi-synonymes, ses dérivés ou les termes évoquant le même cadre sémantique).

Les résultats de cette évaluation nous ont permis d'identifier le modèle et les paramétrisations qui offraient les meilleurs résultats pour différentes relations lexicales. Par exemple, la Table 1, qui présente la MAP maximale atteinte par chaque modèle sur les trois jeux de données de référence, montre que l'AD offre les meilleurs résultats sur les QSYN, tandis que W2V produit les meilleurs résultats sur les DRV. Les résultats sur le jeu FRM montrent que l'AD atteint une MAP plus élevée que W2V, mais l'écart est moins important que sur les autres jeux de données. Par ailleurs, on atteint une MAP beaucoup moins élevée sur ce jeu de données que sur les QSYN ou les DRV. La MAP relativement faible sur les FRM est attribuable à plusieurs facteurs ; nous reviendrons sur ce point dans la discussion de nos résultats.

Modèle	QSYN	DRV	FRM
AD	0.4476	0.3960	0.3183
W2V	0.4147	0.4567	0.3046

TABLE 1: MAP maximale des deux modèles sur les 3 jeux de données.

Dans le cadre de cet article, les résultats de cette évaluation nous ont servi à identifier les 10 modèles (des paramétrisations particulières de l'AD ou de W2V) qui produisaient les meilleurs résultats sur le jeu FRM ; nous n'énumérerons pas les valeurs des (hyper)paramètres de ces 10 modèles, mais soulignons que 5 de ces modèles ont été produits par AD et les 5 autres par W2V.

La MAP minimale, maximale et moyenne atteinte par ces 10 modèles sur les 3 jeux de données sont présentées dans la Table 2. On observe notamment que la MAP varie davantage sur les QSYN et les DRV que sur les FRM. Par ailleurs, en comparant ces données à celles présentées dans la Table 1, on observe que la MAP atteinte par ces 10 modèles, qui donnent les meilleurs résultats sur les FRM, est bien en-dessous de la MAP maximale observée en ce qui concerne les QSYN et les DRV. Cela illustre le fait qu'un modèle doit capter à la fois les QSYN et les DRV pour bien modéliser l'appartenance au même cadre sémantique. En revanche, les modèles qui sont plus « spécialisés », c'est-à-dire qui captent très bien une des deux relations au détriment de l'autre, donnent de moins bons résultats sur les FRM.

Jeu	Min.	Max.	Moy. (écart-type)
QSYN	0.3174	0.3953	0.3519 (\pm 0.0255)
DRV	0.2967	0.4007	0.3692 (\pm 0.0313)
FRM	0.2818	0.3183	0.2993 (\pm 0.0103)

TABLE 2: MAP minimale, maximale et moyenne atteinte par les 10 modèles en fonction du jeu de données.

L'expérience que nous avons réalisée consiste à calculer, pour chaque combinaison possible de deux modèles parmi ces 10 modèles, une mesure de similarité qui combine les mesures de similarité qu'on obtient à partir de chacun des deux modèles. Une matrice de similarité mot-mot est calculée à partir de

4. Voir goo.gl/t3Px5T pour une explication détaillée de la MAP.

5. Pour obtenir la liste triée des voisins distributionnels d'une requête, on calcule la similarité entre le vecteur de la requête et le vecteur de tous les autres mots-cibles représentés par le modèle ; nous utilisons dans ce travail la même mesure de similarité et les mêmes mots-cibles que lors de l'évaluation systématique initiale (Bernier-Colborne & Drouin, 2016).

chacun des deux modèles (au moyen du cosinus de l'angle des vecteurs), puis ces deux matrices, que nous appellerons S_1 et S_2 , sont combinées au moyen de l'une des fonctions suivantes pour obtenir une nouvelle matrice de similarité \hat{S} :

- similarité minimale : $\hat{S} = \min(S_1, S_2)$
- similarité maximale : $\hat{S} = \max(S_1, S_2)$
- similarité moyenne : $\hat{S} = \frac{S_1 + S_2}{2}$
- produit des similarités : $\hat{S} = S_1 * S_2$
- produit des distances : $\hat{S} = 1 - (1 - S_1) * (1 - S_2)$

4 Résultats

La Table 3 montre la MAP qu'on obtient en combinant deux mesures de similarité, en fonction de la méthode utilisée pour les combiner. Les résultats obtenus sur le jeu FRM indiquent que deux des méthodes (similarité moyenne et produit des similarités) produisent, en moyenne, une MAP plus élevée que le maximum atteint par un modèle unique (0.3183) ; le produit des distances offre en moyenne une MAP moins élevée que le produit des similarités, mais atteint un maximum légèrement supérieur. De plus, on atteint une MAP maximale de 0.3488, ce qui représente un gain de plus de 3 points de MAP (ou presque 10%) par rapport au meilleur modèle unique ; cette amélioration est significative ($p < 0.05$) selon un test de Wilcoxon (Hull, 1993). Il est intéressant de noter qu'on obtient également des gains par rapport au meilleur des 10 modèles uniques sur les QSYN et les DRV, ces améliorations étant également significatives selon le même test.

Fonction	QSYN	DRV	FRM
similarité minimale	0.3635 (0.3958)	0.3679 (0.4134)	0.3058 (0.3186)
similarité maximale	0.3426 (0.3929)	0.3739 (0.4042)	0.2954 (0.3154)
similarité moyenne	0.3658 (0.4082)	0.3933 (0.4382)	0.3196 (0.3488)
produit des similarités	0.3690 (0.4142)	0.3930 (0.4372)	0.3205 (0.3410)
produit des distances	0.3628 (0.4026)	0.3912 (0.4393)	0.3161 (0.3416)

TABLE 3: MAP moyenne (et maximale entre parenthèses) obtenue sur les 3 jeux de données au moyen des différentes fonctions utilisées pour calculer la mesure de similarité \hat{S} .

Ces résultats indiquent qu'on arrive effectivement à mieux capter l'appartenance au même cadre sémantique en combinant les mesures de similarité calculées à partir de deux modèles qui captent bien cette propriété.

Parmi les 45 paires de modèles évaluées, les 24 paires qui produisent une MAP supérieure à 0.3183 sur les FRM (le maximum atteint par un modèle unique) lorsqu'on utilise la similarité moyenne sont toutes constituées d'un modèle AD et d'un modèle W2V, et toutes les autres sauf une sont constituées de deux modèles du même type ; si on utilise plutôt le produit des similarités, toutes les paires constituées d'un modèle AD et d'un modèle W2V, et seulement celles-ci, produisent une MAP supérieure à 0.3183. Ainsi, on n'augmente pas la MAP du meilleur modèle AD en le combinant à un autre modèle AD (une paramétrisation légèrement différente), mais on obtient systématiquement une MAP plus élevée que le meilleur modèle AD en combinant un des modèles AD et un des modèles W2V.

Nous nous sommes demandé si les paires de modèles qui produisent les meilleurs résultats sur les FRM sont constituées de modèles qui produisent une MAP similaire (entre eux) sur les QSYN d'une part et sur les DRV d'autre part. Si on prend les trois paires de modèles qui produisent une MAP supérieure à 0,34 (au moyen de la similarité moyenne ou du produit des similarités), on observe qu'elles combinent toutes un modèle AD qui modélise mieux les QSYN que les DRV (et mieux que le modèle W2V) et un modèle W2V qui modélise mieux les DRV que les QSYN (et mieux que le modèle AD). Cela suggère que les paires de modèles qui produisent les meilleurs résultats sont effectivement constituées de modèles qui font des compromis différents entre la modélisation des QSYN et des DRV pour bien modéliser les FRM.

Nous avons donc montré qu'on peut mieux modéliser l'appartenance au même cadre sémantique en combinant plusieurs modèles qui captent bien (mais différemment) cette propriété. On peut se demander pourquoi on n'arrive toujours pas à atteindre une MAP aussi élevée sur les FRM que sur les QSYN ou les DRV. Cet écart est lié à divers facteurs. Un facteur important est la partie du discours des termes compris dans les données de référence. Le jeu FRM reflète un cadre descriptif particulièrement utile pour la description d'unités lexicales prédictives, à savoir la sémantique des cadres, et ce jeu contient effectivement une proportion plus élevée de verbes que le jeu QSYN, par exemple. Or, les verbes sont moins bien modélisés par l'approche distributionnelle que les noms ou les adjectifs, selon les résultats de l'évaluation systématique que nous avons exploités dans ce travail. Ainsi, la proportion élevée de verbes serait un facteur important pour expliquer la MAP relativement faible qu'on obtient sur les FRM.

Dans le cadre de cet article, nous avons évalué des combinaisons de deux modèles offrant chacun de bons résultats sur les FRM, et produisant donc de bons résultats à la fois sur les QSYN et les DRV, mais pas les meilleurs résultats possibles sur l'une ou l'autre de ces relations. Une autre possibilité consisterait à combiner des modèles plus « spécialisés », à savoir des modèles produisant les meilleurs résultats possibles sur les QSYN d'une part et les DRV d'autre part. Nous avons réalisé quelques expériences pour mettre à l'épreuve cette stratégie, et les résultats que nous avons obtenus jusqu'à date suggèrent qu'il est préférable de combiner des modèles qui produisent de bons résultats à la fois sur les QSYN et les DRV.

5 Travaux reliés

Cette étude est reliée aux travaux visant l'identification automatique de cadres sémantiques ou de représentations semblables à ces derniers (Chambers & Jurafsky, 2009, entre autres). Par contre, notre objectif n'est pas d'acquérir des cadres sémantiques complets, mais plutôt d'assister l'identification de termes évoquant le même cadre (ou des cadres reliés). Il est donc plus proche à cet égard de travaux comme celui de De Cao *et al.* (2008), qui vise à augmenter la couverture de FrameNet (Baker *et al.*, 1998) en ajoutant des unités lexicales aux cadres sémantiques encodés dans cette ressource ; les auteurs exploitent d'ailleurs une combinaison de deux mesures de similarité à cette fin, à savoir une mesure de similarité distributionnelle et une mesure basée sur WordNet (Fellbaum, 1998).

Sur le plan de la méthodologie, cette étude est reliée aux travaux exploitant des mélanges de modèles distributionnels. Par exemple, Curran (2002) combine des modèles basés sur différents types de contextes distributionnels au moyen de trois méthodes différentes (la similarité moyenne et deux méthodes basées sur les rangs de similarité plutôt que les mesures de similarité elles-mêmes) et montre que ces méthodes produisent des thésaurus distributionnels de meilleure qualité que les modèles

individuels ; il obtient d'ailleurs de meilleurs résultats en combinant tous les modèles évalués plutôt que les trois meilleurs, ce qui suggère que nous devrions tenter de combiner plus de deux modèles. La complémentarité de différents types de contextes distributionnels a également été observée par Agirre *et al.* (2009, p. 25). Par ailleurs, Ferret (2015) fusionne un thésaurus distributionnel initial et des thésaurus produits au moyen de différentes techniques de réordonnement ; il compare différentes méthodes de fusion et observe qu'elles permettent toutes d'améliorer la qualité des thésaurus.

Ce travail est également relié aux travaux comparant l'AD et les modèles de langue neuronaux implémentés dans l'outil `word2vec` (Baroni *et al.*, 2014; Levy *et al.*, 2015), mais nous ne connaissons aucun travail comparant la capacité de ces deux types de modèles distributionnels à capter des relations lexicales spécifiques, à l'exception de la synonymie. Par ailleurs, soulignons que la possibilité de combiner l'AD et `word2vec` a été évoquée par Baroni *et al.* (2014, p. 245) : « It remains to be seen whether the two types of models are complementary in the errors they make, in which case combined models could be an interesting avenue for further work. » En outre, Schwartz *et al.* (2015) ont montré qu'on peut améliorer la qualité des résultats produits par un modèle distributionnel basé sur les patrons lexico-syntaxiques en le combinant avec un modèle produit au moyen de `word2vec`.

6 Conclusion

Dans le cadre d'une démarche lexicographique outillée reposant sur l'exploitation de modèles sémantiques distributionnels, nous avons cherché à déterminer si la combinaison de différents modèles permet de mieux modéliser l'appartenance au même cadre sémantique. Nous avons montré qu'on peut effectivement mieux détecter les termes qui évoquent le même cadre en combinant des mesures de similarité calculées à partir de deux modèles distributionnels différents, au moyen de fonctions simples telles que la moyenne ou le produit des similarités. Nous constatons systématiquement une augmentation de la précision par rapport au meilleur modèle unique en combinant des modèles produits au moyen de techniques distributionnelles différentes (l'analyse distributionnelle et un modèle neuronal), alors que la combinaison de modèles similaires n'a pas un effet favorable. Une évaluation qualitative des voisinages nous permettrait éventuellement de caractériser ce que les combinaisons de modèles captent mieux que les modèles individuels et de mieux apprécier l'importance de l'amélioration observée dans le cadre de l'application envisagée.

Nous supposons que les gains observés sont liés au fait que les modèles distributionnels qui captent relativement bien l'appartenance au même cadre sémantique doivent faire un compromis entre la modélisation de la dérivation syntaxique et d'autres relations paradigmatiques telles que la synonymie. Ainsi, en combinant des modèles qui font des compromis différents pour obtenir une précision relativement élevée sur cette tâche, on arrive à augmenter cette précision davantage. Il serait intéressant de vérifier si on peut obtenir une amélioration plus importante en utilisant d'autres méthodes pour combiner les modèles, en combinant plus de deux modèles ou en exploitant différents types de contextes distributionnels.

Remerciements

Ce travail a bénéficié du soutien financier du Conseil de recherches en sciences humaines (CRSH) du Canada. Nous remercions également les relecteurs anonymes pour leurs commentaires utiles.

Références

- AGIRRE E., ALFONSECA E., HALL K., KRAVALOVA J., PAŞCA M. & SOROA A. (2009). A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of Human Language Technologies : The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, p. 19–27 : ACL.
- BAKER C. F., FILLMORE C. J. & LOWE J. B. (1998). The Berkeley FrameNet project. In *Proceedings of the 36th Meeting of the Association for Computational Linguistics (ACL) and 17th International Conference on Computational Linguistics (COLING), Volume 1*, p. 86–90 : ACL.
- BARONI M., DINU G. & KRUSZEWSKI G. (2014). Don't count, predict ! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 238–247, Baltimore, Maryland : ACL.
- BERNIER-COLBORNE G. & DROUIN P. (2016). Évaluation des modèles sémantiques distributionnels : le cas de la dérivation syntaxique. In *Actes de la 23ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Paris.
- BULLINARIA J. A. & LEVY J. P. (2012). Extracting semantic representations from word co-occurrence statistics : stop-lists, stemming, and SVD. *Behavior research methods*, **44**(3), 890–907.
- CHAMBERS N. & JURAFSKY D. (2009). Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Volume 2*, p. 602–610 : ACL.
- CURRAN J. R. (2002). Ensemble methods for automatic thesaurus extraction. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 222–229 : ACL.
- DE CAO D., CROCE D., PENNACCHIOTTI M. & BASILI R. (2008). Combining word sense and usage for modeling frame semantics. In *Proceedings of the 2008 Conference on Semantics in Text Processing*, p. 85–101 : ACL.
- FELLBAUM C. (1998). *Wordnet : An Electronic Lexical Database*. Cambridge, MA : MIT Press.
- FERRET O. (2015). Déclasser les voisins non sémantiques pour améliorer les thésaurus distributionnels. In *Actes de la 22ème conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, p. 146–157 : ATALA.
- FILLMORE C. J. (1982). Frame semantics. In THE LINGUISTIC SOCIETY OF KOREA, Ed., *Linguistics in the Morning Calm : Selected Papers from SICOL-1981*, p. 111–137. Seoul : Hanshin Publishing Co.
- HULL D. (1993). Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '93*, p. 329–338, New York, NY, USA : ACM.
- LEVY O., GOLDBERG Y. & DAGAN I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, **3**, 211–225.
- MEL'ČUK I. A., CLAS A. & POLGUÈRE A. (1995). *Introduction à la lexicologie explicative et combinatoire*. Louvain-la-Neuve : Duculot.

- MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013a). Efficient estimation of word representations in vector space. In *Proceedings of Workshop at the International Conference on Learning Representations (ICLR)*.
- MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. S. & DEAN J. (2013b). Distributed representations of words and phrases and their compositionality. In C. BURGESS, L. BOTTOU, M. WELLING, Z. GHAHRAMANI & K. WEINBERGER, Eds., *Advances in Neural Information Processing Systems 26 (NIPS)*, p. 3111–3119. Curran Associates, Inc.
- PROKOPIDIS P., PAPAVALASSIOU V., TORAL A., RIERA M. P., FRONTINI F., RUBINO F. & THURMAIR G. (2012). *Final Report on the Corpus Acquisition & Annotation subsystem and its components*. Rapport interne WP-4.5, PANACEA Project.
- SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*.
- SCHWARTZ R., REICHART R. & RAPPOPORT A. (2015). Symmetric pattern based word embeddings for improved word similarity prediction. In *Proceedings of the 19th Conference on Computational Language Learning (CoNLL)*, p. 258–267.