Exploration de modèles distributionnels au moyen de graphes 1-PPV

Gabriel Bernier-Colborne[1]
(1) OLST, Université de Montréal, CP 6128, succ. Centre-Ville, Montréal (QC) Canada, H3C 3J7
gabriel.bernier-colborne@umontreal.ca

Résumé. Dans cet article, nous montrons qu'un graphe à 1 plus proche voisin (graphe 1-PPV) offre différents moyens d'explorer les voisinages sémantiques captés par un modèle distributionnel. Nous vérifions si les composantes connexes de ce graphe, qui représentent des ensembles de mots apparaissant dans des contextes similaires, permettent d'identifier des ensembles d'unités lexicales qui évoquent un même cadre sémantique. Nous illustrons également différentes façons d'exploiter le graphe 1-PPV afin d'explorer un modèle ou de comparer différents modèles.

Abstract.

Exploring distributional semantic models using a 1-NN graph.
We show how a 1-NN graph can be used to explore the semantic neighbourhoods modeled by distributional models of semantics. We check whether the connected components of the graph, which represent sets of words that occur in similar contexts, can be used to identify sets of lexical units that evoke the same semantic frame. We then illustrate different ways in which the 1-NN graph can be used to explore a model or compare different models.

Mots-clés : Sémantique distributionnelle, sémantique lexicale, graphe, terminologie, sémantique des cadres.

Keywords: Distributional semantics, lexical semantics, graph, terminology, frame semantics.

1 Introduction

Dans le cadre d'un projet visant à décrire le vocabulaire du domaine de l'environnement, nous cherchons à faciliter l'identification de relations sémantiques paradigmatiques telles que la synonymie ainsi que l'identification d'ensembles d'unités lexicales évoquant un même cadre sémantique, suivant le cadre descriptif proposé par Fillmore (1982). Nous exploitons à cette fin des techniques permettant l'identification semi-automatique de relations sémantiques à partir de corpus spécialisés. L'analyse distributionnelle permet d'estimer la similarité sémantique de deux mots en comparant les contextes dans lesquels ils apparaissent dans un corpus, l'hypothèse sous-jacente étant que les mots qui apparaissent dans des contextes similaires ont tendance à présenter des affinités sémantiques (Harris, 1954). La similarité distributionnelle, qui peut être calculée de différentes façons, est souvent utilisée pour construire des thésaurus distributionnels, ressources associant à chaque entrée une liste de ses plus proches voisins (PPV) selon la mesure de similarité.

Un thésaurus distributionnel peut être considéré comme un graphe de k plus proches voisins (graphe k-PPV), graphe dans lequel chaque mot est relié à ses k PPV. Les graphes k-PPV peuvent servir non seulement à représenter le voisinage d'un mot donné, mais aussi à identifier des ensembles de mots sémantiquement reliés. Différentes techniques peuvent être utilisées à cette fin ; une technique simple consiste à calculer les composantes connexes d'un graphe 1-PPV, celles-ci représentant des ensembles de mots distributionnellement similaires. Dans cet article, nous vérifions si le graphe 1-PPV peut faciliter l'identification d'ensembles d'unités lexicales qui évoquent un même cadre sémantique ; nous montrons également que le graphe nous fournit une perspective intéressante sur les voisinages sémantiques captés par un modèle distributionnel. Nous décrirons le graphe 1-PPV à la Section 2 et les ressources que nous avons utilisées à la Section 3. Nous évaluerons le graphe à la Section 4 et présenterons différentes façons d'exploiter le graphe à la Section 5. Enfin, nous présenterons quelques travaux reliés à la Section 6.

2 Le graphe 1-PPV

Tout modèle qui permet d'estimer la similarité de deux mots peut donner lieu à la construction d'un graphe de voisinage qui relie chaque mot à ses PPV. Dans le cadre de ce travail, nous avons utilisé deux modèles différents, à savoir HAL et word2vec (W2V), afin de vérifier si le graphe 1-PPV révèle des différences quant aux voisinages sémantiques qu'ils modélisent. HAL (Lund *et al.*, 1995; Schütze, 1992) est un modèle à fenêtre graphique qui repose sur une matrice mot-mot qui encode la fréquence de cooccurrence des mots. Le modèle de langue neuronal word2vec (Mikolov *et al.*, 2013a,b), qui a été exploité dans plusieurs applications du TAL dans les dernières années, apprend des représentations distribuées de mots qui peuvent servir à estimer la similarité sémantique, entre autres. Dans les deux cas, nous calculons la similarité entre les mots au moyen du cosinus de l'angle de leurs vecteurs ; les PPV d'un mot sont obtenus en calculant la similarité entre ce mot et tous les autres mots dans le vocabulaire.

Le graphe que nous utilisons est un graphe k-PPV symétrique [1], c'est-à-dire un graphe non orienté dans lequel deux mots w_i et w_j sont reliés par une arête si w_i est un des k PPV de w_j ou si w_j est un des k PPV de w_i. Le nombre de composantes connexes [2] dans ce graphe dépend de la valeur de k, entre autres : en faisant varier la valeur de k, nous avons observé que le graphe k-PPV symétrique devient généralement connexe dès que la valeur de k atteint 3 ou 4. Un graphe connexe peut servir à identifier des ensembles de mots similaires au moyen d'une technique permettant de repérer des sous-graphes de forte densité. Dans cet article, nous utilisons plutôt un graphe qui n'est pas connexe, le graphe 1-PPV symétrique, dans lequel deux mots sont reliés si l'un est le PPV de l'autre. Nous utilisons le graphe 1-PPV parce que le calcul de ses composantes connexes nous fournit un moyen simple d'identifier des ensembles de mots distributionnellement similaires. Nous vérifierons ainsi si ce genre de graphe peut faciliter l'identification d'ensembles d'unités lexicales évoquant un même cadre sémantique (ces ensembles seront décrits à la Section 3) dans le cadre de l'élaboration d'une ressource lexicale spécialisée basée sur la sémantique des cadres.

3 Ressources utilisées

Les modèles ont été construits [3] sur le corpus monolingue français PANACEA – domaine de l'environnement (ELRA-W0065), un corpus de pages Web reliées au domaine de l'environnement, contenant 23 514 documents (environ 47 millions de tokens). Ce corpus a été compilé au moyen d'un outil de construction automatique de corpus spécialisés conçu dans le cadre du projet PANACEA et il est distribué librement à des fins de recherche [4]. Le corpus a été lemmatisé au moyen de TreeTagger (Schmid, 1994).

Nous avons également obtenu des données de référence afin d'évaluer le graphe 1-PPV. Ces données ont été extraites du Framed DiCoEnviro [5] (L'Homme & Robichaud, 2014; L'Homme *et al.*, 2014), une ressource lexicale spécialisée décrivant les termes du domaine de l'environnement au moyen de la sémantique des cadres (Fillmore, 1982). La méthodologie utilisée pour construire le Framed DiCoEnviro est inspirée de celle du projet FrameNet (Ruppenhofer *et al.*, 2010) ; certains des cadres sont basés sur ceux dans FrameNet, d'autres ont été élaborés spécifiquement pour le domaine de l'environnement. Nous avons extrait de cette ressource des ensembles d'unités lexicales qui évoquent le même cadre sémantique ; p. ex. le cadre Change_of_Temperature est évoqué par les unités lexicales *réchauffer*, *réchauffement*, *refroidir* et *refroidissement*. Pour chaque cadre, nous avons extrait les unités lexicales qui font partie du vocabulaire pour lequel nous avons construit des modèles, qui est constitué des 10000 formes lemmatisées les plus fréquentes dans le corpus (les mots vides étant exclus au moyen d'un anti-dictionnaire). Parmi les ensembles résultants, nous avons conservé ceux contenant au moins 2 mots, puis nous avons éliminé le recoupement entre les ensembles : lorsque deux ensembles partageaient au moins un mot, nous avons conservé seulement le plus gros ensemble. Nous avons ainsi obtenu 52 ensembles de référence. Le nombre de mots par ensemble varie entre 2 et 10, 42 des 52 ensembles contenant 4 mots ou moins.

1. Il existe différents types de graphes de voisinage, tels que les graphes k-PPV symétriques et mutuels (Maier *et al.*, 2007).
2. Les composantes connexes d'un graphe sont les sous-graphes de taille maximale dans lesquels il existe un chemin entre n'importe quelle paire de sommets.
3. Nous utilisons les librairies python suivantes : gensim (`http://radimrehurek.com/gensim/`) pour l'entraînement des modèles word2vec, scikit-learn (`http://scikit-learn.org/`) pour la SVD et networkx (`http://networkx.github.io/`) pour la construction, l'analyse et la visualisation de graphes.
4. Voir `http://catalog.elra.info/product_info.php?products_id=1186&language=fr` et `http://panacea-lr.eu/`
5. `http://olst.ling.umontreal.ca/dicoenviro/framed/index.php` (en construction)

4 Évaluation

Les ensembles de référence décrits à la section 3 ont été utilisés afin d'évaluer les graphes 1-PPV symétriques créés à partir des modèles distributionnels. Cette évaluation a été réalisée afin de vérifier si les composantes du graphe 1-PPV permettent d'identifier des ensembles d'unités lexicales évoquant un cadre sémantique ; elle a également servi à choisir un modèle pour les exemples présentés dans cet article. Pour chaque modèle, nous avons évalué différentes paramétrisations [6]. Dans le cas du modèle HAL, nous avons fait varier le type, la forme et la taille de la fenêtre de contexte ainsi que la pondération appliquée aux fréquences de cooccurrence ; l'influence de ces paramètres a été analysée par Bullinaria & Levy (2007), entre autres. Pour chaque paramétrisation, nous avons également appliqué la SVD (300 dimensions, algorithme ARPACK) au modèle, suivant Schütze (1992). Nous avons également évalué différentes paramétrisations du modèle word2vec (W2V) en faisant varier certains de ses principaux paramètres : l'architecture du modèle, l'algorithme d'entraînement, la taille de fenêtre et le nombre de dimensions. Nous avons ainsi évalué 40 paramétrisations de chaque modèle (HAL, SVD et W2V).

Pour chacune des paramétrisations, nous avons calculé différentes caractéristiques du graphe 1-PPV résultant et évalué le graphe au moyen de mesures d'évaluation externes calculées en comparant les composantes connexes du graphe aux ensembles de référence décrits à la Section 3. Dans le cadre de cet article, les mesures que nous utilisons sont deux mesures simples de rappel. Le rappel est simplement le pourcentage des ensembles de référence dont tous les mots se retrouvent dans la même composante. Cette mesure permet d'estimer dans quelle mesure les composantes du graphe permettent d'identifier des ensembles d'unités lexicales évoquant un même cadre sémantique. Pour pénaliser les graphes contenant un nombre faible de composantes (ce qui augmente la probabilité que les mots d'un ensemble de référence se retrouveront dans la même composante), nous avons également calculé une mesure de rappel corrigée en fonction du nombre de composantes, de la façon suivante : $R_{corr} = R \cdot \frac{2|C|}{|V|}$, où R est le rappel, $|C|$ est le nombre de composantes dans le graphe et $|V|$ est le nombre de mots dans le vocabulaire.

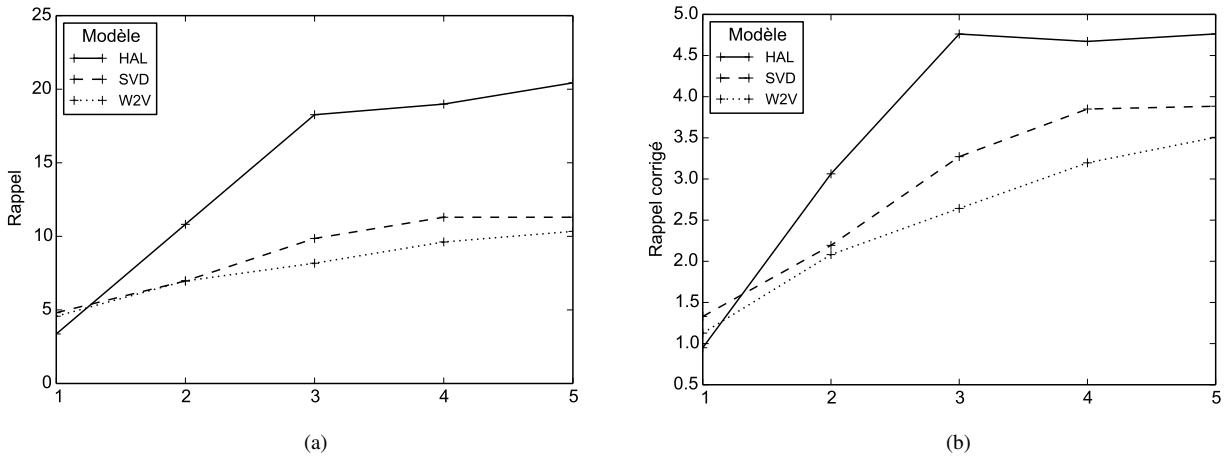


FIGURE 1: Influence de la taille de la fenêtre de contexte sur **(a)** le rappel et **(b)** le rappel corrigé. Tous les points sont des moyennes sur l'ensemble des paramétrisations correspondant à une taille de fenêtre et un modèle particuliers.

Le rappel que nous avons observé est de 14.37% en moyenne pour les modèles HAL, 8.85% pour les modèles SVD et 7.93% pour les modèles W2V. Le rappel maximal est de 25% pour les modèles HAL, 17.31% pour les modèles SVD et 15.38% pour les modèles W2V. Ces résultats suggèrent que les composantes du graphe 1-PPV peuvent servir à identifier des ensembles d'unités lexicales qui évoquent un cadre sémantique dans une certaine mesure, bien qu'une augmentation du rappel demeure souhaitable ; à ce titre, il serait intéressant de vérifier comment ce graphe se compare à d'autres types de graphes de voisinage.

Un des paramètres importants des 3 modèles que nous utilisons est la taille de la fenêtre de contexte. La Figure 1 montre l'influence de la taille de fenêtre et du type de modèle sur le rappel et le rappel corrigé. Lorsque le rappel est corrigé en fonction du nombre de composantes, la différence entre les 3 types de modèles est moins importante parce que les

6. Nous ne décrivons pas en détail chacun des paramètres des différents modèles, parce que nous n'examinons pas l'influence des paramètres dans le cadre de cet article, à l'exception de la taille de fenêtre.

modèles HAL produisent un nombre moins élevé de composantes (1318 en moyenne) que les modèles SVD (1611) et W2V (1534). Le fait que HAL offre un rappel plus élevé que les modèles SVD et W2V est lié à plusieurs facteurs, outre le nombre de composantes ; un de ces facteurs est la nature des données de référence, qui peuvent notamment contenir des unités lexicales de différentes parties du discours. Soulignons que le rappel maximal a été atteint avec une fenêtre de 3 mots pour les modèles HAL, 4 mots pour les modèles SVD et 5 mots pour les modèles W2V ; il serait donc intéressant de tester des fenêtres plus larges pour vérifier si le rappel de W2V continue à augmenter.

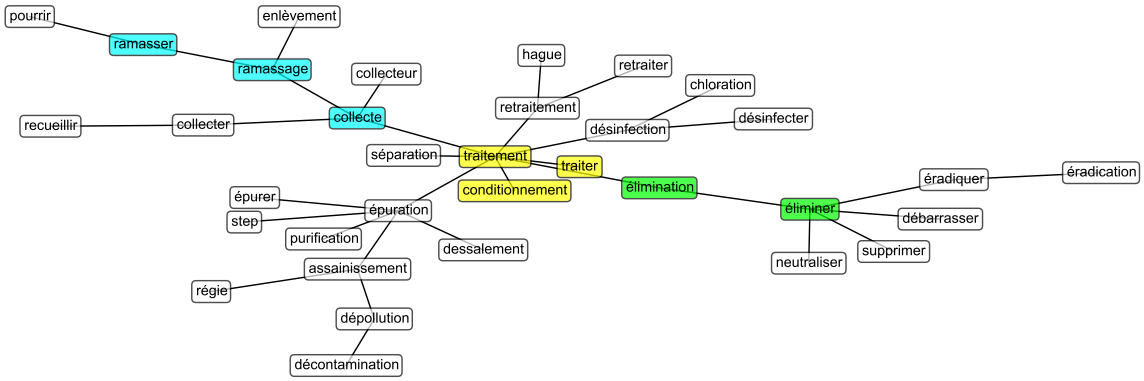
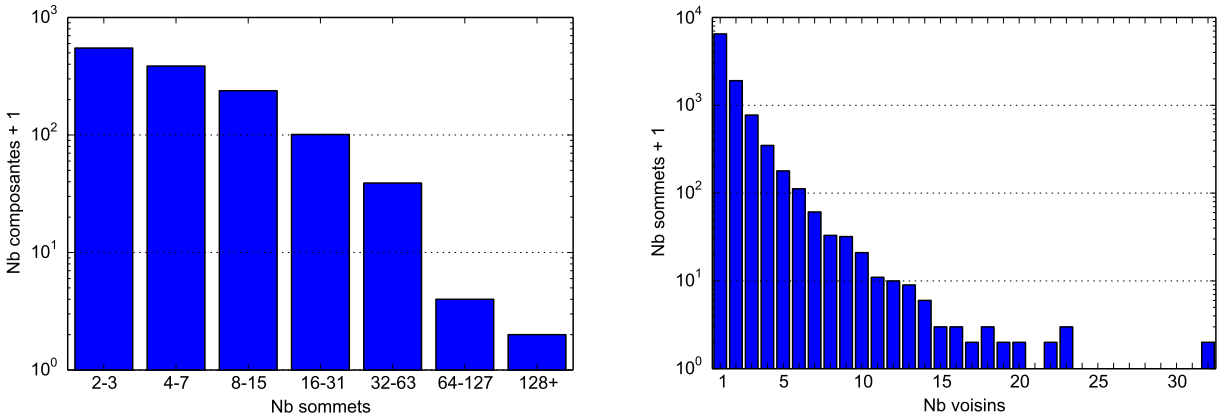


FIGURE 2: Composante du graphe 1-PPV contenant 3 ensembles de référence (mots en bleu, en jaune et en vert).

Le modèle que nous avons retenu pour les exemples présentés dans cet article (sauf indication contraire) est celui qui maximise à la fois le rappel (25%) et le rappel corrigé (6.54) ; il s'agit d'un des modèles HAL calculés au moyen d'une fenêtre de 3 mots. La Figure 2 montre une composante du graphe 1-PPV correspondant. Cette composante contient 3 des ensembles de référence ; ceux-ci sont constitués d'unités lexicales qui évoquent 3 cadres sémantiques liés à la gestion des matières résiduelles (Collecting, Processing_materials et Removing).

La Figure 3 illustre des caractéristiques de ce graphe, à savoir la distribution du nombre de sommets par composante et la distribution du nombre de voisins par sommet. Les graphes 1-PPV contiennent un petit nombre de grosses composantes (la plus grosse contenant en moyenne 179 sommets) et un nombre élevé de petites composantes, dont beaucoup ne contiennent que 2 sommets ; ces composantes d'ordre 2 représentent des PPV *réciproques* ou *mutuels* (paires de mots dont chacun est le PPV de l'autre). Le nombre moyen de composantes d'ordre 2 était de 342 pour les modèles HAL, 360 pour les modèles SVD et 389 pour les modèles W2V, ce qui suggère que W2V produit un nombre plus élevé de PPV mutuels.



(a) (b)

FIGURE 3: **(a)** Distribution du nombre de sommets par composante. **(b)** Distribution du nombre de voisins par sommet.

Rang	Sommets	Mots les plus proches du vecteur moyen
1	196	pouvoir, falloir, aller, vouloir, devoir, jamais, faire, déjà, commencer, ...
2	107	of, for, and, the, research, environmental, policy, water, to, ...
3	86	espèce, oiseau, mammifère, animal, poisson, menacer, population, sauvage, rare, ...
4	78	michel, jacques, françois, alain, patrick, paul, dominique, pierre, andré, ...
5	61	substance, toxique, chimique, produit, contenir, polluant, molécule, composé, dangereux, ...
6	61	maïs, céréale, culture, soja, blé, riz, cultiver, colza, pomme, ...
7	61	philippe, jean, bernard, christian, daniel, pascal, jean-pierre, jean-claude, gérard, ...
8	58	déterminer, définir, établir, fixer, évaluer, examiner, décrire, étudier, mesurer, ...
9	55	penser, croire, savoir, peur, apprendre, imaginer, rêver, oublier, douter, ...
10	54	paris, lyon, édition, lille, rennes, toulouse, marseille, bordeaux, montpellier, ...
11	54	autorisation, permis, autoriser, agrément, délivrer, certificat, dérogation, interdiction, interdire, ...
12	53	arbre, plante, feuille, fleur, arbuste, herbe, racine, végétal, graine, ...
13	51	réduire, diminuer, augmenter, limiter, accroître, réduction, baisser, minimiser, croître, ...
14	50	organiser, participer, lancer, réunir, rassembler, annoncer, initier, regrouper, dérouler, ...
15	49	chercheur, scientifique, expert, biologiste, spécialiste, équipe, économiste, auteur, journaliste, ...

TABLE 1: Mots centraux des composantes comprenant le plus grand nombre de sommets.

5 Applications du graphe

Dans la section précédente, nous avons montré que le calcul des composantes connexes du graphe 1-PPV, un moyen simple d'obtenir des ensembles de mots distributionnellement similaires, permet dans certains cas d'identifier des ensembles d'unités lexicales évoquant un même cadre sémantique. Le graphe peut également être utilisé de différentes façons afin d'explorer ou de caractériser les voisinages sémantiques captés par un modèle distributionnel. Par exemple, on peut observer les composantes qui contiennent le plus grand nombre de sommets. Les plus grosses composantes du graphe que nous avons retenu sont présentées dans le Tableau 1, chaque composante étant illustrée au moyen des mots les plus proches du vecteur moyen des mots dans la composante.

Une autre caractéristique intéressante du graphe 1-PPV est le degré (nombre de voisins) des sommets. Par exemple, les sommets ayant le plus de voisins dans le graphe que nous avons retenu sont des mots très fréquents : *pouvoir* (32 voisins), *gens* (23), *aller* (23), *philippe* (22), *oiseau* (20), *of* (19), *insecte* (18), *croire* (18), *falloir* (17) et *chose* (16). En revanche, si on applique la SVD à ce modèle, les sommets qui ont le plus de voisins dans le graphe 1-PPV comprennent des mots beaucoup moins fréquents, dont plusieurs prénoms ; soulignons aussi que le degré maximal du modèle HAL est deux fois plus élevé que celui du modèle SVD correspondant. Dans le cas du modèle W2V qui obtient le meilleur rappel, les sommets ayant le plus de voisins sont des mots à très faible fréquence : *connerie* (26 voisins), *poésie* (16), *sabine* (15), *objectivité* (13), *inadmissible* (12), *cruauté* (12), *economic* (12), *michèle* (12), *with* (11), *cendrer* (11).

Le graphe est également utile à des fins de visualisation. Par exemple, pour visualiser le voisinage d'un mot, on peut vérifier à quelle composante il appartient et produire une figure comme la Figure 2. Si la composante qui contient une requête donnée ne contient pas suffisamment de sommets pour illustrer adéquatement le voisinage de la requête, on peut vérifier quelles autres composantes sont proches de la requête [7]. Par exemple, la composante contenant le mot *absorber* contient 13 mots ; en calculant les 3 autres composantes les plus proches du vecteur de *absorber*, on obtient les 4 composantes illustrées dans la Figure 4.

Le graphe 1-PPV offre de nombreuses possibilités. Par exemple, il serait possible de calculer un deuxième graphe 1-PPV sur les vecteurs moyens des composantes du graphe 1-PPV pour obtenir une représentation plus abstraite des voisinages sémantiques, une possibilité que nous avons commencé à explorer. On pourrait également imaginer différentes façons d'améliorer la cohésion sémantique des ensembles de mots similaires que représentent les composantes du graphe 1-PPV.

7. La distance entre une composante et la requête peut être estimée de différentes façons ; nous utilisons la similarité entre le vecteur de la requête et le vecteur moyen des mots dans une composante donnée, mesurée au moyen du cosinus de l'angle des vecteurs.

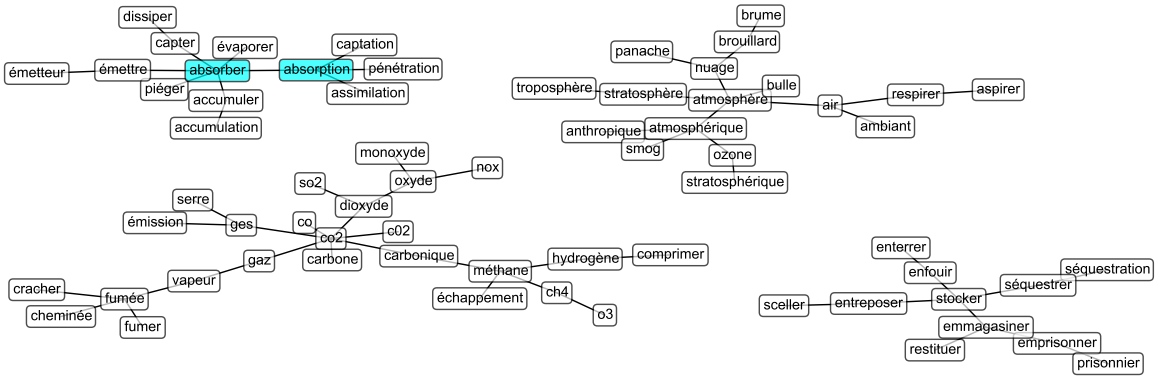


FIGURE 4: Composante contenant la requête *absorber*, accompagnée des 3 autres composantes les plus proches du vecteur de la requête. Les mots *absorber* et *absorption* (en bleu), forment un ensemble de référence (cadre Soaking_up).

6 Travaux reliés

À notre connaissance, il existe peu de travaux qui ont cherché à exploiter les graphes de voisinage afin d'explorer des modèles distributionnels ou d'identifier des ensembles de mots sémantiquement reliés. Gyllensten & Sahlgren (2015) soulignent que la méthode généralement utilisée pour interroger un modèle distributionnel, qui consiste à obtenir une liste ordonnée de voisins pour un mot donné, ne rend pas compte de la structure interne du voisinage du mot. Ils proposent d'utiliser un graphe de voisinage relatif pour décrire le voisinage d'un mot d'une façon qui rend compte de ses différents sens ; ils utilisent notamment cette méthode pour comparer les propriétés de différents modèles distributionnels. Des méthodes basées sur graphe ont été utilisées dans plusieurs travaux visant à découvrir les différents sens ou usages des mots à partir de corpus ; ces méthodes exploitent généralement un graphe de cooccurrence (Dorow & Widdows, 2003; Véronis, 2003; Biemann, 2006; Di Marco & Navigli, 2013), mais des graphes de similarité distributionnelle ont également été utilisés (Ferret, 2004). Morardo & Villemonte de La Clergerie (2013) présentent une plateforme de production, de visualisation et de validation de ressources lexicales dont une des composantes principales permet de construire des réseaux lexicaux basés sur la similarité distributionnelle des termes. En ce qui concerne les réseaux lexicaux, Steyvers & Tenenbaum (2005) ont analysé la structure de trois réseaux différents afin de modéliser l'acquisition et l'évolution du lexique, et ont comparé les propriétés de ces réseaux à celles des graphes de voisinage produits au moyen de l'analyse sémantique latente (Landauer & Dumais, 1997). En outre, Claveau *et al.* (2014) considèrent les thésaurus distributionnels comme des graphes k-PPV et exploitent l'information contenue dans ces graphes afin d'améliorer la qualité des thésaurus. Enfin, nous ne connaissons aucun travail portant spécifiquement sur les graphes 1-PPV des modèles distributionnels.

7 Conclusion

Dans cet article, nous avons montré que les composantes connexes d'un graphe 1-PPV symétrique offrent différents moyens d'explorer les voisinages sémantiques captés par un modèle distributionnel et de comparer différents modèles. Nous avons montré que ces composantes, qui représentent des ensembles de mots distributionnellement similaires, permettent dans certains cas d'identifier des ensembles d'unités lexicales qui évoquent un même cadre sémantique. Une évaluation plus approfondie serait nécessaire pour déterminer plus précisément dans quelle mesure les graphes de voisinage distributionnel peuvent faciliter l'identification de ces ensembles. À ce titre, nous comptons mettre à l'épreuve différents types de graphes de voisinage et continuer à développer notre méthodologie d'évaluation afin de mieux évaluer l'efficacité de ces méthodes dans le cadre de l'élaboration de ressources lexicales spécialisées.

Remerciements

Ce projet bénéficie du soutien financier du Conseil de recherches en sciences humaines (CRSH) du Canada.

Références

BIEMANN C. (2006). Chinese whispers : an efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of the first workshop on graph based methods for natural language processing*, p. 73–80 : ACL.

BULLINARIA J. A. & LEVY J. P. (2007). Extracting semantic representations from word co-occurrence statistics : A computational study. *Behavior research methods*, **39**(3), 510–526.

CLAVEAU V., KIJAK E. & FERRET O. (2014). Explorer le graphe de voisinage pour améliorer les thésaurus distributionnels. In B. BIGI, Ed., *Actes de TALN 2014 (Traitement automatique des langues naturelles)*, p. 220–231, Marseille : ATALA LPL.

DI MARCO A. & NAVIGLI R. (2013). Clustering and diversifying web search results with graph-based word sense induction. *Computational Linguistics*, **39**(3), 709–754.

DOROW B. & WIDDOWS D. (2003). Discovering corpus-specific word senses. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics – Volume 2*, p. 79–82 : ACL.

FERRET O. (2004). Découvrir des sens de mots à partir d'un réseau de cooccurrences lexicales. In P. BLACHE, Ed., *Actes de TALN 2004 (Traitement automatique des langues naturelles)*, Fès, Maroc : ATALA LPL.

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.

GYLLENSTEN A. C. & SAHLGREN M. (2015). Navigating the semantic horizon using relative neighborhood graphs. *CoRR*, **abs/1501.02670**.

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

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.

L'HOMME M.-C. & ROBICHAUD B. (2014). Frames and terminology : Representing predicative terms in the field of the environment. In *Proceedings of CogALex*, p. 186–197, Dublin : ACL, DCU.

L'HOMME M.-C., ROBICHAUD B. & SUBIRATS RÜGGEBERG C. (2014). Discovering frames in specialized domains. In *Proceedings of LREC*, p. 1364–1371, Reykjavik : ELRA.

LUND K., BURGESS C. & ATCHLEY R. A. (1995). Semantic and associative priming in high-dimensional semantic space. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society*, p. 660–665.

MAIER M., HEIN M. & VON LUXBURG U. (2007). Cluster identification in nearest-neighbor graphs. In *Algorithmic Learning Theory*, p. 196–210 : Springer.

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

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

MORARDO M. & VILLEMONTE DE LA CLERGERIE É. (2013). Vers un environnement de production et de validation de ressources lexicales sémantiques. In *Actes de SemDis 2013 : Enjeux actuels de la sémantique distributionnelle*, p. 167–180, Les Sables d'Olonne, France.

RUPPENHOFER J., ELLSWORTH M., PETRUCK M. R. L., JOHNSON C. R. & SCHEFFCZYK J. (2010). FrameNet II : Extended theory and practice. `http://framenet2.icsi.berkeley.edu/docs/r1.5/book.pdf`.

SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*.

SCHÜTZE H. (1992). Dimensions of meaning. In *Proceedings of the 1992 ACM/IEEE Conference on Supercomputing (Supercomputing'92)*, p. 787–796 : IEEE Computer Society Press.

STEYVERS M. & TENENBAUM J. B. (2005). The large-scale structure of semantic networks : Statistical analyses and a model of semantic growth. *Cognitive science*, **29**(1), 41–78.

VÉRONIS J. (2003). Cartographie lexicale pour la recherche d'information. In B. DAILLE, Ed., *Actes de TALN 2003 (Traitement automatique des langues naturelles)*, p. 265–274, Batz-sur-mer : ATALA IRIN.