Analyse distributionnelle de corpus spécialisés pour l'identification de relations lexico-sémantiques

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. Nous décrivons une étude visant à repérer automatiquement des relations lexico-sémantiques à partir de corpus spécialisés au moyen d'une méthode d'analyse distributionnelle. Les résultats obtenus montrent qu'un modèle non structuré, basé sur la cooccurrence des mots dans le corpus, permet d'obtenir, pour un terme donné, des termes reliés sur le plan paradigmatique (quasi-synonymes, antonymes, hyponymes). Nous discuterons la méthodologie d'évaluation et de sélection des paramètres, qui exploite des données extraites d'un dictionnaire spécialisé. Nous analyserons l'influence de paramètres tels que la forme et la taille de la fenêtre de contexte, la pondération des statistiques et l'utilisation d'une technique de réduction de dimension. Nous comparerons également les relations identifiées dans deux corpus, un portant sur le domaine de l'environnement et l'autre, sur le traitement automatique de la langue.

Abstract. We describe an experiment wherein a word space model is used to automatically extract lexico-semantic relations from specialized corpora. Results show that an unstructured model, which exploits basic word cooccurrence information, can effectively identify paradigmatically related terms (near synonyms, antonyms, hyponyms) given a target term. We discuss the parameter selection and evaluation methodologies, which rely on data extracted from a specialized dictionary. We analyze the impact of parameters such as the shape and size of the context window, the weighting scheme and the use of dimensionality reduction. We also compare the relations identified in two specialized corpora, one dealing with the environment and the other pertaining to natural language processing.

Mots-clés : Sémantique distributionnelle, sémantique computationnelle, relations lexico-sémantiques, corpus spécialisé, terminologie.

Keywords: Distributional semantics, computational semantics, lexico-semantic relations, specialized corpora, terminology.

1 Introduction

Dans le cadre d'un projet portant sur l'identification de thématiques en corpus spécialisé, nous cherchons à extraire des relations lexico-sémantiques à partir de données textuelles. Notre objectif est d'obtenir, à partir d'un terme donné, des termes dont le sens est relié à celui de la requête ; dans cet article, nous nous intéresserons particulièrement à une classe de relations paradigmatiques classiques, à savoir la (quasi-)synonymie, l'antonymie et l'hyponymie. Il n'est pas important, du moins à cette étape du projet, que les relations extraites soient étiquetées, seulement qu'elles concernent des termes du domaine ciblé pour le projet et qu'elles appartiennent à cette classe particulière de relations.

Les techniques de la sémantique distributionnelle apparaissent comme un moyen efficace de réaliser cette tâche. Celles-ci sont basées sur l'hypothèse distributionnelle, d'abord formulée par (Harris, 1954), selon laquelle les mots apparaissant dans des contextes similaires ont tendance à présenter des affinités sémantiques. Ces techniques ont d'abord été déployées sur des corpus spécialisés, "puisque c'est précisément pour traiter des données de ce type qu'a été formulée l'hypothèse distributionnelle" (Morlane-Hondère & Fabre, 2012, p. 1001). La tendance actuelle consiste plutôt à utiliser des corpus les plus gros possibles, provenant souvent de sources hétérogènes, dont le nombre de mots dépasse souvent le milliard. (Adam *et al.*, 2013) soulignent cette tendance, et optent délibérément pour un corpus de taille plus modeste ; de même, (Ferret, 2010) utilise un corpus relativement petit parce que la taille des corpus qu'il est possible de construire dépend de la langue et du domaine ciblés. Les corpus utilisés dans ces travaux contiennent tout de même des centaines de millions de

mots. Ainsi, il est difficile de déterminer dans quelle mesure les techniques de la sémantique distributionnelle permettront d'identifier des relations lexico-sémantiques dans un corpus spécialisé contenant quelques millions de mots seulement.

En lien avec la question de la taille et de la nature des corpus se pose celle du type de modèle utilisé, ou plus précisément la nature des contextes utilisés pour construire le modèle. À notre connaissance, les travaux décrivant l'application de méthodes distributionnelles à des corpus spécialisés (Grefenstette, 1992; Nazarenko *et al.*, 1997; Bourigault, 2002) ont surtout exploité des modèles structurés, à savoir des modèles qui exploitent des contextes de nature syntaxique plutôt que la simple cooccurrence. Nous avons plutôt opté pour un modèle non structuré, tout comme (Ferret, 2010), qui justifie ce choix par le fait que les analyseurs syntaxiques robustes ne sont pas disponibles pour toutes les langues. Par ailleurs, l'auteur observe que les résultats qu'il obtient sont comparables à ceux obtenus au moyen d'un modèle structuré sur la même tâche (WordNet-Based Synonymy Test). Puisque l'auteur utilise un corpus de plusieurs centaines de millions de mots, nous ne pouvons pas conclure d'emblée qu'un modèle non structuré produira de bons résultats sur un petit corpus spécialisé. Voilà une des questions auxquelles nous tenterons de répondre dans cet article, à savoir si un modèle non structuré permet d'identifier des relations lexico-sémantiques dans un corpus spécialisé de petite taille.

À cette fin, nous avons construit des modèles sur un corpus du domaine de l'environnement et comparé les voisinages identifiés à des données extraites d'un dictionnaire spécialisé du même domaine. Comme le soulignent (Adam *et al.*, 2013), ce type d'évaluation ne permet pas d'évaluer la qualité de tous les liens de voisinage distributionnel, qui peuvent correspondre à des relations qui ne sont pas décrites dans la ressource lexicale. Dans cette optique, nous avons réalisé une évaluation manuelle portant sur les voisins considérés comme incorrects lors de l'évaluation automatique, ce qui permet non seulement une mesure plus exacte de la précision des résultats, mais aussi une estimation de la capacité du modèle à améliorer la couverture de la ressource lexicale, aspect occulté par l'évaluation automatique.

La 2e tâche de cette édition de SemDis nous fournit l'occasion d'examiner les résultats obtenus sur ce corpus, puis de les comparer à ceux que l'on obtient sur un corpus comparable quant à sa taille et sa nature spécialisée, mais portant sur un domaine différent, à savoir le traitement automatique de la langue. L'approche que nous avons adoptée consiste à déterminer les paramètres optimaux du modèle en explorant systématiquement l'espace des paramètres et en évaluant les modèles résultants sur les données de référence. Par la suite, nous construisons un nouveau modèle sur le corpus TALN en utilisant les mêmes paramètres, et comparons les résultats obtenus sur les deux corpus.

Une partie de cet article sera donc consacrée à la sélection des paramètres du modèle, sujet qui a fait l'objet de nombreux travaux sur la sémantique distributionnelle. Par exemple, (Sahlgren, 2006) a examiné l'influence du type d'information contextuelle exploité par le modèle (segments textuels dans le cas de la LSA, cooccurrents dans le cas de HAL), et l'influence de la distance ou mesure de similarité entre vecteurs a été examinée par (Weeds *et al.*, 2004; Ferret, 2010). En ce qui concerne HAL, la méthode que nous employons dans ce travail, (Bullinaria & Levy, 2007) ont évalué l'influence de plusieurs des paramètres de ce modèle, y compris certains des paramètres sur lesquels nous nous pencherons dans cet article : taille, forme et type de fenêtre de contexte ; pondération des statistiques ; choix d'une technique de sélection d'attributs ou de réduction de dimension. Ils se sont d'ailleurs intéressés à la question de la taille du corpus, et ont montré que la sélection de certains paramètres tels que la pondération et la mesure de similarité entre vecteurs peut exercer une influence particulièrement importante lorsque le corpus est de petite taille (en l'occurrence 4,6 millions de mots). Plus récemment, (Kiela & Clark, 2014) ont réalisé une évaluation systématique de la plupart des paramètres de ce modèle sur plusieurs jeux de données en utilisant des corpus de différentes tailles ; une des conclusions intéressantes de ce travail est que l'utilisation de contextes de nature syntaxique n'est pas forcément bénéfique, l'utilisation d'une fenêtre de cooccurrence étroite sur un gros corpus produisant de meilleurs résultats que les contextes syntaxiques sur la plupart des jeux de données utilisés pour l'évaluation. Soulignons finalement l'étude de (Padró *et al.*, 2014), qui compare quelques pondérations et mesures de similarité, et qui souligne l'influence importante du seuil de fréquence minimale utilisé pour choisir les mots-cibles du modèle.

Dans la section suivante, nous décrirons les ressources que nous avons utilisées. La section 3 portera sur la construction et les paramètres du modèle. Dans la section 4, nous décrirons la procédure de sélection des paramètres, qui repose sur une évaluation automatique. Les résultats obtenus seront analysés à la section 5 ; entre autres, nous y présenterons les résultats obtenus sur le corpus TALN et les comparerons à ceux obtenus sur le corpus du domaine de l'environnement.

2 Ressources utilisées

2.1 Corpus et prétraitements

Deux corpus ont été utilisés dans le cadre de ce travail. Le premier est le corpus TALN (Boudin, 2013), qui regroupe des articles parus dans les actes de TALN/RECITAL entre 2007 et 2013, totalisant environ 2 millions de mots. Puisqu'une version analysée syntaxiquement à l'aide de l'analyseur Talismane (Urieli & Tanguy, 2013) a été mise à la disposition des participants à SemDis, nous l'avons utilisée afin de reconstruire une version lemmatisée du corpus. Nous n'avons pas exploité les autres renseignements résultant de l'analyse, notamment les liens de dépendance syntaxique, puisque nous avons opté pour un modèle non structuré ; aucun prétraitement supplémentaire n'a été effectué.

Le deuxième corpus est le corpus monolingue français PANACEA – domaine de l'environnement (ELRA-W0065), un corpus de documents Web portant sur divers aspects du domaine de l'environnement. Ce corpus a été compilé au moyen de l'outil de construction automatique de corpus spécialisés conçu dans le cadre du projet PANACEA [1], et il est distribué librement à des fins de recherche [2]. Il contient plus de 23 000 documents totalisant plus de 47 millions de mots.

Le prétraitement de ce corpus se décline en plusieurs étapes. Nous avons d'abord extrait le contenu textuel des documents XML qui forment le corpus. Dans ces documents, un attribut (*crawlinfo*) indique, pour chaque segment textuel, s'il est dans une langue autre que celle du corpus, s'il est considéré comme trop court ou s'il correspond à du "boilerplate". Tous ces segments ont été supprimés, puis chaque document a été converti en texte ordinaire. Quelques opérations de normalisation ont ensuite été appliquées, portant sur les URL et adresses courriel, entre autres. Puis, le corpus a été lemmatisé à l'aide de TreeTagger (Schmid, 1994) [3].

Comme nous l'avons souligné dans l'introduction, les méthodes de la sémantique distributionnelle sont sensibles à la taille des corpus, donc il nous semblait important d'utiliser un corpus de taille comparable à celle du corpus TALN. À cette fin, nous avons extrait, au moyen d'une technique de recherche d'information [4], un sous-ensemble du corpus PANACEA portant sur les changements climatiques et les énergies renouvelables, deux thématiques importantes du dictionnaire dont nous avons extrait les données de référence (voir section 2.2) ; ce sous-corpus contient 1200 documents totalisant ~2,1 millions de tokens. Nous avions alors à notre disposition deux corpus spécialisés de taille comparable. De plus, pour le corpus du domaine de l'environnement, nous avons obtenu de données de référence pouvant servir à évaluer la qualité des modèles construits sur ce corpus, que nous décrirons à la section 2.2.

Bien que la taille des corpus soit comparable, il est important de noter qu'ils présentent des différences importantes à d'autres égards, notamment quant au niveau de spécialisation. Contrairement au corpus TALN, le corpus PANACEA est constitué de documents provenant de différentes sources : sites d'organismes gouvernementaux ou non gouvernementaux, sites de vulgarisation scientifique, encyclopédies, journaux, blogues et répertoires de sites Web, entre autres. De plus, une analyse sommaire d'un échantillon du corpus suggère que la majorité des documents ne sont pas destinés à des experts, bien que le public visé varie d'une source à l'autre. En outre, la taille des documents est extrêmement variable : dans le sous-corpus que nous avont extrait, le nombre de tokens varie d'une centaine à plusieurs dizaines de milliers, le nombre moyen de tokens par document étant ~1800.

2.2 Données de référence

Les données de référence que nous avons utilisées afin d'évaluer les modèles et de déterminer les paramètres optimaux ont été extraites du DiCoEnviro [5], un dictionnaire spécialisé du domaine de l'environnement élaboré à l'Observatoire de linguistique Sens-Texte. Le DiCoEnviro vise à décrire le sens et le fonctionnement des termes du domaine de l'environnement, en particulier du sous-domaine des changements climatiques, des énergies renouvelables et de la gestion des matières résiduelles, et à expliciter les différents liens qui existent entre ces termes.

1. http://panacea-lr.eu/
2. http://catalog.elra.info/product_info.php?products_id=1186&language=fr
3. Le fait d'utiliser des analyseurs différents pourrait avoir une incidence sur les résultats obtenus sur chaque corpus, mais nous supposons que celle-ci ne sera pas très importante, puisque nous n'utilisons les analyseurs que pour la lemmatisation.
4. Nous ne décrirons pas cette technique, puisqu'elle n'entre pas dans les objectifs de cet atelier.
5. En construction. Le dictionnaire peut être consulté à l'adresse http://olst.ling.umontreal.ca/cgi-bin/dicoenviro/search_enviro.cgi.

Les entrées du DiCoEnviro appartiennent à différentes parties du discours, à savoir le nom, le verbe, l'adjectif, ainsi que certaines locutions ; par exemple, il contient des articles pour les termes *biodiversité, climat, climatique, composter, gaz à effet de serre, polluer* et *polluant*. Ces termes sont repérés dans un corpus spécialisé en fonction des critères lexico-sémantiques de sélection de termes proposés par (L'Homme, 2004) ; il est important de noter que le corpus utilisé pour la compilation du dictionnaire est distinct du corpus PANACEA utilisé dans cette étude, que nous utilisons parce qu'il est distribué librement. Les différentes acceptions d'un même terme, distinguées au moyen de tests lexico-sémantiques, ont chacune leur propre article, mais dans le cadre de l'expérience que nous avons réalisée, nous ne faisons pas de distinction entre les différentes acceptions d'un terme.

La fiche de chaque terme contient sa structure actancielle ainsi que de nombreux liens lexicaux, qui peuvent être de nature syntagmatique ou paradigmatique. Dans ce travail, nous nous sommes intéressés à certaines relations paradigmatiques précises, à savoir :

— les quasi-synonymes et autres sens voisins (p. ex. *extinction* → *disparition*, *pollueur* → *polluant*) ;
— les antonymes ou sens contraires (p. ex. *réchauffement* → *refroidissement*) ;
— les hyponymes ou sortes de (p. ex. *activité* → *agriculture*).

Pour chaque terme faisant l'objet d'un article dans le dictionnaire, nous avons extrait tous les termes voisins entretenant avec l'entrée une de ces trois relations.

Les relations d'hyponymie ("sortes de") ont été divisées en deux catégories de la façon suivante : si le terme voisin est un terme complexe qui contient le terme faisant l'objet de l'article (l'entrée) ainsi qu'un modificateur, nous considérons qu'il s'agit plutôt d'une relation syntagmatique entre le terme en entrée et le modificateur, à savoir une collocation ; si le terme voisin ne contient pas l'entrée, nous considérons qu'il s'agit d'une relation paradigmatique entre les deux termes. Ainsi, la paire <énergie, énergie hydroélectrique> a été exclue des données de référence, tandis que la paire <carburant, biogaz> a été retenue.

Par ailleurs, nous avons exclu tous les termes complexes, qu'il s'agisse de l'entrée de l'article ou du terme voisin. Nous avons ainsi obtenu une liste de paires <entrée, terme relié> constituées de deux termes simples participant à une relation paradigmatique (sens voisin, contraire ou sorte de). Parmi les paires extraites, nous avons éliminé celles où un des deux termes n'était pas inclus dans le vocabulaire utilisé pour construire le modèle (voir section 3.1), ce qui représentait environ 15% des paires. Restaient environ 630 paires [6]. Parmi celles-ci, nous en avons conservé 600 choisies au hasard, dont 400 ont servi pour faire la sélection des paramètres du modèle, et 200 ont été réservées pour une évaluation finale du meilleur modèle, ainsi qu'une analyse manuelle des résultats.

3 Construction du modèle

Pour nos expériences, nous utilisons le modèle Hyperspace Analogue to Language, ou HAL (Lund *et al.*, 1995; Lund & Burgess, 1996). HAL fait partie de la famille des modèles dits *non structurés*, qui n'exploitent pas d'information syntaxique. Dans ce modèle, la représentation vectorielle d'un mot est basée sur la fréquence à laquelle d'autres mots apparaissent près de lui dans un corpus ; on appellera les mots pour lesquels on construit des vecteurs *mots-cibles* et ceux qui servent d'attributs *mots-contextes*. Ainsi, des mots partageant des cooccurrents auront une représentation semblable. Une mesure de similarité est ensuite utilisée pour comparer les vecteurs et calculer leur distance dans l'espace sémantique que définit le modèle HAL.

En ce qui concerne la notation utilisée dans la suite de cet article, la matrice de cooccurrence qui contient les représentations vectorielles des mots-cibles sera dénotée par $\mathbf{X}$. Le vocabulaire sera noté W et sera indexé par i lorsque nous désignons un mot-cible et par j lorsque nous désignons un mot-contexte ; le nombre de mots dans le vocabulaire sera noté m. Les vecteurs des mots-cibles et mots-contextes seront donc dénotés respectivement par $\mathbf{x}_{i:}$ et $\mathbf{x}_{:j}$, et les cellules de la matrice par x_{ij}.

6. Les données ont été récupérées au début mars 2014. Le nombre de relations décrites dans le DiCoEnviro augmente à mesure qu'il est enrichi.

La matrice de cooccurrence $\mathbf{X}$ est construite en plaçant une fenêtre de contexte autour de chaque occurrence d'un mot-cible et en incrémentant chaque fois, dans le vecteur du mot-cible, la fréquence de cooccurrence des autres mots qui se trouvent à l'intérieur de la fenêtre. Dans la matrice $\mathbf{X}$, chaque cellule x_{ij} indique donc la fréquence à laquelle le mot-contexte w_j apparait dans la fenêtre de contexte du mot-cible w_i. L'incrémentation de x_{ij} peut être pondérée par l'inverse de la distance entre w_i et w_j dans un contexte donné [7] ; dans ce cas, nous dirons que la fenêtre de contexte est *triangulaire*, suivant (Bullinaria & Levy, 2007). En revanche, dans une fenêtre de contexte *rectangulaire*, la fréquence de tous les mots-contextes dans la fenêtre est incrémentée de 1.

3.1 Sélection du vocabulaire

Chaque mot dans le vocabulaire pour lequel nous construisons l'espace sémantique correspond à la fois à une rangée ($\mathbf{x}_{i:}$) et à une colonne ($\mathbf{x}_{:j}$) de la matrice de cooccurrence. L'ensemble des mots-cibles est donc le même que celui des mots-contextes [8]. Nous déterminons ce vocabulaire (W) en fonction de la fréquence globale des mots dans le corpus. Il est courant de déterminer le vocabulaire au moyen d'un seuil fixe de fréquence, souvent fixé à 100 (Anguiano & Denis, 2011) ; nous utilisons un critère de sélection de vocabulaire semblable, mais qui dépend moins de la taille du corpus ; en effet, les mots ayant au moins 100 occurrences dans les corpus que nous utilisons sont peu nombreux. Nous éliminons d'abord des mots vides au moyen d'une liste d'exclusion, ainsi que les chaînes qui ne sont pas constituées exclusivement de caractères alphabétiques. Parmi les mots restants, nous conservons les m mots les plus fréquents. Ce nombre a été fixé de sorte à assurer une bonne couverture des données qui serviraient à l'évaluation du modèle. En conservant les 5000 mots les plus fréquents, seulement ~15% des paires extraites du DiCoEnviro (voir section 2.2) contenaient un mot absent de W. Le vocabulaire est donc de taille relativement petite, car on utilise fréquemment des vocabulaires de plusieurs dizaines de milliers de mots ou plus, mais il offre une bonne couverture des termes décrits dans la ressource utilisée pour l'évaluation ; d'ailleurs, les corpus spécialisés de petite taille contiennent beaucoup moins de formes distinctes que les corpus contenant des centaines de millions de mots.

3.2 Forme, type et taille de la fenêtre de contexte

Comme nous l'avons souligné ci-dessus, la fenêtre de contexte peut avoir une forme rectangulaire ou triangulaire [9]. Les fenêtres se distinguent également selon qu'on prend en compte le contexte à gauche du mot-cible, le contexte à droite ou les deux. HAL exploite une fenêtre de contexte dite *directionnelle* (Sahlgren, 2006) : lors de la construction de la matrice de cooccurrence, seuls les cooccurrents à gauche du mot-cible sont comptabilisés, de sorte que pour chaque mot $w_i \in W$, la rangée $\mathbf{x}_{i:}$ indique la fréquence à laquelle chaque mot-contexte apparait avant w_i, et la colonne $\mathbf{x}_{:i}$ indique la fréquence à laquelle les mots-contextes apparaissent après w_i. Une fois cette matrice construite, on concatène $\mathbf{x}_{i:}$ et $\mathbf{x}_{:i}$, ce qui produit un vecteur de dimension $2m$ contenant la fréquence de cooccurrence dans le contexte à gauche de w_i ainsi que celle dans le contexte à droite de w_i.

La fenêtre de contexte peut aussi être symétrique, comme dans le modèle proposé par (Schütze, 1992), qui a précédé HAL ; dans ce cas, aucune distinction n'est faite entre les cooccurrents apparaissant à gauche et à droite du mot-cible. Si on construit la matrice initiale de la manière décrite ci-dessus, en n'observant que les cooccurrents à gauche du mot-cible, on peut obtenir un contexte symétrique en prenant la somme (plutôt que la concaténation) de $\mathbf{x}_{i:}$ et $\mathbf{x}_{:i}$ pour chaque mot-cible w_i ; la dimension des vecteurs résultants est donc m plutôt que $2m$. De plus, il est possible de n'utiliser que les cooccurrents à gauche ($\mathbf{x}_{i:}$) ou seulement ceux à droite ($\mathbf{x}_{:i}$). (Bullinaria & Levy, 2007) appellent ces quatre types de contexte *gauche&droite*, *gauche+droite*, *gauche* et *droite* respectivement.

Enfin, la taille de la fenêtre de contexte a une influence considérable sur les résultats obtenus. Nous vérifierons l'influence de la forme, du type et de la taille de la fenêtre de contexte dans l'expérience décrite ci-dessous.

Soulignons finalement que nous permettons à la fenêtre de contexte de sauter les frontières de phrases, et que les mots dans le corpus qui ne font pas partie du vocabulaire W ne sont pas supprimés ; ils ne sont tout simplement pas comptabilisés lors de la construction de la matrice de cooccurrence.

7. D'autres pondérations en fonction de la distance sont possibles ; par exemple, (Sahlgren, 2006) utilise 2^{1-L} au lieu de $\frac{1}{L}$, où L est la distance entre les 2 mots.
8. Il serait possible de définir ces deux vocabulaires de façons distinctes, mais il est courant d'utiliser un seul et même vocabulaire.
9. D'autres formes sont possibles, telle qu'une fenêtre gaussienne.

3.3 Pondération des fréquences

La matrice de cooccurrence contient, pour chaque paire de mots w_i et w_j, la fréquence à laquelle w_j co-occurre avec w_i. Ces fréquences de cooccurrence peuvent être pondérées de différentes façons, notamment pour diminuer l'influence des mots-contextes très fréquents, mais peu discriminants.

Une pondération simple, que nous appelerons DAMP, consiste à prendre le logarithme de la fréquence :

$$\text{DAMP}(x_{ij}) = \log(x_{ij} + 1)$$

(Lavelli *et al.*, 2004) décrivent une variante de TF-IDF pour les modèles exploitant une matrice de cooccurrence plutôt qu'une matrice terme-document, qu'ils appellent TF-ITF. Nous avons implémenté une version légèrement modifiée de cette pondération, que nous formulons de la façon suivante :

$$\text{TF-ITF}(x_{ij}) = \log(x_{ij} + 1) \cdot \log \frac{m}{||\mathbf{x}_{:j}||_0}$$

où m est la taille du vocabulaire et $||\mathbf{x}_{:j}||_0$ est le nombre d'éléments non nuls dans la colonne $\mathbf{x}_{:j}$, autrement dit le nombre de mots-cibles avec lesquels le mot-contexte w_j co-occurre au moins une fois.

Une pondération particulièrement efficace selon (Bullinaria & Levy, 2007) est la Positive Pointwise Mutual Information (PPMI), que nous formulons de la façon suivante, suivant (Turney & Pantel, 2010) :

$$\begin{aligned}
p_{ij} &= \frac{x_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{m} x_{ij}} \\
p_{i:} &= \frac{\sum_{j=1}^{m} x_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{m} x_{ij}} \\
p_{:j} &= \frac{\sum_{i=1}^{m} x_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{m} x_{ij}} \\
\text{PMI}(x_{ij}) &= \log \frac{p_{ij}}{p_{i:} \cdot p_{:j}} \\
\text{PPMI}(x_{ij}) &= \begin{cases} \text{PMI}(x_{ij}) & \text{si PMI}(x_{ij}) > 0. \\ 0 & \text{sinon.} \end{cases}
\end{aligned}$$

où p_{ij} estime la probabilité que w_j co-occurre avec w_i, $p_{i:}$ estime la probabilité du mot-cible w_i et $p_{:j}$ estime la probabilité du mot-contexte w_j.

3.4 Sélection d'attributs ou réduction de dimension

Dans le modèle HAL original, la dimension des représentations vectorielles des mots-cibles est réduite en éliminant les colonnes à faible variance pour n'en conserver que quelques centaines ; il est également possible de faire la sélection d'attributs en fonction d'autres critères, tels que la fréquence du mot-contexte, mais ces deux critères étant corrélés (Bullinaria & Levy, 2007, p. 519), ils produiraient des résultats semblables.

Nous évaluons l'influence de cette technique et la comparons à une technique de réduction de dimension appelée décomposition en valeurs singulières [10] (SVD), qu'exploite notamment une autre méthode de sémantique distributionnelle, la LSA (Landauer & Dumais, 1997). (Schütze, 1992) décrit l'utilisation de la SVD sur une matrice de cooccurrence semblable à celle qu'exploite HAL ; cette technique n'améliore pas les résultats qu'il obtient, mais l'auteur s'en sert tout de même pour réduire la dimension des représentations de mots et accélérer leur traitement subséquent. Nous vérifierons si la SVD permet d'obtenir de meilleurs résultats sur les données que nous utilisons.

10. Nous utilisons l'implémentation de la SVD (algorithme ARPACK) offerte dans le toolkit scikit-learn (Pedregosa *et al.*, 2011) pour Python.

3.5 Distance ou mesure de similarité

Enfin, une distance ou une mesure de similarité est utilisée pour comparer les vecteurs et déterminer leur proximité dans l'espace sémantique. (Lund & Burgess, 1996) utilisent des distances de la famille Minkowski (Manhattan, euclidienne, etc.). Nous avons plutôt opté, comme (Schütze, 1992), pour le cosinus de l'angle des vecteurs, une mesure de similarité courante dans le domaine de la sémantique distributionnelle. En outre, (Bullinaria & Levy, 2007) montrent que le cosinus produit les meilleurs résultats sur plusieurs tâches, particulièrement lorsqu'on pondère les fréquences au moyen de l'information mutuelle, ce qui concorde d'ailleurs avec les observations de (Ferret, 2010).

4 Évaluation automatique et sélection des paramètres

Nous avons réalisé une expérience visant à déterminer la valeur optimale de certains des paramètres du modèle HAL, à savoir la fenêtre de contexte (forme, type et taille), la pondération des statistiques et la réduction de dimension ou sélection d'attributs. Nous n'avons pas évalué l'influence d'autres facteurs tels que les prétraitements linguistiques (parce que nous nous intéressons ici à un modèle non structuré, qui exige peu de prétraitement) ou la distance ou mesure de similarité entre vecteurs, étant donné que plusieurs travaux ont montré que le cosinus est une mesure de similarité efficace en ce qui concerne les modèles distributionnels.

Les modèles ont été construits sur le corpus du domaine de l'environnement et évalués sur les données de référence décrites à la section 2.2. Les paramètres pouvaient prendre les valeurs suivantes :

- Taille de la fenêtre de contexte : entre 1 et 15 mots (un contexte gauche de 2 mots signifie qu'on observe les 2 mots à gauche du mot-cible ; un contexte gauche&droite de 2 mots signifie qu'on observe 2 mots à gauche et 2 mots à droite).
- Forme de la fenêtre de contexte : triangulaire (TRI) ou rectangulaire (RECT).
- Type de fenêtre de contexte : gauche&droite (G&D), gauche+droite (G+D), gauche seulement (G) ou droite seulement (D).
- Pondération : aucune, DAMP, TF-ITF ou PPMI.
- Réduction :
  - Sélection d'attributs par variance (SEL) avec nombre d'attributs $\in \{500, 1000, \ldots, 4500\}$; ce nombre est doublé dans le cas de la fenêtre de contexte G&D.
  - SVD avec nombre de composantes $\in \{50, 100, \ldots, 500\}$.
  - Aucune.

4.1 Évaluation automatique

La sélection des paramètres a été réalisée au moyen d'une évaluation automatique, la tâche consistant à prédire le terme relié dans chacune des 400 paires <entrée, terme relié> étant donné l'entrée. La mesure utilisée pour comparer les modèles est le rappel au rang n (nous utiliserons parfois l'abréviation $R@n$). Le rappel au rang n correspond au pourcentage des paires <entrée, terme relié> pour lesquelles le terme relié se trouve parmi les n plus proches voisins de l'entrée selon le modèle. Ainsi, le rappel au rang 1 (R@1) correspond au pourcentage des exemples pour lesquels le terme relié correspond au plus proche voisin (PPV) de l'entrée, et le rappel au rang 10 (R@10), au pourcentage des exemples pour lesquels le terme relié est parmi ses 10 plus proches voisins. Bien que cette mesure est exprimée sous la forme d'un pourcentage, elle ne peut pas toujours atteindre 100 % (notamment au rang 1) puisqu'il y a souvent plus d'un terme relié par entrée. Or, étant donné que les données de référence contiennent généralement 1 ou 2 termes reliés par entrée (c'est le cas pour ~70 % des entrées) et que le nombre de termes reliés par entrée varie considérablement (de 1 à 8), il nous semble préférable d'utiliser le rappel plutôt que la précision pour comparer les modèles.

Le meilleur rappel au rang 1, de 17,25%, a été atteint par cinq modèles, présentés dans le Tableau 1. Il est intéressant de noter est que ces modèles exploitent tous la pondération TF-ITF ; de plus, quatre de ces modèles exploitent un contexte symétrique (G+D). En revanche, les modèles qui maximisent le rappel au rang 10, présentés dans le Tableau 2, exploitent tous un contexte G&D et la pondération PPMI. Dans les deux cas, tous les meilleurs modèles exploitent la réduction par SVD et un contexte étroit, de deux ou trois mots.

Fenêtre			Pondération	Réduction (nb dimensions)	R@10	R@1
Taille	Forme	Type				
3	TRI	G&D	TF-ITF	SVD (350)	49,75	17,25
2	RECT	G+D	TF-ITF	SVD (250)	48	17,25
2	RECT	G+D	TF-ITF	SVD (150)	47,75	17,25
2	RECT	G+D	TF-ITF	SVD (200)	47,25	17,25
2	RECT	G+D	TF-ITF	SVD (300)	46,75	17,25

TABLE 1 – 5 meilleurs modèles (triés en fonction du rappel au rang 1).

Fenêtre			Pondération	Réduction (nb dimensions)	R@10	R@1
Taille	Forme	Type				
2	RECT	G&D	PPMI	SVD (250)	54	16,5
2	TRI	G&D	PPMI	SVD (300)	54	15,75
2	TRI	G&D	PPMI	SVD (250)	53,75	16,5
2	RECT	G&D	PPMI	SVD (300)	53	15,5
2	RECT	G&D	PPMI	SVD (400)	53	15,25

TABLE 2 – 5 meilleurs modèles (triés en fonction du rappel au rang 10).

Étant donné les différences observées entre les modèles qui maximisent R@1 et ceux qui maximisent R@10, nous avons cherché à vérifier si certaines paramétrisations seraient plus adaptées pour des relations spécifiques parmi les trois relations ciblées. Nous avons donc comparé, au moyen de l'évaluation automatique, deux paramétrisations identiques sauf en ce qui concerne le type de fenêtre de contexte et la pondération (les 2 paramètres qui semblent favoriser soit R@1 soit R@10). Les deux modèles exploitent une fenêtre de contexte rectangulaire de deux mots et la réduction par SVD (250 composantes). Le premier modèle, un des cinq qui maximisent R@1, exploite une fenêtre G+D et la pondération TF-ITF. L'autre modèle, qui maximise R@10, exploite une fenêtre G&D et la pondération PPMI. Pour chaque modèle, nous avons calculé R@1 et R@10 sur les trois sous-ensembles des 400 paires <entrée, terme relié> correspondant aux trois relations possibles entre l'entrée et le terme relié : sorte de (23 paires), contraire (103 paires) et sens voisin (274 paires).

Les résultats de cette comparaison, présentés dans le Tableau 3, ne suggèrent pas qu'une des deux paramétrisations est particulièrement adaptée à une des trois relations : les deux modèles captent mieux les sens voisins que les contraires et les contraires mieux que les sortes de (peut-être parce que les paramètres ont été optimisés sur des données qui comprennent plus de sens voisins que de contraires et plus de contraires que de sortes de). De plus, pour toutes les relations, le premier modèle maximise le rappel au rang 1, et le deuxième modèle, le rappel au rang 10. Nous examinerons systématiquement l'influence des paramètres à la section 5.1, mais sans faire de distinction entre les trois relations.

Relation entre entrée et terme relié	R@1 (%)		R@10 (%)	
	Modèle A	Modèle B	Modèle A	Modèle B
Sorte de	**13,04**	8,70	34,78	**43,48**
Contraire	**14,56**	12,62	42,72	**47,57**
Sens voisin	**18,61**	**18,61**	51,09	**57,30**

TABLE 3 – Évaluation en fonction de la relation entre l'entrée et le terme relié. Le modèle A exploite une fenêtre G+D et la pondération TF-ITF. Le modèle B exploite une fenêtre G&D et la pondération PPMI.

5 Analyse des résultats

Dans cette section, nous examinerons l'influence de divers paramètres du modèle, puis nous présenterons les résultats d'une évaluation manuelle du meilleur modèle, enfin nous construirons un modèle sur le corpus TALN en utilisant les mêmes paramètres et comparerons les résultats à ceux obtenus sur le corpus du domaine de l'environnement.

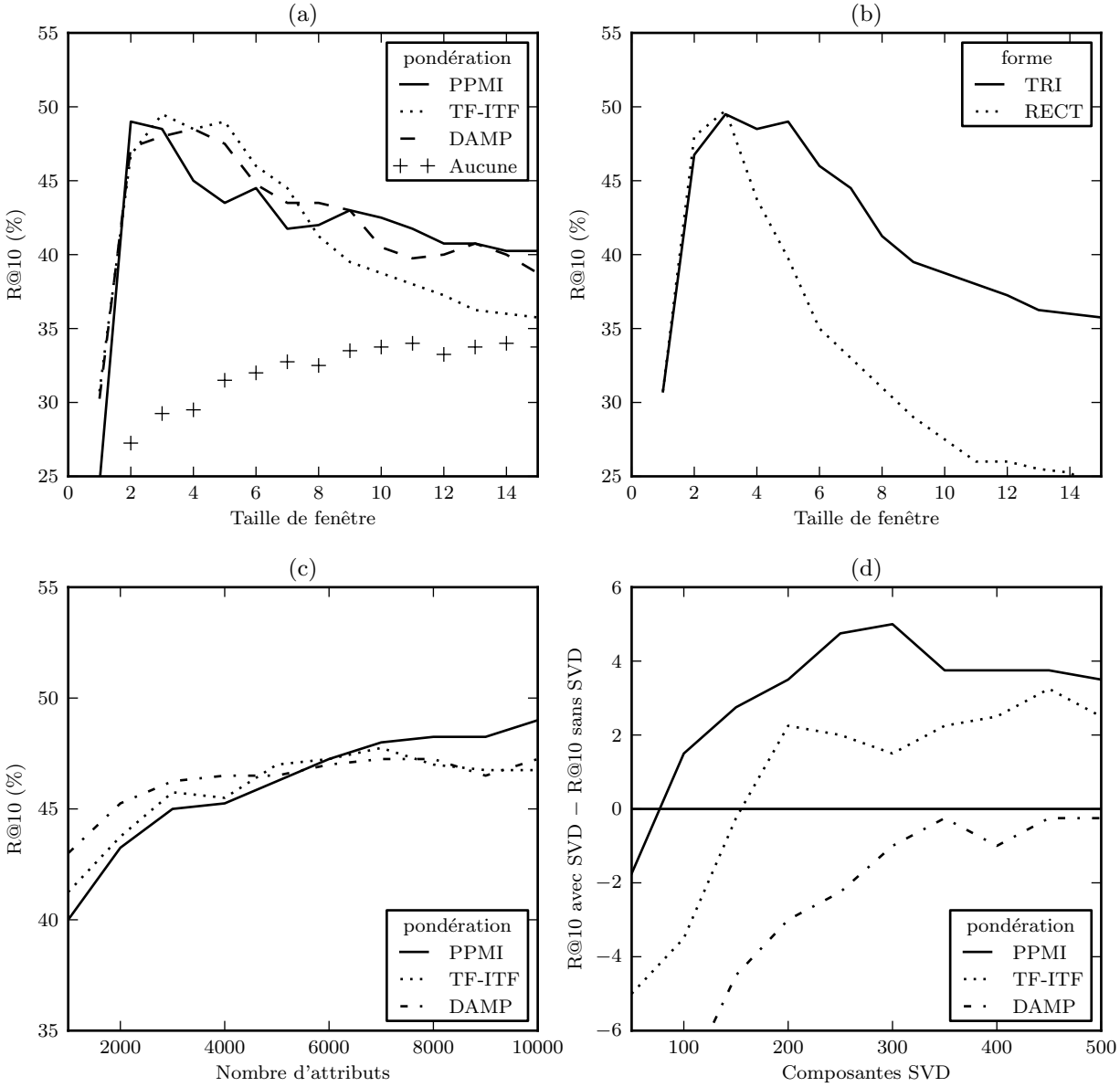

FIGURE 1 – Influence des paramètres du modèle. **(a)** Influence de la taille de fenêtre et de la pondération (paramètres fixes : fenêtre G&D triangulaire, aucune réduction de dimension). **(b)** Influence de la forme de la fenêtre de contexte (fenêtre G&D, pondération TF-ITF, aucune réduction). **(c)** Influence de la sélection d'attributs : rappel au rang 10 en fonction du nombre d'attributs conservés (fenêtre G&D triangulaire de 2 mots). **(d)** Augmentation du rappel au rang 10 lorsqu'on applique la réduction par SVD, en fonction du nombre de composantes (fenêtre G&D triangulaire de 2 mots).

5.1 Influence des paramètres

La Figure 1 illustre l'influence de certains paramètres du modèle ; les paramètres qui ne varient pas (p. ex. le type et la forme de la fenêtre de contexte dans le cas de la Figure 1-a) sont généralement ceux qui maximisent soit R@1 soit R@10 ; par contre, dans les Figures 1-a et 1-b, aucune réduction de dimension n'est appliquée, afin d'illustrer plus clairement l'influence des autres paramètres. La Figure 1-a montre que les 3 pondérations implémentées sont très efficaces, et qu'il n'y a pas une différence très importante entre le meilleur résultat atteint par chacune des pondérations. Il est aussi intéressant de noter que lorsque les fréquences ne sont pas pondérées, plus le contexte est large, plus la précision augmente ; en revanche, lorsque les fréquences sont pondérées, une fenêtre étroite (de 2 à 4 mots selon la pondération) donne les meilleurs résultats. Ces résultats concordent avec ceux de (Ferret, 2010) et de (Bullinaria & Levy, 2007), ces derniers obtenant les meilleurs résultats sur 3 des 4 tâches utilisées pour l'évaluation au moyen d'une fenêtre symétrique rectangulaire de taille 1 (pour l'anglais) et de la pondération PPMI.

La Figure 1-b montre que lorsque la fenêtre est triangulaire, la précision diminue moins rapidement à mesure qu'on augmente la taille du contexte, mais qu'elle n'améliore pas le meilleur résultat. Cette figure montre le cas où les fréquences sont pondérées par TF-ITF ; lorsque nous appliquons la pondération PPMI, nous observons la même tendance, mais la différence entre les deux courbes est moins importante. Ces résultats concordent avec ceux de (Bullinaria & Levy, 2007), qui observent que les fenêtres triangulaires ont tendance à produire des résultats similaires à ceux qu'on obtient avec des fenêtres rectangulaires de plus petite taille.

Les Figures 1-c et 1-d concernent la réduction de dimension. La Figure 1-c montre qu'il est possible d'éliminer plusieurs milliers d'attributs à faible variance tout en maintenant une précision élevée, mais que cette technique ne permet pas d'augmenter la précision d'une manière significative (du moins pas dans le cas d'une fenêtre de contexte de taille 2). Enfin, la Figure 1-d montre que la SVD permet, dans certains cas, d'améliorer la précision tout en diminuant la dimension des vecteurs. Par contre, nous avons observé que le nombre optimal de composantes varie beaucoup en fonction des autres paramètres du modèle, notamment la taille de fenêtre ; de plus, l'amélioration observée diminue lorsqu'on augmente la taille de fenêtre. Dans certains cas (notamment lorsque la pondération DAMP est utilisée, comme le montre la figure), la SVD diminue la précision. Cette technique de réduction ne semble donc pas très robuste, mais soulignons de nouveau que nos meilleurs modèles exploitent tous la SVD.

Nous ne montrons pas ici l'influence du type de fenêtre de contexte, mais soulignons que la fenêtre G&D maximise à la fois les mesures R@1 et R@10, bien que la fenêtre G+D atteigne également le meilleur rappel au rang 1.

5.2 Évaluation manuelle

L'évaluation automatique décrite à la section 4.1 estime la capacité d'un modèle à capter des relations lexico-sémantiques paradigmatiques à partir d'un corpus. Or, étant donné que les données de référence n'offrent pas une couverture complète de toutes les relations paradigmatiques qu'il serait possible de repérer au sein de ce corpus, nous avons procédé à une évaluation manuelle des voisins identifiés par le modèle ayant produit les meilleurs résultats lors de l'évaluation automatique, plus précisément celui qui maximise le rappel au rang 10. Ce modèle exploite une fenêtre G&D rectangulaire de 2 mots, la pondération PPMI et la réduction de dimension par SVD (250 composantes).

L'évaluation manuelle a été effectuée sur 200 paires <entrée, terme relié> qui n'ont pas servi lors de la sélection des paramètres. Dans un premier temps, nous avons vérifié si le modèle offrait une précision aussi élevée sur les nouvelles données de référence, au moyen de l'évaluation automatique. Les résultats obtenus sur ces 200 paires étaient de R@1 = 12,5% et R@10 = 47%. On observe donc une légère baisse par rapport aux résultats obtenus sur les 400 paires utilisées pour la sélection des paramètres.

Puis, nous avons évalué manuellement la précision des voisins identifiés par le modèle sur ces 200 exemples, en observant le plus proche voisin pour chaque entrée. D'abord, l'évaluation automatique indique que pour 25 des 200 exemples, le PPV correspond au terme relié. Or, puisque le DiCoEnviro contient souvent plus d'un terme relié paradigmatiquement pour une entrée donnée, il se peut que le voisin soit valide même s'il ne correspond pas au terme relié inclus dans une paire particulière. C'est effectivement le cas pour 54 des 200 exemples. Donc, pour 79 des 200 exemples, le PPV est effectivement un terme relié paradigmatiquement selon les données de référence.

Comme nous l'avons souligné, les données de référence utilisées ne peuvent pas offrir une couverture complète des termes de l'environnement et des relations lexico-sémantiques auxquelles ils participent. Il se peut donc que la précision soit plus élevée que le suggère l'évaluation automatique, et que les PPV considérés comme incorrects soient en fait des termes reliés qui pourraient être ajoutés aux données de référence. Pour cette raison, les 121 exemples restants ont fait l'objet d'une évaluation manuelle.

Le critère utilisé pour juger la validité d'un PPV est le suivant : si au moins une des acceptions de l'entrée et au moins un des sens du PPV participent à une relation valide (sens voisin, contraire, sorte de), le voisin est valide. Au lieu de juger la pertinence des PPV de façon binaire (oui/non), nous avons défini trois jugements possibles : le PPV est valide, il participe avec l'entrée à un autre type de relation (notamment des relations syntagmatiques, de dérivation ou de méronymie) ou il n'est pas pertinent du tout. Des exemples illustrant ces deux derniers jugements sont présentés dans le Tableau 4.

<**entrée, terme relié**>	**PPV**	**Jugement**
<globe, monde>	mer	Le PPV et l'entrée participent à un autre type de relation (méronymie).
<jeter, recycler>	verre	Le PPV et l'entrée participent à un autre type de relation : *verre* est la réalisation d'un des actants de *jeter*.
<influer, peser>	influent	Le PPV et l'entrée participent à un autre type de relation (dérivation).
<météorologique, climatique>	extrême	Le PPV n'est pas pertinent. *météorologique* et *extrême* modifient les mêmes noms, mais *extrême* ne serait pas décrit dans l'article de *météorologique*.
<localement, globalement>	normalement	Le PPV n'est pas pertinent.
<amplification, intensification>	rouille	Le PPV n'est pas pertinent.

TABLE 4 – Exemples illustrant l'évaluation manuelle (PPV signifie plus proche voisin).

L'évaluation des voisinages a été confiée à une terminologue élaborant des dictionnaires et a été réalisée en fonction de l'intérêt qu'ils peuvent présenter du point de vue de leur description dans le dictionnaire du domaine de l'environnement. C'est d'ailleurs pour mieux représenter l'intérêt que présentent les voisinages identifiés qu'une catégorie intermédiaire (autres relations) a été prise en compte lors de l'évaluation manuelle. Par exemple, comme le montre le Tableau 4, *verre* serait ajouté dans le dictionnaire comme réalisation d'un des actants de *jeter* (il serait également décrit dans l'article de *recycler*). Il présenterait donc un certain intérêt pour le terminologue qui élabore l'article du verbe *jeter*, bien que la relation qu'il entretient avec ce verbe ne fasse pas partie des relations ciblées dans le cadre de ce travail. En revanche, même si *extrême*, *météorologique* et *climatique* peuvent modifier le même type de nom (p. ex. *évènement, phénomène*) et pourraient tous apparaître dans l'article de ces noms, *extrême* ne serait pas décrit dans l'article de *météorologique* (ni dans celui de *climatique*, par ailleurs) ; il est donc considéré comme incorrect.

Situation	**Nombre d'exemples**
Le PPV est un des termes reliés	79
Le PPV est valide, mais n'est pas dans le dictionnaire	71
Le PPV est relié à l'entrée, mais par un autre type de relation	31
Le PPV n'est pas pertinent	19
Total	200

TABLE 5 – Résultats de l'évaluation manuelle.

Les résultats de l'évaluation manuelle sont résumés dans le Tableau 5. Ces résultats suggèrent que, si l'on ne prend en considération que le plus proche voisin de chaque entrée, le niveau de bruit se situerait soit autour de 10%, soit autour de 25% si on ne considère pas comme valides les voisins qui participent avec l'entrée à un autre type de relation lexico-sémantique. La présence de voisinages non pertinents est liée à plusieurs facteurs, notamment la fréquence relative des mots-cibles (Weeds *et al.*, 2004; Ferret, 2010). La polysémie est un autre facteur qui pourrait expliquer certains voisinages non pertinents. Par exemple, l'article du terme *vert* dans le DiCoEnviro donne comme termes reliés *écologique*, *environnemental* et *propre* ; en revanche, les voisins identifiés par le modèle comprennent *forestier* et *agricole*, indiquant

un sens différent (caractérisé par une quantité importante de végétation). Le modèle HAL, qui apprend une seule représentation prototypique par mot-cible, ne permet pas de modéliser explicitement les différents sens d'un mot, mais il existe différentes techniques pour ce faire, telles que le modèle à prototypes multiples proposé par (Reisinger & Mooney, 2010).

5.3 Comparaison avec le corpus TALN

Après avoir réalisé la sélection des paramètres du modèle, nous avons construit un modèle identique sur le corpus TALN et extrait les voisins des 8 mots à l'étude dans le cadre de la 2e tâche de cette édition de SemDis. Le Tableau 6 présente les 10 plus proches voisins de ces 8 mots. Ces voisins comprennent une quantité importante de quasi-synonymes ou sens voisins ainsi que des antonymes (*complexe* → *simple*) et des méronymes (*graphe* → *noeud*), entre autres.

Il est intéressant de noter le cas du terme *sémantique*, qui est ambigu quant à sa partie du discours : 9 des 10 voisins de ce terme sont reliés sémantiquement à l'adjectif, tandis que le dernier voisin, *sens*, pourrait être interprété comme un quasi-synonyme du nom *sémantique*.

calculer	complexe	précis	fréquence	méthode	trait	sémantique	graphe
estimer	long	riche	probabilité	algorithme	élément	syntaxique	arbre
mesurer	simple	détaillé	poids	approche	indice	lexical	réseau
obtenir	fréquent	général	proportion	stratégie	attribut	morphologique	grammaire
déterminer	rare	spécifique	longueur	technique	catégorie	linguistique	dépendance
évaluer	court	particulier	valeur	système	structure	conceptuel	noeud
définir	difficile	fin	score	procédure	étiquette	grammatical	lexique
comparer	riche	systématique	nombre	processus	classe	temporel	structure
pondérer	spécifique	complet	distance	modèle	propriété	fonctionnel	automate
maximiser	utile	strict	taille	tâche	information	formel	vecteur
combiner	proche	exact	coût	méthodologie	forme	sens	transducteur

TABLE 6 – Voisins obtenus pour les 8 mots à l'étude en utilisant le corpus TALN.

Parmi les 8 mots à l'étude, 5 sont également présents dans le vocabulaire du modèle construit sur le corpus du domaine de l'environnement : *calculer, complexe, précis, fréquence* et *méthode*. Les 10 plus proches voisins de ces mots sont présentés dans le Tableau 7. La présence de ces mots dans les 2 vocabulaires pourrait indiquer qu'ils appartiennent à ce que l'on a appelé le *vocabulaire général d'orientation scientifique* (Phal, 1971) ou le *lexique scientifique transdisciplinaire* (Tutin, 2007; Drouin, 2007). Certains de ces mots semblent avoir le même sens dans les deux domaines ; par exemple, le verbe *calculer* a des voisins très semblables dans les 2 modèles. En revanche, le terme *fréquence* présente des voisins très différents, ce terme étant associé à la notion d'évènements météorologiques extrêmes dans un domaine, et au nombre d'occurrences d'une unité linguistique dans l'autre.

calculer	complexe	précis	fréquence	méthode
mesurer	physique	détailler	intensité	technologie
déterminer	simple	quantitatif	accentuation	procédé
évaluer	biologique	clair	multiplication	technique
exprimer	régir	contraignant	survenue	outil
simuler	déterminant	fiable	violence	méthodologie
atteindre	difficile	rigoureux	cas	pratique
prédire	chimique	complet	épisode	mode
comptabiliser	naturel	relatif	occurrence	système
estimer	essentiel	ambitieux	sécheresse	dispositif
comparer	écologique	spécifique	gravité	approche

TABLE 7 – Voisins obtenus pour 5 des mots à l'étude en utilisant le corpus du domaine de l'environnement.

6 Conclusion

Dans cet article, nous avons mis en application une technique d'analyse distributionnelle afin d'identifier des relations lexico-sémantiques à partir de corpus spécialisés de petite taille. De plus, nous avons systématiquement optimisé certains des paramètres du modèle afin de cibler une famille spécifique de relations. Ces paramètres, qui concernent la fenêtre de contexte, la pondération des statistiques et la réduction de dimension, ont été optimisés au moyen d'une évaluation automatique exploitant des données extraites d'un dictionnaire spécialisé.

Les résultats de l'expérience que nous avons réalisée montrent qu'un modèle non structuré permet d'identifier, pour un terme donné, des termes reliés sur le plan paradigmatique à partir d'un corpus spécialisé. Une évaluation manuelle a montré que le modèle capte bien les relations de quasi-synonymie, d'antonymie et d'hyponymie décrites dans le dictionnaire dont nous avons extrait les données de référence, et qu'il pourrait servir à l'enrichir. Les voisinages observés comprennent aussi, mais dans une plus faible proportion, d'autres types de relations lexico-sémantiques, notamment des relations syntagmatiques, de dérivation ou de méronymie. On pourrait envisager de déployer d'autres techniques d'analyse distributionnelle sur les données que nous avons utilisées afin de comparer leur capacité à repérer des relations lexico-sémantiques à partir de corpus spécialisés.

Étant donné la qualité des résultats obtenus sur le corpus PANACEA, un corpus provenant de sources hétérogènes et destiné à des publics variés, il serait intéressant de vérifier quel degré de précision on peut atteindre en exploitant un corpus plus homogène et destiné à des experts, mais portant sur le même domaine. Nous envisageons également de vérifier dans quelle mesure ce modèle peut faciliter une description terminologique basée sur la sémantique des cadres (Fillmore, 1982; Ruppenhofer *et al.*, 2010).

Remerciements

Nous remercions Marie-Claude L'Homme, Patrick Drouin et les relecteurs anonymes pour leurs commentaires et suggestions, et nous remercions Mme L'Homme d'avoir réalisé l'évaluation manuelle. Ce projet bénéficie du soutien financier du Conseil de recherches en sciences humaines (CRSH) du Canada.

Références

ADAM C., FABRE C. & MULLER P. (2013). Évaluer et améliorer une ressource distributionnelle : Protocole d'annotation de liens sémantiques en contexte. *TAL*, **54**(1), 71–97.

ANGUIANO E. H. & DENIS P. (2011). FreDist : Automatic construction of distributional thesauri for French. In *Actes de la 18e conférence sur le traitement automatique des langues naturelles (TALN)*, p. 119–124, Montpellier.

BOUDIN F. (2013). TALN Archives : Une archive numérique francophone des articles de recherche en traitement automatique de la langue. In *Actes de la 20e conférence sur le traitement automatique des langues naturelles (TALN)*, p. 507–514, Les Sables d'Olonne.

BOURIGAULT D. (2002). Upery : Un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. In *Actes de la 9e conférence sur le traitement automatique des langues naturelles (TALN)*, p. 75–84, Nancy.

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.

DROUIN P. (2007). Identification automatique du lexique scientifique transdisciplinaire. *Revue française de linguistique appliquée*, **12**(2), 45–64.

FERRET O. (2010). Similarité sémantique et extraction de synonymes à partir de corpus. In *Actes de la 17e conférence sur le traitement automatique des langues naturelles (TALN)*, Montréal.

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.

GREFENSTETTE G. (1992). Sextant : Exploring unexplored contexts for semantic extraction from syntactic analysis. In *Proceedings of the 30th annual meeting on Association for Computational Linguistics*, p. 324–326 : Association for Computational Linguistics.

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

KIELA D. & CLARK S. (2014). A systematic study of semantic vector space model parameters. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC) @ EACL 2014*, p. 21–30 : Association for Computational Linguistics.

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.

LAVELLI A., SEBASTIANI F. & ZANOLI R. (2004). Distributional term representations : An experimental comparison. In *Proceedings of the thirteenth ACM international conference on information and knowledge management*, p. 615–624 : ACM.

L'HOMME M.-C. (2004). *La terminologie : Principes et techniques*. Montréal : Presses de l'Université de Montréal.

LUND K. & BURGESS C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, **28**(2), 203–208.

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*, volume 17, p. 660–665.

MORLANE-HONDÈRE F. & FABRE C. (2012). Le test de substituabilité à l'épreuve des corpus : Utiliser l'analyse distributionnelle automatique pour l'étude des relations lexicales. In *Actes du Congrès mondial de linguistique française (CMLF) 2012*, p. 1001–1015.

NAZARENKO A., ZWEIGENBAUM P., BOUAUD J. & HABERT B. (1997). Corpus-based identification and refinement of semantic classes. In *Proceedings of the AMIA Annual Fall Symposium*, p. 585–589 : American Medical Informatics Association.

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

PEDREGOSA F., VAROQUAUX G., GRAMFORT A., MICHEL V., THIRION B., GRISEL O., BLONDEL M., PRETTENHOFER P., WEISS R., DUBOURG V., VANDERPLAS J., PASSOS A., COURNAPEAU D., BRUCHER M., PERROT M. & DUCHESNAY E. (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

PHAL A. (1971). *Vocabulaire général d'orientation scientifique (V.G.O.S.) – Part du lexique commun dans l'expression scientifique*. Paris : Didier, Crédif.

REISINGER J. & MOONEY R. J. (2010). Multi-prototype vector-space models of word meaning. In *Human Language Technologies : The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, p. 109–117 : Association for Computational Linguistics.

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`.

SAHLGREN M. (2006). *The word-space model : Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. PhD thesis, Stockholm University.

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.

TURNEY P. D. & PANTEL P. (2010). From frequency to meaning : Vector space models of semantics. *Journal of artificial intelligence research*, **37**(1), 141–188.

TUTIN A. (2007). Traitement sémantique par analyse distributionnelle des noms transdisciplinaires des écrits scientifiques. In *Actes de la 14e conférence sur le traitement automatique des langues naturelles (TALN)*, p. 283–292, Toulouse.

URIELI A. & TANGUY L. (2013). L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions : Études de cas avec l'analyseur Talismane. In *Actes de la 20e conférence sur le traitement automatique des langues naturelles (TALN)*, p. 188–201, Les Sables d'Olonne.

WEEDS J., WEIR D. & MCCARTHY D. (2004). Characterising measures of lexical distributional similarity. In *Proceedings of the 20th international conference on Computational Linguistics*, p. 1015 : Association for Computational Linguistics.