# Sélection non supervisée de relations sémantiques pour améliorer un thésaurus distributionnel

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

RÉSUMÉ

Les travaux se focalisant sur la construction de thésaurus distributionnels ont montré que les relations sémantiques qu'ils recèlent sont principalement fiables pour les mots de forte fréquence. Dans cet article, nous proposons une méthode pour rééquilibrer de tels thésaurus en faveur des mots de fréquence faible sur la base d'un mécanisme d'amorçage : un ensemble d'exemples et de contre-exemples de mots sémantiquement similaires sont sélectionnés de façon non supervisée et utilisés pour entraîner un classifieur supervisé. Celui-ci est ensuite appliqué pour réordonner les voisins sémantiques du thésaurus utilisé pour sélectionner les exemples et contre-exemples. Nous montrons comment les relations entre les constituants de noms composés similaires peuvent être utilisées pour réaliser une telle sélection et comment conjuguer ce critère à un critère déjà expérimenté sur la symétrie des relations sémantiques. Nous évaluons l'intérêt de cette procédure sur un large ensemble de noms en anglais couvrant un vaste spectre de fréquence.

ABSTRACT

**Unsupervised selection of semantic relations for improving a distributional thesaurus**

Work about distributional thesauri has shown that the relations in these thesauri are mainly reliable for high frequency words. In this article, we propose a method for improving such a thesaurus through its re-balancing in favor of low frequency words. This method is based on a bootstrapping mechanism : a set of positive and negative examples of semantically similar words are selected in an unsupervised way and used for training a supervised classifier. This classifier is then applied for reranking the semantic neighbors of the thesaurus used for example selection. We show how the relations between the mono-terms of similar nominal compounds can be used for performing this selection and how to associate this criterion with an already tested criterion based on the symmetry of semantic relations. We evaluate the interest of the global procedure for a large set of English nouns with various frequencies.

MOTS-CLÉS : Sémantique lexicale, similarité sémantique, thésaurus distributionnels.

KEYWORDS: Lexical semantics, semantic similarity, distributional thesauri.

## 1 Introduction

Le travail présenté dans cet article s'inscrit dans le contexte de la construction automatique de thésaurus à partir de corpus. Dans le prolongement de (Grefenstette, 1994) ou (Curran

et Moens, 2002), une manière largement répandue d'aborder ce problème est d'utiliser une mesure de similarité sémantique pour extraire les voisins sémantiques de chacune des entrées pressenties du thésaurus. Trois principales approches peuvent être distinguées pour construire une telle mesure. La première repose sur des ressources construites manuellement abritant des relations sémantiques clairement identifiées, généralement de nature paradigmatique. Les travaux exploitant des réseaux lexicaux de type WordNet pour élaborer des mesures de similarité sémantique, tels que (Budanitsky et Hirst, 2006) ou (Pedersen *et al.*, 2004), entrent pleinement dans cette catégorie. Ces mesures s'appuient typiquement sur la structure hiérarchique de ces réseaux, fondée sur des relations d'hyperonymie. La deuxième approche pour construire une telle mesure fait appel à une source de connaissances concernant les mots moins structurée que la précédente : les descriptions textuelles de leur sens. Les *gloses* de WordNet ont ainsi été utilisées pour mettre en œuvre des mesures de type Lesk dans (Banerjee et Pedersen, 2003) et plus récemment, des mesures ont été définies à partir de Wikipédia ou des définitions des Wiktionaries (Gabrilovich, 2007). La dernière option pour la construction d'une mesure de similarité sémantique prend appui sur un corpus en généralisant l'hypothèse distributionnelle : chaque mot est caractérisé par l'ensemble des contextes dans lesquels il apparaît pour un corpus donné et la similarité sémantique de deux mots est évaluée sur la base de la proportion de contextes que ces deux mots partagent. Cette perspective, initialement adoptée par (Grefenstette, 1994) et (Lin, 1998), a fait l'objet d'études approfondies, notamment dans (Curran et Moens, 2002), (Weeds, 2003) ou (Heylen *et al.*, 2008).

Le problème de l'amélioration des résultats d'une implémentation « classique » de l'approche distributionnelle telle qu'elle est réalisée dans (Curran et Moens, 2002) a déjà fait l'objet d'un certain nombre de travaux. Une partie d'entre eux se sont focalisés sur la pondération des éléments constituant les contextes distributionnels, à l'instar de (Broda *et al.*, 2009), qui transforme les poids au sein de des contextes en rangs, ou de (Zhitomirsky-Geffet et Dagan, 2009), repris et étendu par (Yamamoto et Asakura, 2010), qui propose une méthode fondée sur l'amorçage pour modifier les poids des éléments des contextes en s'appuyant sur les voisins sémantiques trouvés au moyen d'une mesure de similarité distributionnelle initiale. Des approches plus radicalement différentes ont également vu le jour. L'utilisation de méthodes de réduction de dimensions, comme l'Analyse Sémantique Latente dans (Padó et Lapata, 2007), les modèles de type multi-prototype (Reisinger et Mooney, 2010) ou la redéfinition de l'approche distributionnelle dans un cadre bayésien dans (Kazama *et al.*, 2010) se rangent dans cette seconde catégorie.

Le travail que nous présentons dans cet article s'appuie comme (Zhitomirsky-Geffet et Dagan, 2009) sur un mécanisme d'amorçage mais adopte une perspective différente, initiée dans (Ferret, 2012) : au lieu d'utiliser les « meilleurs » voisins sémantiques pour adapter directement les poids des éléments constituant les contextes distributionnels des mots, l'idée est de sélectionner de façon non supervisée un ensemble restreint de mots jugés sémantiquement similaires pour entraîner, à l'instar de (Hagiwara, 2008), un classifieur statistique supervisé capable de modéliser la notion de similarité sémantique. La sélection de cet ensemble d'apprentissage est réalisée plus précisément en associant deux critères faibles fondés sur la similarité distributionnelle des mots : le premier, déjà expérimenté dans (Ferret, 2012), exploite la symétrie de la relation de similarité sémantique ; le second, nouvellement introduit ici, fait l'hypothèse que les constituants de mots composés sémantiquement similaires sont eux-mêmes susceptibles d'entretenir des liens de similarité sémantique. Nous montrons que le classifieur ainsi construit est utilisable pour réordonner les voisins sémantiques trouvés par la mesure de similarité initiale et corriger certaines de ses insuffisances du point de vue de la construction d'un thésaurus distributionnel.

# 2 Construction d'un thésaurus distributionnel initial

L'utilisation de l'amorçage implique dans notre cas de construire un thésaurus initial dont la qualité, au moins pour un sous-ensemble de celui-ci, soit suffisamment élevée pour servir de marchepied à une amélioration plus globale. Compte tenu du mode de construction de ce type de thésaurus, cet objectif prend la forme de la définition d'une mesure de similarité distributionnelle obtenant des performances, telles qu'elles peuvent être évaluées au travers de tests de type TOEFL (Landauer et Dumais, 1997) par exemple, compatibles avec cette exigence. (Ferret, 2010) s'est attaché à la sélection d'une telle mesure. Nous reprenons ici les conclusions de ce travail.

## 2.1 Définition d'une mesure de similarité distributionnelle

Bien que notre langue cible soit l'anglais, nous avons choisi de limiter le niveau des traitements linguistiques appliqués au corpus source de nos données distributionnelles à l'étiquetage morpho-syntaxique et à la lemmatisation, de manière à faciliter la transposition du travail à des langues moins dotées. Cette approche apparaît à cet égard comme un compromis raisonnable entre l'approche de (Freitag *et al.*, 2005), dans laquelle aucune normalisation n'est faite, et l'approche plus largement répandue consistant à utiliser un analyseur syntaxique, à l'instar de (Curran et Moens, 2002). Plus précisément, nous nous sommes appuyés sur l'outil *TreeTagger* (Schmid, 1994) pour assurer le prétraitement du corpus AQUAINT-2 qui est à la base de ce travail. Ce corpus comprenant environ 380 millions de mots est composé d'articles de journaux.

Les paramètres d'extraction des données distributionnelles et les caractéristiques de la mesure de similarité sont quant à eux issus de la sélection opérée dans (Ferret, 2010) :

- contextes distributionnels constitués de cooccurrents graphiques : noms, verbes et adjectifs collectés grâce à une fenêtre de taille fixe centrée sur chaque occurrence du mot cible ;
- taille de la fenêtre = 3 (un mot à droite et un mot à gauche du mot cible), c'est-à-dire des cooccurrents de très courte portée ;
- filtrage minimal des contextes : suppression des seuls cooccurrents de fréquence égale à 1 ;
- fonction de pondération des cooccurrents dans les contextes = *Information mutuelle* entre le mot cible et son cooccurrent ;
- mesure de similarité entre contextes, pour évaluer la similarité sémantique de deux mots = mesure *Cosinus*.

Un filtre fréquentiel est en outre appliqué à la fois aux mots cibles et à leurs cooccurrents puisque seuls les mots de fréquence supérieure à 10 sont considérés.

## 2.2 Construction et évaluation du thésaurus initial

La construction de notre thésaurus distributionnel initial à partir de la mesure de similarité définie ci-dessus a été réalisée comme dans (Lin, 1998) ou (Curran et Moens, 2002) en extrayant les plus proches voisins sémantiques de chacune de ses entrées. Plus précisément, cette mesure a été calculée entre chaque entrée et l'ensemble de ses voisins possibles. Ces voisins ont ensuite été ordonnés selon l'ordre décroissant des valeurs de cette mesure et les $N$ premiers voisins ($N = 100$) ont été conservés en tant que voisins sémantiques de l'entrée. Les entrées du thésaurus

de même que leurs voisins possibles étaient constitués des noms du corpus AQUAINT-2 de fréquence supérieure à 10. À titre illustratif, nous donnons les premiers voisins de deux entrées de ce thésaurus, *aid* et *procurator*, avec leur poids :

| | |
|---|---|
| aid | assistance [0,41] relief [0,34] funding [0,29] grant [0,27] fund [0,26] donation [0,26] ... |
| procurator | justiceship [0,31] amadou [0,27] commmission [0,26] pamphleteer [0,22] ... |

Le tableau 1 montre quant lui les résultats de l'évaluation du thésaurus distributionnel obtenu, réalisée en comparant les voisins sémantiques extraits à deux ressources de référence complémentaires : les synonymes de WordNet [W], dans sa version 3.0, qui permettent de caractériser une similarité fondée sur des relations paradigmatiques et le thésaurus Moby [M], qui regroupe des mots liés par des relations plus diverses. Comme l'illustre la $4^{ème}$ colonne du tableau, ces deux ressources sont aussi très différentes en termes de richesse. Le but étant d'évaluer la capacité à extraire des voisins sémantiques, elles sont filtrées pour en exclure les entrées et les voisins non présents dans le vocabulaire du corpus AQUAINT-2 (cf. la différence entre le nombre de mots de la $1^{ère}$ colonne et le nombre de mots effectivement évalués de la $3^{ème}$ colonne). Une fusion de ces deux ressources a également été faite [WM]. La fréquence des mots étant une donnée importante des approches distributionnelles, les résultats globaux sont différenciés suivant deux tranches fréquentielles de même effectif (7 335 mots chacune) : *hautes* pour les mots de fréquence > à la fréquence médiane (249) et *basses* pour les autres. Ces résultats se déclinent sous la forme de différentes mesures, à commencer à la $5^{ème}$ colonne par le taux de rappel par rapport aux ressources considérées pour les 100 premiers voisins de chaque mot. Ces voisins

| **fréq.** | **réf.** | **#mots éval.** | **#syn. /mot** | **rappel** | **R-préc.** | **MAP** | **P@1** | **P@5** | **P@10** | **P@100** |
|---|---|---|---|---|---|---|---|---|---|---|
| | W | 10 473 | 2,9 | 24,6 | 8,2 | 9,8 | 11,7 | 5,1 | 3,4 | 0,7 |
| toutes | M | 9 216 | 50,0 | 9,5 | 6,7 | 3,2 | 24,1 | 16,4 | 13,0 | 4,8 |
| 14 670 | WM | 12 243 | 38,7 | 9,8 | 7,7 | 5,6 | 22,5 | 14,1 | 10,8 | 3,8 |
| | W | 5889 | 3,3 | 29,4 | 11,8 | 13,5 | 17,4 | 7,5 | 4,9 | 1,0 |
| hautes | M | 5751 | 60,5 | 11,2 | 9,4 | 4,6 | 35,9 | 24,2 | 18,9 | 6,8 |
| 7 335 | VM | 6754 | 52,6 | 11,4 | 11,1 | 7,4 | 36,4 | 22,8 | 17,5 | 6,0 |
| | W | 4584 | 2,3 | 16,0 | 3,7 | 5,1 | 4,2 | 2,0 | 1,4 | 0,4 |
| basses | M | 3465 | 32,5 | 4,4 | 2,3 | 0,9 | 4,4 | 3,4 | 3,1 | 1,4 |
| 7 335 | WM | 5489 | 21,6 | 5,1 | 3,6 | 3,4 | 5,5 | 3,3 | 2,7 | 1,1 |

TABLE 1 – Évaluation de l'extraction des voisins sémantiques (mesures données en pourcentage)

étant ordonnés, il est en outre possible de réutiliser les métriques d'évaluation classiquement adoptées en recherche d'information en faisant jouer aux mots cibles le rôle de requêtes et aux voisins celui des documents. Les dernières colonnes du tableau 1 rendent compte de ces mesures : la R-précision (R-préc.) est la précision obtenue en se limitant aux R premiers voisins, *R* étant le nombre de synonymes dans la ressource de référence pour l'entrée considérée ; la MAP (Mean Average Precision) est la moyenne des précisions pour chacun des rangs auxquels un synonyme de référence a été identifié ; enfin, sont données les précisions pour différents seuils de nombre de voisins sémantiques examinés (précision après examen des 1, 5, 10 et 100 premiers voisins). Les résultats du tableau 1 suscitent trois principales observations. En premier lieu, il faut constater que les résultats sont globalement faibles. Cette faiblesse touche à la fois la proportion des synonymes et mots liés trouvés et leur rang parmi les voisins sémantiques. Bien que les comparaisons avec d'autres travaux soient rendues difficiles par la diversité des conditions de

construction et d'évaluation des thésaurus, il est néanmoins possible d'affirmer que cette faiblesse ne nous est pas spécifique. (Muller et Langlais, 2011) ont ainsi évalué le thésaurus construit dans (Lin, 1998) avec les mêmes mesures et les mêmes références que les nôtres et trouvent des résultats assez comparables en tenant compte du fait que le corpus de (Lin, 1998) était beaucoup plus gros que le nôtre, 3 milliards de mots, et que les données distributionnelles étaient extraites sur la base de cooccurrences syntaxiques. À titre indicatif, l'utilisation de WordNet comme référence pour des fréquences $> 5\,000$ donnent ainsi les valeurs suivantes pour les données de Lin : P@1 = 16,5 ; P@5 = 5,0 ; P@10 = 3,5 ; MAP = 9,2 ; R-préc. = 16,7. Par rapport aux fréquences hautes du tableau 1, configuration la plus directement comparable, on constate qu'en dehors de la R-précision, plus élevée dans le cas des données de Lin, les autres mesures donnent des valeurs proches de celles rapportées dans (Muller et Langlais, 2011).

Le deuxième point que laisse apparaître ce tableau est la forte dépendance des résultats vis-à-vis de la fréquence des entrées du thésaurus. Les meilleurs résultats sont ainsi obtenus par les mots de la tranche de fréquences supérieure tandis que les mesures d'évaluation diminuent de façon très significative pour la tranche fréquentielle la plus basse. Le dernier constat a trait à l'impact de la référence utilisée pour l'évaluation du thésaurus. WordNet est ainsi caractérisé par un nombre restreint de synonymes pour chaque nom tandis que le thésaurus Moby contient pour chaque entrée un large ensemble de synonymes et de mots liés. La conséquence de cette différence s'observe clairement au niveau des précisions à différents rangs dans le tableau 1 : les valeurs sont nettement supérieures pour Moby par rapport à WordNet alors que la mesure de similarité sous-jacente est la même. Seule la richesse de la référence varie. Ce phénomène est également illustré dans (Ferret, 2010) au travers de la comparaison avec (Curran et Moens, 2002).

# 3 Amélioration d'un thésaurus distributionnel

## 3.1 Principes

L'évaluation de notre thésaurus distributionnel initial montre que les voisins sémantiques obtenus sont significativement meilleurs pour certaines entrées que pour d'autres. Une telle configuration est *a priori* favorable à un mécanisme de type amorçage dans la mesure où il est envisageable de s'appuyer sur les résultats des « bonnes » entrées pour obtenir une amélioration plus globale. (Zhitomirsky-Geffet et Dagan, 2009) a déjà fait appel à l'amorçage dans un contexte proche du nôtre, l'acquisition de relations d'implication textuelle entre mots. Cependant, des expérimentations rapportées dans (Ferret, 2010) ont montré que la transposition de cette approche à notre problème n'était pas concluante. Ainsi, au lieu d'utiliser les résultats d'une mesure de similarité initiale pour modifier directement les poids des éléments constitutifs des contextes distributionnels, nous avons adopté une approche plus indirecte, fondé sur (Hagiwara, 2008).

(Hagiwara, 2008) a en effet montré qu'il est possible d'entraîner et d'appliquer avec un bon niveau de performance un classifieur statistique, en l'occurrence de type Machine à Vecteurs de Support (SVM), pour décider si deux mots sont ou ne sont pas synonymes, au sens large du terme. Par ailleurs, ce travail montre également que la valeur de la fonction de décision caractérisant les SVM, dont on n'utilise que le signe dans le cas d'une classification binaire, peut jouer, pour l'ordonnancement des voisins sémantiques, le même rôle que la valeur d'une mesure de similarité telle que celle définie à la section 2.

À la différence de (Hagiwara, 2008), nous ne disposons pas d'un ensemble d'exemples et de contre-exemples étiquetés manuellement pour réaliser l'entraînement d'un tel classifieur. En revanche, les voisins sémantiques obtenus en appliquant la mesure de similarité de la section 2 peuvent être exploités pour construire un tel ensemble. Cette mesure n'offre pas de critère évident pour discriminer les mots sémantiquement liés[1]. Cependant, elle peut être utilisée plus indirectement pour sélectionner un ensemble d'exemples et de contre-exemples de façon non supervisée en minimisant le nombre d'erreurs. Ces erreurs correspondent à des exemples considérés comme positifs mais en réalité négatifs et d'exemples considérés comme négatifs mais en fait positifs. Dans cette optique, nous proposons d'entraîner un classifieur SVM grâce à ces ensembles et de l'appliquer ensuite pour réordonner les voisins sémantiques obtenus précédemment. L'ensemble de la démarche peut être résumée par la procédure suivante :

– définition d'une mesure de similarité distributionnelle ;
– application de cette mesure pour la construction d'un thésaurus distributionnel par le biais de l'extraction de voisins sémantiques ;
– sélection non supervisée d'un ensemble d'exemples et de contre-exemples de mots sémantiquement similaires grâce aux résultats de l'application de la mesure de similarité ;
– entraînement d'un classifieur statistique à partir de l'ensemble d'exemples constitué ;
– application du classifieur entraîné au réordonnancement des voisins du thésaurus initial.

Le point clé de l'amélioration des résultats par ce moyen est de sélectionner de façon non supervisée un nombre suffisant d'exemples et de contre-exemples en minimisant les erreurs propres à une telle sélection. Dans la section 4, nous proposons d'associer deux méthodes faibles, à la fois au sens de la productivité et de la validité des résultats, pour accomplir cette tâche.

### 3.2 Représentation des exemples

Avant de présenter plus en détail ce processus de sélection, il convient de préciser la nature des exemples et des contre-exemples. Nous reprenons de ce point de vue la conception développée dans (Hagiwara, 2008) : un exemple est constitué d'un couple de mots considérés comme synonymes ou plus généralement sémantiquement liés ; un contre-exemple est formé d'un couple de mots entre lesquels un tel lien sémantique n'existe pas. La représentation de ces couples pour un classifieur de type SVM s'effectue en associant leurs représentations distributionnelles. Cette association s'effectue pour chaque couple ($M_1$, $M_2$) en sommant le poids des cooccurrents communs aux mots $M_1$ et $M_2$. Les cooccurrents de $M_x$ non présents dans $M_y$ se voient attribuer un poids nul. Chaque exemple ou contre-exemple a donc la même forme que la représentation distributionnelle d'un mot, c'est-à-dire un vecteur de mots pondérés.

## 4 Sélection des exemples et des contre-exemples

Du point de vue de la sélection des exemples et des contre-exemples de mots sémantiquement liés, le tableau 1 offre une image claire : trouver des exemples est beaucoup plus problématique que trouver des contre-exemples dans la mesure où le nombre de mots sémantiquement liés à

[1] Fixer pour ce faire un seuil sur les valeurs de similarité produit de mauvais résultats du fait de la variabilité de ces valeurs d'une entrée à l'autre. Ce constat a motivé notre choix d'utiliser un SVM en classification plutôt qu'en régression.

une entrée du thésaurus diminue très fortement dès que l'on considère ses voisins de rang un peu élevé. Dans les expérimentations de la section 5, nous avons ainsi construits nos contre-exemples à partir de nos exemples en créant pour chaque exemple (A,B) deux contre-exemples de la forme : (A, *voisin de rang 10 de A*) et (B, *voisin de rang 10 de B*). Le choix d'un rang supérieur garantirait un nombre plus faible de faux contre-exemples (*i.e.* couples de synonymes) et donc *a priori*, de meilleurs résultats. En pratique, l'utilisation de voisins du mot cible de rang assez faible conduit à une performance supérieure, sans doute parce que ceux-ci sont plus utiles en termes de discrimination, étant plus proches de la zone de transition entre exemples et contre-exemples. Nous avons par ailleurs constaté expérimentalement que le rapport entre contre-exemples et exemples dans (Hagiwara, 2008), égal 6,5 et donc fortement déséquilibré en faveur des contre-exemples, n'était pas nécessaire dans notre situation et pouvait se ramener à 2.

Pour la sélection des exemples, le tableau 1 impose un double constat : trouver un voisin sémantiquement proche est d'autant plus probable que la fréquence de l'entrée du thésaurus considérée est élevée et que le rang du voisin est faible. La forme extrême de cette logique conduirait à retenir comme exemples tous les couples de mots (*entrée de haute fréquence*, *voisin de rang 1*), ce qui donne un large nombre d'exemples – 7 335 – mais un taux d'erreur (*i.e.* nombre de couples de mots non liés sémantiquement) également élevé – 63,6% dans le cas le plus favorable (référence WM). Nous avons donc proposé une approche plus sélective pour choisir nos exemples parmi les entrées fréquentes du thésaurus afin d'aboutir à une solution plus équilibrée entre le nombre d'exemples et leur taux d'erreur. Cette approche associe deux méthodes de sélection non supervisées produisant chacune un nombre limité d'exemples mais avec un meilleur taux d'erreur. Nous présentons ces méthodes dans les deux sections suivantes en détaillant plus spécifiquement celle fondée sur les mots composés, nouvelle proposition de cet article.

## 4.1 Sélection fondée sur les relations de symétrie dans le thésaurus

Notre première méthode de sélection d'exemples de mots sémantiquement similaires a été introduite dans (Ferret, 2012). Elle est fondée sur l'hypothèse que les relations de similarité sémantique sont symétriques, ce qui est strictement vrai dans le cas des synonymes de WordNet mais l'est moins pour les mots liés de Moby. En accord avec cette hypothèse, nous avons considéré que si une entrée A du thésaurus initial a pour voisin un mot B, ce voisin a d'autant plus de chances d'être sémantiquement similaire à A que A est lui-même un voisin de B en tant qu'entrée du thésaurus. Plus précisément, les résultats du tableau 1 nous ont conduit à limiter l'application de ce principe aux voisins de rang 1 et aux entrées de haute fréquence, dont les voisins sont eux-mêmes généralement des noms de haute fréquence. Nous avons donc appliqué ce principe aux 7 335 entrées dites de haute fréquence du thésaurus, obtenant des cas de symétrie entre entrée et voisin de rang 1 pour 1 592 entrées. 796 exemples de mots sémantiquement similaires ont finalement été produits puisque les couples (A,B) et (B,A) représentent un même exemple.

## 4.2 Sélection fondée sur les mots composés

### 4.2.1 Construction d'un thésaurus distributionnel de noms composés

La seconde méthode que nous proposons pour la sélection de couples de mots sémantiquement similaires repose sur l'hypothèse que les mono-termes de deux mots composés sémantiquement

 

similaires occupant dans ces deux termes le même rôle syntaxique sont eux-mêmes susceptibles d'être sémantiquement similaires. Par exemple, les noms composés *movie_director* et *film_director* étant trouvés similaires et les têtes syntaxiques de ces deux composés étant identiques, il est vraisemblable que la similarité sémantique observée entre *film* et *movie* dans le thésaurus initial soit véritable. Le point de départ de cette hypothèse étant la similarité sémantique des mots composés, nous avons commencé par construire un thésaurus distributionnel de noms composés pour l'anglais, à l'image du thésaurus de la section 2 pour les noms simples. Cette construction a été réalisée à partir du même corpus et avec les mêmes paramètres que pour les mono-termes, à l'exception bien entendu de l'ajout d'une étape dans le prétraitement linguistique des documents du corpus pour l'identification des noms composés. Cette identification a été réalisée en deux étapes : un ensemble de noms composés ont d'abord été extraits du corpus AQUAINT-2 sur la base d'un nombre limité de patrons morpho-syntaxiques ; les plus fréquents de ces composés ont ensuite été utilisés comme référence dans un processus d'indexation contrôlée.

La première étape a été mise en œuvre grâce à l'outil *mwetoolkit* (Ramisch *et al.*, 2010), qui permet d'extraire efficacement des mots composés d'un corpus à partir du résultat d'un étiqueteur morpho-syntaxique, le *TreeTagger* dans notre cas, en s'appuyant sur un ensemble de patrons morpho-syntaxiques. Nous nous sommes limités aux trois patrons de noms composés suivants : *<nom> <nom>*, *<adjectif> <nom>*, *<nom> <préposition> <nom>*. Un ensemble de 3 246 401 noms composés ont ainsi été extraits du corpus AQUAINT-2 parmi lesquels seuls les 30 121 termes de fréquence supérieure à 100 ont été retenus, pour des raisons à la fois de fiabilité et de limitation du vocabulaire pour la construction du thésaurus. L'identification de ces termes de référence dans les textes a ensuite été réalisée en appliquant la stratégie de l'appariement maximal à la sortie lemmatisée du *TreeTagger*. Finalement, des contextes distributionnels constitués à la fois de mots simples et de termes complexes ont été construits suivant les principes de la section 2 et des voisins ont été trouvés pour 29 174 noms composés.

| **réf.** | **#mots éval.** | **#syn. /mot** | **rappel** | **R-préc.** | **MAP** | **P@1** | **P@5** | **P@10** | **P@100** |
|---|---|---|---|---|---|---|---|---|---|
| W | 608 | 1,2 | 82,0 | 41,5 | 50,0 | 43,4 | 14,3 | 8,0 | 1,0 |
| M | 241 | 2,3 | 38,0 | 9,0 | 12,2 | 11,2 | 6,5 | 4,2 | 0,9 |
| WM | 813 | 1,6 | 63,5 | 32,7 | 39,5 | 34,9 | 12,3 | 7,1 | 1,0 |

TABLE 2 – Évaluation du thésaurus distributionnel pour les noms composés

Le tableau 2 donne les résultats de l'évaluation des voisins sémantiques trouvés en prenant comme précédemment en tant que référence WordNet, le thésaurus Moby et la fusion des deux. Le premier constat pouvant être fait est la proportion très faible, par rapport aux mono-termes, d'entrées ayant pu être évaluées : seulement 2,8% des entrées, à comparer à 83,5% des entrées pour les mono-termes. De ce fait, les résultats de cette évaluation doivent être considérés avec prudence, même si le nombre d'entrées évaluées est globalement plus élevé que le nombre d'entrées considérées dans les évaluations standards : 70 pour (Curran et Moens, 2002) ou 353 pour (Gabrilovich, 2007). Cette prudence est particulièrement de mise pour les mots liés de Moby : les résultats, à l'exception du rappel, sont très significativement inférieurs à ceux obtenus avec les mono-termes mais le nombre d'entrées évaluées – 241 – est aussi faible. À l'inverse, les performances obtenues pour les synonymes de WordNet sont très nettement supérieures sur tous les plans à celles caractérisant les mono-termes, ces résultats étant obtenus pour un nombre d'entrées – 608 – nettement supérieur. Cette différence ne s'expliquant pas par un biais concernant

la fréquence des entrées évaluées vis-à-vis respectivement de WordNet et de Moby, il semble donc que le comportement des noms composés soit, du point de vue des similarités distributionnelles, l'inverse de celui des noms simples, favorisant les relations sémantiques paradigmatiques par rapport aux relations syntagmatiques. La plus faible ambiguïté sémantique des noms composés serait une explication possible de ce phénomène qui demanderait néanmoins une étude plus approfondie avec une base d'évaluation plus large.

#### 4.2.2 Sélection d'exemples à partir de noms composés

La sélection d'exemples de mots simples sémantiquement similaires à partir de noms composés s'appuie sur la structure syntaxique de ces noms composés. Compte tenu des patrons utilisés pour l'extraction des termes, cette structure prend la forme de l'un des trois grands schémas suivants : <nom>$_{expansion}$ <nom>$_{tête}$, <adjectif>$_{expansion}$ <nom>$_{tête}$, <nom>$_{tête}$ <préposition> <nom>$_{expansion}$.

Chaque nom composé $C_i$ a ainsi été représenté sous la forme d'un couple de noms $(T_i, E_i)$, dans lequel $T_i$ représente la tête syntaxique de $C_i$ et $D_i$, son expansion, au sens des grammaires de dépendance. Conformément au principe sous-tendant notre méthode sélection, si un nom composé $(T_2, E_2)$ est un voisin sémantique d'un nom composé $(T_1, E_1)$ (au plus, son $c^{ième}$ voisin), il est probable que $T_1$ et $T_2$ ou $E_1$ et $E_2$ soient sémantiquement similaires[2]. Comme le montre le tableau 2, notre thésaurus distributionnel de noms composés est cependant loin d'être parfait. Pour limiter les erreurs, nous avons ajouté des contraintes sur l'appariement des constituants des noms composés similaires en nous appuyant sur la similarité distributionnelle de ces constituants. Au final, nous sélectionnons des exemples de noms simples sémantiquement similaires (couples de noms suivant $\rightarrow$) en appliquant les trois règles suivantes, dans lesquelles $E_1 = E_2$ signifie que $E_1$ et $E_2$ sont identiques et $T_1 \equiv T_2$ signifie que $T_2$ est au plus le $n^{ième}$ voisin de $T_1$ dans notre thésaurus de noms simples :

(1) $T_1 \equiv T_2$ et $E_1 = E_2 \rightarrow (T_1, T_2)$
(*crash*, *accident*) issu de *car_crash* et *car_accident* ; (*boat*, *vessel*) de *fishing_vessel* et *fishing_boat*
(2) $E_1 \equiv E_2$ et $T_1 = T_2 \rightarrow (E_1, E_2)$
(*ocean*, *sea*) de *ocean_floor* et *sea_floor* ; (*jail*, *prison*) de *prison_cell* et *jail_cell*
(3) $E_1 \equiv E_2$ et $T_1 \equiv T_2 \rightarrow (T_1, T_2), (E_1, E_2)$
(*increase*, *rise*) et (*salary*, *pay*) de *salary_increase* et *pay_rise*

# 5 Expérimentations et évaluation

## 5.1 Sélection des exemples de mots sémantiquement similaires

Le tableau 3 fait une synthèse des résultats de nos deux méthodes de sélection de mots sémantiquement similaires en donnant le pourcentage des couples sélectionnés trouvés dans chacune de nos ressources (W, M et WM) ainsi que la taille de chaque ensemble d'exemples. Dans le cas de la seconde méthode, ces mesures sont également déclinées au niveau de chacune des trois

[2]Notons que nous ne nous intéressons pas ici à la similarité entre $E1$ et $E2$ lorsque ce sont des adjectifs.

règles de sélection. Les chiffres donnés entre crochets représentent quant à eux les pourcentages d'erreurs parmi les exemples de mots non similaires. Ces résultats ont été obtenus en fixant expérimentalement la taille du voisinage considéré pour les entrées à 3 pour les noms composés ($c$) et à 1 pour les noms simples ($n$). En outre, ces trois règles de sélection ont été appliquées avec l'ensemble des entrées du thésaurus des noms composés et les entrées du thésaurus des noms simples dites de haute fréquence. Les valeurs des paramètres $c$ et $n$ ne résultent pas d'une optimisation sophistiquée mais répondent plutôt une logique induite des évaluations réalisées : pour les mono-termes, seul le premier voisin est retenu du fait de la faiblesse des résultats alors que pour les multi-termes, le voisinage peut être légèrement élargi du fait d'une meilleure fiabilité des voisins. Il est à noter par ailleurs que l'association de deux ensembles d'exemples sélectionnés par des méthodes différentes rend les résultats plus stables vis-à-vis des valeurs de $c$ et $n$.

| **méthode** | **W** | | **M** | | **WM** | | **# exemples** |
|---|---|---|---|---|---|---|---|
| symétrie | 36,6 | [2,0] | 55,5 | [14,4] | 59,7 | [12,4] | 796 |
| règle (1) | 19,3 | | 56,1 | | 56,9 | | 921 |
| règle (2) | 16,2 | | 42,4 | | 44,7 | | 308 |
| règle (3) | 13,5 | | 45,9 | | 46,2 | | 40 |
| *règles (1,2)* | *17,8* | [2,5] | *52,2* | [16,8] | *53,0* | [16,1] | *1 115* |
| règles (1,2,3) | 17,6 | | 51,7 | | 52,4 | | 1 131 |
| symétrie + *règles (1,2)* | *23,5* | [2,3] | *52,5* | [16,3] | *54,3* | [15,0] | *1 710* |
| symétrie + règles (1,2,3) | 23,3 | | 52,1 | | 53,9 | | 1 725 |

TABLE 3 – Résultats de la sélection des exemples

L'évaluation de la seconde méthode de sélection montre d'abord que la règle (3), qui est *a priori* la moins fiable des trois, ne produit effectivement qu'un petit nombre d'exemples tendant à dégrader les résultats. De ce fait, seule la combinaison des règles (1) et (2) a été utilisée dans ce qui suit. Cette évaluation montre en outre que les têtes de deux noms composés sémantiquement liés ont davantage tendance à être elles-mêmes similaires si leurs expansions sont similaires que n'ont tendance à être similaires des expansions de deux noms composés dont les têtes sont similaires. Ce résultat n'était pas évident *a priori* dans la mesure où l'on s'attend à ce que la tête d'un composé soit davantage représentatif de son sens que son expansion. Plus globalement, le tableau 3 laisse apparaître que la première méthode de sélection est supérieure à la seconde mais que leur association produit un compromis intéressant entre le nombre d'exemples, 1 710, et son taux d'erreur, 45,7% avec WM comme référence. Cette complémentarité est également illustrée par le faible nombre d'exemples – 201 – qu'elles partagent.

## 5.2 Mise en œuvre du réordonnancement des voisins

La mise en œuvre effective de notre approche de réordonnancement des voisins sémantiques nécessite de fixer un certain nombre de paramètres liés aux SVM. De même que (Hagiwara, 2008), nous avons adopté un noyau RBF et une stratégie de type *grid search* pour l'optimisation du paramètre $\gamma$ fixant la largeur de la fonction gaussienne du noyau RBF et du paramètre $C$ d'ajustement entre la taille de la marge et le taux d'erreur. Cette optimisation a été réalisée pour chaque ensemble d'apprentissage considéré en se fondant sur la mesure de précision calculée dans le cadre d'une validation croisée divisant ces ensembles en 5 parties. Chaque modèle SVM correspondant a été construit en utilisant l'outil LIBSVM puis appliqué à la totalité des 14 670

noms cibles de notre évaluation initiale. Plus précisément, pour chaque nom cible $NC$, une représentation d'exemple a été construite pour chaque couple ($NC$, voisin de $NC$) et a été soumise au modèle SVM considéré en mode classification. L'ensemble de ces voisins ont ensuite été réordonnés suivant la valeur de la fonction de décision ainsi calculée pour chaque voisin.

## 5.3 Évaluation

Le tableau 4 donne les résultats globaux du réordonnancement réalisé sur la base des exemples sélectionnés par chacune des deux méthodes présentées tandis que les résultats détaillés du tableau 5 correspondent au réordonnancement fondé sur l'association des deux méthodes de sélection. Chacun des ces trois thésaurus a été évalué selon les mêmes principes qu'à la section 2.2. La valeur de chaque mesure se voit associer sa différence avec la valeur correspondante pour le thésaurus initial dans le tableau 1. Enfin, comme l'évaluation s'applique au résultat d'un réordonnancement, les mesures de rappel et de précision au rang le plus lointain ne changent pas et ne sont pas rappelées.

| méthode | réf. | R-préc. | | MAP | | P@1 | | P@5 | | P@10 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| symétrie | W | 7,8 | (-0,4) | 9,4 | (-0,4) | 11,2 | (-0,5) ‡ | 5,0 | (-0,1) ‡ | 3,3 | (-0,1) ‡ |
| | M | 7,1 | (0,4) | 3,4 | (0,2) | 27,3 | (3,2) | 17,6 | (1,2) | 13,7 | (0,7) |
| | WM | 8,0 | (0,3) | 5,7 | (0,1) | 24,6 | (2,1) | 14,9 | (0,8) | 11,4 | (0,6) |
| composés | W | 7,2 | (-1,0) | 8,8 | (-1,0) | 10,4 | (-1,3) | 4,6 | (-0,5) | 3,1 | (-0,3) |
| | M | 7,1 | (0,4) | 3,3 | (0,1) | 26,8 | (2,7) | 17,4 | (1,0) | 13,5 | (0,5) |
| | WM | 7,8 | (0,1) | 5,5 | (-0,1) | 24,0 | (1,5) | 14,6 | (0,5) | 11,2 | (0,4) |

TABLE 4 – Réordonnancement des voisins sémantiques de toutes les entrées du thésaurus initial pour chaque méthode de sélection d'exemples

La tendance générale est claire : le processus de réordonnancement conduit à une amélioration significative des résultats à l'échelle globale (tableau 4 et lignes *tous* du tableau 5) pour les références M et WM[3]. Parallèlement, une diminution des résultats est observée pour la référence W, diminution statistiquement non significative pour le tableau 5. En d'autres termes, par rapport au thésaurus initial, la procédure de réordonnancement tend à favoriser les mots similaires au détriment des synonymes. Cette tendance n'est pas surprenante compte tenu du principe de ce réordonnancement : les premiers sont en effet mieux représentés que les seconds dans les exemples sélectionnés du fait même de leur meilleure représentation au niveau global. Les modèles SVM appris ne font en l'occurrence qu'amplifier un état de fait déjà présent initialement. Ce biais est particulièrement fort pour la méthode de sélection fondée sur les noms composés, comme l'illustre le tableau 4. Cependant, les résultats du tableau 5 montrent clairement l'intérêt de l'association des deux méthodes de sélection, la méthode de sélection fondée sur la symétrie des relations venant rééquilibrer ce biais au bénéfice des résultats globaux. Par ailleurs, en associant la partie du thésaurus initial correspondant aux fréquences hautes et la partie du thésaurus après réordonnancement correspondant aux fréquences basses (cf. ligne *hybride* du tableau 5), on obtient un thésaurus hybride dont les résultats sont supérieurs à ceux du thésaurus initial pour toutes les conditions.

[3]La significativité statistique des différences a été évaluée grâce à un test de Wilcoxon avec un seuil de 0,05, les échantillons étant appariés. Seules les différences suivies du signe ‡ sont considérées comme non significatives.

| fréq. | réf. | R-préc. | | MAP | | P@1 | | P@5 | | P@10 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| toutes | W | 7,9 | (-0,3) ‡ | 9,5 | (-0,3) ‡ | 11,5 | (-0,2) ‡ | 5,1 | (0,0) ‡ | 3,4 | (0,0) ‡ |
| | M | 7,2 | (0,5) | 3,5 | (0,3) | 27,9 | (3,8) | 18,1 | (1,7) | 14,1 | (1,1) |
| | WM | 8,0 | (0,3) | 5,8 | (0,2) | 25,3 | (2,8) | 15,3 | (1,2) | 11,7 | (0,9) |
| hautes | W | 9,9 | (-1,9) | 11,7 | (-1,8) | 15,1 | (-2,3) | 6,8 | (-0,7) | 4,5 | (-0,4) |
| | M | 9,4 | (0,0) | 4,5 | (-0,1) ‡ | 37,5 | (1,6) | 24,3 | (0,1) ‡ | 19,0 | (0,1) ‡ |
| | WM | 10,5 | (-0,6) ‡ | 6,8 | (-0,6) | 36,7 | (0,3) ‡ | 22,5 | (-0,3) ‡ | 17,4 | (-0,1) ‡ |
| basses | W | 5,4 | (1,7) | 6,8 | (1,7) | 6,9 | (2,7) | 3,0 | (1,0) | 2,0 | (0,6) |
| | M | 3,5 | (1,2) | 1,7 | (0,8) | 12,0 | (7,6) | 7,8 | (4,4) | 5,9 | (2,8) |
| | WM | 5,0 | (1,4) | 4,6 | (1,2) | 11,3 | (5,8) | 6,5 | (3,2) | 4,7 | (2,0) |
| toutes (hybride) | W | 9,0 | (0,8) | 10,6 | (0,8) ‡ | 12,8 | (1,1) | 5,6 | (0,5) | 3,6 | (0,2) |
| | M | 7,2 | (0,5) | 3,5 | (0,3) ‡ | 26,9 | (2,8) | 18,1 | (1,7) | 14,1 | (1,1) |
| | WM | 8,3 | (0,6) | 6,1 | (0,5) ‡ | 25,1 | (2,6) | 15,5 | (1,4) | 11,8 | (1,0) |

TABLE 5 – Réordonnancement du thésaurus initial avec les deux méthodes de sélection d'exemples

L'analyse des résultats du tableau 5 en termes de fréquence des mots met en évidence une seconde grande tendance : l'amélioration produite par le réordonnancement est d'autant plus sensible que la fréquence de l'entrée du thésaurus est faible. Ainsi, pour les noms de faible fréquence, cette amélioration s'observe quelle que soit la référence tandis que pour les noms de forte fréquence, la variation est négative pour certaines références et mesures et positive pour d'autres. Ce constat montre que le réordonnancement tend ainsi à rééquilibrer le thésaurus initial, très fortement biaisé vers les fortes fréquences. Enfin, l'évaluation de ces trois thésaurus confirment les résultats du tableau 3 à propos de chaque ensemble d'exemples sélectionnés : le thésaurus construit à partir des exemples de la première méthode de sélection est meilleur que celui construit à partir des exemples de la seconde méthode de sélection et les deux sont nettement dépassés par le thésaurus construit à partir de la fusion des deux ensembles d'exemples.

| | |
|---|---|
| **WordNet** | respect, admiration, regard |
| Moby | admiration, appreciation, acceptance, dignity, regard, respect, account, adherence, consideration, estimate, estimation, fame, greatness, homage, honor, prestige, prominence, reverence, veneration + 74 mots liés supplémentaires |
| initial | cordiality, gratitude, **admiration**, comradeship, back-scratching, perplexity, **respect**, ruination, appreciation, neighbourliness, trust, empathy, suffragette, goodwill ... |
| après réordonnancement | **respect**, **admiration**, trust, recognition, gratitude, confidence, affection, understanding, solidarity, dignity, appreciation, **regard**, sympathy, acceptance ... |

TABLE 6 – Impact du réordonnancement pour l'entrée *esteem*

Enfin, le tableau 6 illustre pour une entrée spécifique du thésaurus initial, en l'occurrence le mot *esteem*, l'impact du réordonnancement fondé sur les deux méthodes de sélection d'exemples. Ce tableau donne d'abord pour cette entrée ses synonymes dans **WordNet** et les premiers mots qui lui sont liés dans Moby. Il fait ensuite apparaître que dans notre thésaurus *initial*, les deux premiers voisins de cette entrée apparaissant dans une de nos deux ressources de référence sont les mots *admiration*, au rang 3, et le mot *respect*, au rang 7. Le *réordonnancement* améliore significativement la situation puisque ces deux mots deviennent les deux premiers voisins tandis

que le 3$^{ème}$ synonyme donné par WordNet passe du rang 22 au rang 12. Par ailleurs, le nombre de voisins présents parmi les 14 premiers mots liés de Moby passe de 3 à 6.

# 6 Conclusion et perspectives

Dans cet article, nous avons présenté une méthode fondée sur l'amorçage pour améliorer un thésaurus distributionnel. Plus précisément, cette méthode se fonde sur le réordonnancement des voisins sémantiques de ce thésaurus par le biais d'un classifieur SVM. Ce classifieur est entraîné à partir d'un ensemble d'exemples et de contre-exemples sélectionnés de façon non supervisée en combinant deux critères faibles fondés sur la similarité distributionnelle. L'un exploite la symétrie des relations sémantiques tandis que l'autre s'appuie sur l'appariement des constituants de noms composés similaires. Les améliorations apportées par cette méthode sont plus particulièrement notables pour les noms de fréquence faible ou intermédiaire et pour des mots similaires plutôt que pour de stricts synonymes.

Nous envisageons plusieurs pistes d'extension de ce travail. Tout d'abord, nous souhaitons appliquer, tout en conservant une sélection d'exemples non supervisée, des techniques de sélection de caractéristiques afin de mettre en évidence les traits les plus intéressants du point de vue de la similarité sémantique, en particulier pour améliorer les thésaurus distributionnels produits en construisant des modèles plus généraux de cette similarité. L'élargissement des critères de sélection non supervisée d'exemples est une deuxième extension assez directe du travail présenté. Alors que les techniques de sélection expérimentées reposent toutes deux sur des thésaurus distributionnels, des critères s'attachant aux occurrences des mots et à leur environnement plutôt qu'à une représentation distributionnelle sont également envisageables, comme l'utilisation de patrons linguistiques classiques d'extraction de synonymes par exemple. Sur un autre plan, l'évaluation menée, fondée sur la comparaison avec des ressources de référence, pourrait être complétée avec profit par une évaluation *in vivo* permettant de juger de l'impact des améliorations du thésaurus distributionnel sur une tâche auquel il contribue. Parmi les nombreuses tâches possibles, nous serions particulièrement intéressés par celle de segmentation thématique, dans le prolongement de (Adam et Morlane-Hondère, 2009). Enfin, nous planifions d'appliquer la méthode décrite au français en nous appuyant sur des thésaurus distributionnels comme *freDist* (Anguiano et Denis, 2011).

# Références

ADAM, C. et MORLANE-HONDÈRE, F. (2009). Détection de la cohésion lexicale par voisinage distributionnel : application à la segmentation thématique. *In RECITAL'09*, Senlis, France.

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

BANERJEE, S. B. et PEDERSEN, T. (2003). Extended gloss overlaps as a measure of semantic relatedness. *In Eighteenth International Conference on Artificial Intelligence (IJCAI-03)*, Mexico.

BRODA, B., PIASECKI, M. et SZPAKOWICZ, S. (2009). Rank-Based Transformation in Measuring Semantic Relatedness. *In 22$^{nd}$ Canadian Conference on Artificial Intelligence*, pages 187–190.

 

BUDANITSKY, A. et HIRST, G. (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

CURRAN, J. et MOENS, M. (2002). Improvements in automatic thesaurus extraction. *In Workshop of the ACL Special Interest Group on the Lexicon (SIGLEX)*, pages 59–66, Philadelphia, USA.

FERRET, O. (2010). Similarité sémantique et extraction de synonymes à partir de corpus. *In TALN 2010*.

FERRET, O. (2012). Combining bootstrapping and feature selection for improving a distributional thesaurus. *In 20$^{th}$ European Conference on Artificial Intelligence (ECAI 2012)*, pages 336–341.

FREITAG, D., BLUME, M., BYRNES, J., CHOW, E., KAPADIA, S., ROHWER, R. et WANG, Z. (2005). New experiments in distributional representations of synonymy. *In CoNLL 2005*, pages 25–32.

GABRILOVICH, Evgeniyand Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. *In IJCAI 2007*, pages 6–12.

GREFENSTETTE, G. (1994). *Explorations in automatic thesaurus discovery*. Kluwer Academic Publishers.

HAGIWARA, M. (2008). A supervised learning approach to automatic synonym identification based on distributional features. *In ACL-08, student session*, Columbus, Ohio.

HEYLEN, K., PEIRSMANY, Y., GEERAERTS, D. et SPEELMAN, D. (2008). Modelling Word Similarity : An Evaluation of Automatic Synonymy Extraction Algorithms. *In LREC 2008*, Marrakech, Morocco.

KAZAMA, J., DE SAEGER, S., KURODA, K., MURATA, M. et TORISAWA, K. (2010). A bayesian method for robust estimation of distributional similarities. *In ACL 2010*, pages 247–256.

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

LIN, D. (1998). Automatic retrieval and clustering of similar words. *In ACL-COLING'98*, pages 768–774.

MULLER, P. et LANGLAIS, P. (2011). Comparaison d'une approche miroir et d'une approche distributionnelle pour l'extraction de mots sémantiquement reliés. *In TALN 2011*.

PADÓ, S. et LAPATA, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

PEDERSEN, T., PATWARDHAN, S. et MICHELIZZI, J. (2004). Wordnet : :similarity - measuring the relatedness of concepts. *In HLT-NAACL 2004, demonstration papers*, pages 38–41.

RAMISCH, C., VILLAVICENCIO, A. et BOITET, C. (2010). mwetoolkit : a Framework for Multiword Expression Identification. *In LREC'10*, Valetta, Malta.

REISINGER, J. et MOONEY, R. J. (2010). Multi-prototype vector-space models of word meaning. *In HLT-NAACL 2010*, pages 109–117.

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

WEEDS, J. (2003). *Measures and Applications of Lexical Distributional Similarity*. Thèse de doctorat, Department of Informatics, University of Sussex.

YAMAMOTO, K. et ASAKURA, T. (2010). Even unassociated features can improve lexical distributional similarity. *In Second Workshop on NLP Challenges in the Information Explosion Era (NLPIX 2010)*, pages 32–39, Beijing, China.

ZHITOMIRSKY-GEFFET, M. et DAGAN, I. (2009). Bootstrapping Distributional Feature Vector Quality. *Computational Linguistics*, 35(3):435–461.