TALN 2010, Montréal, 19–23 juillet 2010

Similarité sémantique et extraction de synonymes à partir de corpus

Olivier Ferret
CEA, LIST, Laboratoire Vision et Ingénierie des Contenus,
Fontenay-aux-Roses, F-92265 France.
olivier.ferret@cea.fr

Résumé. La définition de mesures sémantiques au niveau lexical a fait l'objet de nombreux travaux depuis plusieurs années. Dans cet article, nous nous focalisons plus spécifiquement sur les mesures de nature distributionnelle. Bien que différentes évaluations ont été réalisées les concernant, il reste difficile à établir si une mesure donnant de bons résultats dans un cadre d'évaluation peut être appliquée plus largement avec le même succès. Dans le travail présenté, nous commençons par sélectionner une mesure de similarité sur la base d'un test de type TOEFL étendu. Nous l'appliquons ensuite au problème de l'extraction de synonymes à partir de corpus en comparant nos résultats avec ceux de (Curran & Moens, 2002). Enfin, nous testons l'intérêt pour cette tâche d'extraction de synonymes d'une méthode d'amélioration de la qualité des données distributionnelles proposée dans (Zhitomirsky-Geffet & Dagan, 2009).

Abstract. The definition of lexical semantic measures has been the subject of lots of works for many years. In this article, we focus more specifically on distributional semantic measures. Although several evaluations about this kind of measures were already achieved, it is still difficult to determine if a measure that performs well in an evaluation framework can be applied more widely with the same success. In the work we present here, we first select a similarity measure by testing it against an extended TOEFL test. Then, we apply this measure for extracting automatically synonyms from a corpus and we compare our results to those of (Curran & Moens, 2002). Finally, we test the interest for synonym extraction of a method proposed in (Zhitomirsky-Geffet & Dagan, 2009) for improving the quality of distributional data.

Mots-clés : extraction de synonymes, similarité sémantique, méthodes distributionnelles.

Keywords: synonym extraction, semantic similarity, distributional methods.

1 Introduction

Cet article s'inscrit dans le champ de la sémantique lexicale et plus précisément de ce que l'on nomme « similarité sémantique lexicale ». L'objectif des travaux menés dans ce domaine de recherche est de déterminer dans quelle mesure deux mots sont proches sur le plan sémantique et, lorsque leur similarité est suffisamment forte, d'expliciter le type de la relation sémantique qui les unit. Une partie de ces travaux (cf. (Zesch & Gurevych, 2010) pour en avoir un panorama) exploitent pour ce faire des sources de connaissances plus ou moins structurées, tels que des dictionnaires. Dans cet article, nous nous focaliserons plus particulièrement sur les approches à base de corpus. La plupart d'entre elles s'appuient sur l'hypothèse distributionnelle selon laquelle des mots se trouvant dans des contextes similaires tendent à avoir des sens similaires (Firth, 1957). Dans le prolongement de (Grefenstette, 1994) et de (Lin, 1998), cette hypothèse

est généralement mise en œuvre en collectant des cooccurrences à partir de corpus de taille importante et en caractérisant chaque terme *T* de ces corpus par le vecteur de ses cooccurrents. Ceux-ci sont pondérés en fonction de la force de leur lien avec *T*. La similarité sémantique entre deux termes est évaluée quant à elle en calculant une mesure de similarité entre les vecteurs qui leur sont associés. Cette perspective a été adoptée et explorée en profondeur par des travaux tels que (Curran & Moens, 2002) ou (Weeds, 2003) en testant un nombre important de mesures de similarité et de fonctions de pondération des cooccurrents.

Quelques variantes de ce schéma de base ont été proposées, sans néanmoins sortir du cadre distributionnel. L'une d'elles est de nature probabiliste : chaque terme y est caractérisé par une distribution de probabilité sur ses cooccurrents et la similarité sémantique entre deux termes est évaluée par une distance entre leurs distributions respectives (Weeds, 2003). L'utilisation de méthodes de réduction de dimensions couvre un autre ensemble de variantes dans le cadre desquelles la similarité entre deux termes est évaluée dans l'espace sémantique issu de la réduction de dimensions réalisée. L'*Analyse Sémantique Latente* (Landauer & Dumais, 1997) et le *Random Indexing* (Salgren, 2006) sont les principaux représentants de ce courant auquel peut se rattacher indirectement les vecteurs conceptuels (Schwab *et al.*, 2007).

Les travaux concernant la similarité sémantique lexicale se définissent également par la façon dont ils évaluent les mesures de similarité sémantique qu'ils proposent. Une manière répandue de réaliser cette évaluation, utilisée initialement par (Landauer & Dumais, 1997), est d'appliquer ces mesures aux questions de synonymie de tests de type TOEFL. Ces questions sont constituées d'un mot cible et de quatre mots candidats parmi lesquels un synonyme du mot cible doit être identifié. Les systèmes développés ayant atteint un haut niveau de performance sur les questions issues du TOEFL (Turney *et al.*, 2003), diverses extensions de cette approche ont été explorées, soit par l'utilisation de questions issues d'autres tests similaires, comme l'ESL (Moraliyski & Dias, 2007), soit par la construction automatique d'un ensemble beaucoup plus large de questions en s'appuyant sur une ressource de référence telle que WordNet (Freitag *et al.*, 2005; Piasecki *et al.*, 2007), soit enfin par l'extension des questions à des relations de nature plus large que la synonymie comme les relations d'analogie présentes dans le test SAT (Turney, 2008).

Un autre mode commun d'évaluation des mesures de similarité sémantique est la comparaison de leurs résultats à une ressource de référence. Des jugements humains portant sur la similarité de couples de mots sont parfois utilisés dans cet esprit (Weeds, 2003) mais de tels jugements constituent en pratique des ressources rares et de petite taille. De ce fait, un mode d'évaluation plus indirect est généralement adopté (Curran & Moens, 2002; Lin, 1998) : les mesures de similarité à tester sont appliquées pour trouver les plus proches voisins sémantiques d'un mot et la pertinence de ces voisins est évaluée en les comparant à un ensemble de synonymes de référence issus de ressources telles que WordNet ou le thésaurus Roget.

L'objectif global du travail que nous présentons ici est d'extraire des synonymes de noms à partir de corpus en s'appuyant sur l'hypothèse distributionnelle, ce qui nécessite en premier lieu de choisir une mesure de similarité sémantique adéquate. En dépit du nombre significatif de travaux déjà réalisés dans ce domaine, comme nous avons pu le voir ci-dessus, il est difficile en pratique de transposer leurs résultats à notre problème : beaucoup d'entre eux ont été évalués sur des tests de type TOEFL, tâche moins exigeante que la nôtre, et les comparaisons avec des ressources de référence sont souvent données pour des ensembles de mots très fréquents (Curran & Moens, 2002) ou portent sur un ensemble de relations de proximité sémantique plus large que la simple synonymie (van der Plas & Bouma, 2004). Dans cet article, nous commençons par présenter les expérimentations que nous avons menées en anglais afin de trouver la mesure de similarité sémantique la plus efficace dans le cadre des contraintes que nous nous fixons en nous fondant un test de type TOEFL étendu. Nous rendons compte ensuite de l'application de cette mesure à l'extraction de synonymes à partir de corpus pour des noms. Enfin, nous étudions si des méthodes d'amé-

lioration de la caractérisation distributionnelle des mots fondées sur l'amorçage sont opérantes dans notre cas de figure. Notre objectif est ainsi d'avoir une vue plus globale de la notion de similarité sémantique, à l'instar de (Turney, 2008) ou de (Baroni & Lenci, 2009).

2 Test de mesures de similarité sémantique

2.1 Définition des mesures de similarité sémantique

Une mesure de similarité sémantique fondée sur l'hypothèse distributionnelle dépend fortement à la fois du corpus à partir duquel les données distributionnelles sont constituées et des moyens utilisés pour réaliser leur extraction. Bien que les corpus utilisés dans ce cadre tendent à devenir de plus en plus gros, ainsi que l'illustre (Pantel *et al.*, 2009), nous avons choisi délibérément un corpus de taille moyenne, le corpus AQUAINT-2, comprenant environ 380 millions de mots et constitué d'articles de journaux. Ce choix est motivé par le fait que la collecte de très gros corpus, outre les moyens que leur traitement induit, n'est pas toujours possible dans tous les domaines et pour toutes les langues.

Mesure de similarité des vecteurs de contexte		**Fonction de pondération des cooccurrents**	
Cosinus	$\frac{\sum_i poids(x_i)\cdot poids(y_i)}{\sqrt{\sum_j poids(x_j)^2 \cdot \sum_j poids(y_j)^2}}$	Information mutuelle (im)	$\log(\frac{p(x,c)}{p(x)\cdot p(c)})$
Jaccard	$\frac{\sum_i min(poids(x_i),poids(y_i))}{\sum_j max(poids(x_j),poids(y_j))}$	T-test	$\frac{p(x,c)-p(x)\cdot p(c)}{\sqrt{p(x)\cdot p(c)}}$
Jaccard†	$\frac{\sum_i min(poids(x_i),poids(y_i))}{\sum_i max(poids(x_i),poids(y_i))}$	Tf.Idf	$N(x,c)\cdot\log(\frac{N_x}{N_{x,c}})$
Dice	$\frac{2\cdot\sum_i min(poids(x_i),poids(y_i))}{\sum_j poids(x_j)+\sum_j poids(y_j)}$	$c : cooccurrent$	
Dice†	$\frac{2\cdot\sum_i min(poids(x_i),poids(y_i))}{\sum_i poids(x_i)+poids(y_i)}$		
Lin	$\frac{\sum_i poids(x_i)+poids(y_i)}{\sum_j poids(x_j)+\sum_j poids(y_j)}$		

TAB. 1 – Mesures de similarité des contextes et fonctions de pondération de leurs constituants testées[1]

Concernant l'extraction des données distributionnelles, nous avons opté là aussi pour une approche peu exigeante quant aux moyens utilisés. Bien qu'un certain nombre de travaux utilisent des analyseurs syntaxiques de surface, suivant en cela (Grefenstette, 1994) et (Lin, 1998), nous nous sommes limités à un pré-traitement linguistique des documents prenant la forme d'une lemmatisation et d'une sélection des mots pleins (noms, verbes et adjectifs) en nous appuyant sur l'outil *TreeTagger* (Schmid, 1994). La facilité d'accès à un analyseur syntaxique de surface pour l'anglais ne doit pas cacher en effet que ce type d'outils n'est pas encore largement répandu pour la plupart des autres langues. Les données distributionnelles que nous associons à chaque nom *N* représentatif du corpus AQUAINT-2[1] prennent donc la forme d'un vecteur de cooccurrents obtenu en comptabilisant les cooccurrences observées entre *N* et les noms,

[1]En pratique, seuls les mots de fréquence strictement supérieure à 10 sont retenus, aussi bien pour les noms cibles de nos évaluations que pour les cooccurrents constituant les vecteurs qui leur sont associés.

verbes et adjectifs d'une fenêtre de taille fixe centrée sur toutes les occurrences de *N* dans le corpus. Nous dénommons ce vecteur un *vecteur de contexte*.

Dans ce cadre, nous définissons une mesure de similarité sémantique entre un mot x et un mot y par le biais des quatre caractéristiques suivantes :

– une mesure de similarité des vecteurs de contexte associés à x et à y ;
– une fonction de pondération estimant l'importance d'un cooccurrent au sein d'un vecteur de contexte ;
– la taille de la fenêtre utilisée pour collecter les cooccurrents ;
– le seuil appliqué pour éliminer au sein des vecteurs de contexte les mots cooccurrant trop rarement avec le mot cible.

Le Tableau 1 donne la définition des différentes mesures de similarité des vecteurs de contexte et des différentes fonctions de pondération des cooccurrents en leur sein que nous avons testées. S'y ajoute la mesure proposée par (Ehlert, 2003) qui, de par sa nature probabiliste, échappe au schéma ci-dessus puisqu'elle repose sur la probabilité des cooccurrents et non sur un poids défini de façon externe.

2.2 Résultats et évaluation

Comme indiqué en introduction, notre sélection d'une mesure de similarité sémantique en vue de l'extraction de synonymes s'est opérée sur la base d'un test de type TOEFL étendu, et plus précisément du WordNet-Based Synonymy Test (WBST), proposé par (Freitag *et al.*, 2005)[2]. Le WBST a été produit en générant automatiquement un large ensemble de questions de type TOEFL à partir des synonymes de WordNet. (Freitag *et al.*, 2005) montre que ce test est plus difficile que le test originel du TOEFL dont les 80 questions ont été initialement utilisées dans (Landauer & Dumais, 1997). La partie du WBST se limitant aux noms, auxquels nous nous restreignons dans ce travail, comprend 9 887 questions. Toutes les combinaisons possibles entre les mesures de similarité des vecteurs de contexte et les fonctions de pondération des cooccurrents ont été testées avec des tailles de fenêtre comprises entre 1 et 5 et des seuils fréquentiels sur les cooccurrents allant de 1 à 5. En pratique, pour chaque question du test, la mesure de similarité testée est calculée entre l'entrée constituant la question et les quatre choix possibles. Ces choix sont ensuite triés selon l'ordre décroissant de leur score et celui ayant la similarité la plus forte est retenu comme candidat synonyme. Dans les rares cas où les données distributionnelles ne permettent pas de départager les différents choix (entre 3,7 et 6,7% des cas selon les mesures), un tirage aléatoire est réalisé. La mesure d'évaluation utilisée est tout simplement le pourcentage de candidats synonymes exacts, ce que l'on peut aussi voir comme la précision au rang 1 puisque nos mesures ordonnancent les choix. Le Tableau 2 donne pour chaque mesure de similarité entre vecteurs de contexte les trois autres paramètres permettant d'obtenir les meilleurs résultats sur le WBST.

La première observation notable à propos de cette évaluation est que pour toutes les mesures de similarité entre vecteurs de contexte, les meilleurs résultats sont obtenus pour une taille de fenêtre et un seuil sur les cooccurrents égaux à 1[3]. Ceci laisse donc à penser que la notion de similarité sémantique est plutôt caractérisée par des cooccurrents de très courte portée parmi lesquels seuls les cooccurrents dont la présence est la plus probablement fortuite sont écartés. Le deuxième constat tiré du Tableau 2 est que le couple

[1] i : index sur les cooccurrents communs à x et y ; j : index sur tous les cooccurrents de x et de y ; $N(x, c)$: fréquence de c comme cooccurrent de x ; N_x : nombre de mots ; $N_{x,c}$: nombre de mots ayant c comme cooccurrent.

[2] Disponible à l'adresse `http://www.cs.cmu.edu/~dayne/wbst-nanews.tar.gz`.

[3] Ce qui conduit à supprimer globalement la moitié environ des cooccurrents.

taille fenêtre		1			3			5		
seuil fréquence		1	3	5	1	3	5	1	3	5
cosinus	im	**71,6**	69,7	67,6	65,7	63,7	62,8	62,5	60,6	59,4
	t-test	**68,9**	66,7	65,0	65,4	64,6	63,8	63,3	62,9	62,0
	tf.idf	**64,0**	63,1	62,0	63,3	62,9	62,5	62,6	62,4	61,7
ehlert	–	**70,2**	68,5	66,2	68,9	67,2	65,9	66,9	65,9	64,4
jaccard	im	**64,8**	63,0	61,7	57,1	55,0	54,1	54,6	52,6	51,3
	t-test	**68,1**	65,8	63,9	61,3	58,8	57,7	58,4	55,9	54,6
	tf.idf	**54,2**	53,9	53,6	49,7	49,6	49,3	48,0	47,9	47,4
dice	im	**64,8**	63,0	61,7	57,1	55,0	54,1	54,6	52,6	51,3
	t-test	**68,1**	65,8	63,9	61,3	58,8	57,7	58,4	55,9	54,6
	tf.idf	**54,2**	53,9	53,6	49,7	49,6	49,3	48,0	47,9	47,4
lin	im	**65,6**	63,5	61,7	57,0	54,6	53,6	54,2	52,1	51,1
	t-test	**67,3**	65,3	63,3	61,0	59,5	58,9	58,5	57,3	55,9
	tf.idf	**60,6**	59,6	58,3	57,9	56,6	55,9	56,6	54,9	53,9
dice†	im	**65,0**	63,2	61,5	58,7	57,5	57,0	56,5	55,9	55,3
	t-test	**66,0**	64,3	62,3	59,7	57,9	57,0	57,5	56,0	55,1
	tf.idf	51,6	52,3	**52,7**	48,4	47,9	48,3	47,2	47,2	46,6
jaccard†	im	**56,1**	54,7	53,2	54,3	54,3	53,4	54,0	54,3	53
	t-test	39,6	37,9	38,2	46,7	43,7	42,2	**48,1**	45,7	43,0
	tf.idf	35,3	34,3	34,4	40,2	38,1	37,3	**41,4**	39,7	38,4

TAB. 2 – Évaluation des mesures de similarité sémantique

Information mutuelle – Cosinus et la mesure de *Ehlert* obtiennent toutes deux les meilleurs résultats, conformément aux conclusions de (Freitag *et al.*, 2005), établies également avec des cooccurrences « graphiques ». Néanmoins, (Freitag *et al.*, 2005) donnait l'avantage à *Ehlert* par rapport au *Cosinus* et nous observons la tendance inverse. Plus précisément, notre meilleur résultat pour le *Cosinus* est égal à leur meilleur résultat pour *Elhert* (en dehors d'une optimisation supervisée également proposée). Par ailleurs, les performances rapportées dans (Freitag *et al.*, 2005) ont été obtenues avec un corpus d'un milliard de mots environ, c'est-à-dire beaucoup plus grand que le nôtre, et la fréquence des noms du WBST dans leur corpus était au moins égale à 1000 tandis que nous avons écarté seulement les mots de fréquence inférieure à 11. Enfin, les performances de notre meilleure mesure se comparent favorablement à celles de (Broda *et al.*, 2009), qui s'appuie sur des cooccurrences syntaxiques : pour les noms de fréquence supérieure à 10 dans leur corpus de référence, le British National Corpus (BNC), un corpus de 100 millions de mots, un pourcentage d'exactitude de 68,04 est obtenu sur un ensemble de 14 376 noms faisant partie du WBST.

3 Extraire des synonymes grâce à une mesure sémantique

Les résultats ci-dessus montrent que nous avons construit une mesure de similarité sémantique distributionnelle dont les performances, dans un cadre d'évaluation standard pour des mesures de ce type, sont au moins aussi élevées que celles de l'état de l'art, tout en mobilisant des moyens moindres. Nous rendons compte maintenant dans cette section de l'application de cette mesure au problème de l'extraction

automatique de synonymes à partir de corpus.

Notre processus d'extraction est simple : les synonymes possibles d'un mot sont trouvés en recherchant les *N* plus proches voisins de ce mot selon notre mesure de similarité. Plus précisément, cette recherche consiste à appliquer cette mesure de similarité entre ce mot cible et tous les autres mots du vocabulaire considéré ayant la même catégorie morpho-syntaxique (ici, les noms d'AQUAINT-2). Finalement, tous ces mots sont triés suivant leur valeur de similarité et seuls les *N* plus proches voisins, avec $N = 100$ dans nos expérimentations, sont conservés en tant que candidats synonymes[4]. Puisque nous nous appuyons sur le *Cosinus* au niveau de notre mesure de similarité sémantique, il serait possible de rendre cette recherche linéaire plus efficace, sans perdre en qualité de résultat, en utilisant une méthode du type « recherche exhaustive des paires similaires » telle que celle présentée dans (Bayardo *et al.*, 2007), ou pour un domaine plus proche du nôtre, dans (Pantel *et al.*, 2009) pour une transposition aux très gros volumes de textes.

fréq.	**réf.**	**# mots**	**# syn. cibles**	**% syn. trouvés**	**R-préc.**	**MAP**	**P@1**	**P@5**	**P@10**	**P@100**
> 10	W	10473	29947	24,6	0,082	0,098	0,117	0,051	0,034	0,007
(tous)	M	9216	460923	9,5	0,067	0,032	0,241	0,164	0,130	0,048
# 14670	WM	12243	473833	9.8	0,077	0,056	0,225	0,140	0,108	0,038
> 1000	W	3690	13509	28,3	**0,111**	**0,125**	0,171	0,077	0,051	0,010
# 4378	M	3732	258836	11,4	0,102	0,049	**0,413**	**0.280**	**0,219**	**0,079**
	WM	4164	263216	11,5	0,110	0,065	**0.413**	0,268	0,208	0,073
$100 <$	W	3732	9562	28,6	0,104	**0,125**	0,136	0,058	0,037	0,007
≤ 1000	M	3306	136467	9,3	0,064	0,031	0,187	0,131	0,104	0,038
# 5175	WM	4392	140750	9,8	0,092	0,073	0,209	0,123	0,093	0,031
≤ 100	W	3051	6876	11,9	0,021	0,033	0,026	0,012	0,009	0,003
# 5117	M	2178	65620	2,8	0,012	0,005	0,025	0,015	0,015	0,008
	WM	3687	69867	3,5	0,021	0,024	0,033	0,017	0,015	0,007

TAB. 3 – Évaluation de l'extraction des synonymes

Le Tableau 3 montre les résultats de l'application de notre mesure de similarité sémantique à l'extraction de synonymes. Nous avons pris comme référence deux ressources : WordNet (W), dans sa version 3.0, et le thésaurus Moby (M). Notre but étant en premier lieu d'évaluer la capacité d'une mesure sémantique distributionnelle à extraire des synonymes d'un corpus, nous avons filtré ces ressources en éliminant en leur sein les termes ne faisant pas partie du vocabulaire des noms simples retenus pour construire nos données distributionnelles. Nous avons également créé une ressource fusionnant WordNet et le thésaurus Moby (WM). Il est à noter que si les synsets de WordNet sont par définition constitués de synonymes, les entrées du thésaurus Moby contiennent également des mots dits liés. Dans ce cas, notre évaluation s'étend donc à la notion de proximité sémantique, au delà de la stricte synonymie.

La fréquence des mots en relation avec la taille des corpus étant une donnée importante dans les approches distributionnelles, nous donnons des résultats globaux mais nous les différencions également suivant trois tranches fréquentielles à peu près équilibrées en termes d'effectifs (cf. $1^{ère}$ colonne) : les mots très fréquents (fréquence > 1000), moyennement fréquents ($100 <$ fréquence ≤ 1000) et faiblement fréquents

[4]De manière indicative, cette recherche est réalisée approximativement en 4 heures sur 48 cœurs d'un cluster.

($10 <$ fréquence ≤ 100). La $3^{ème}$ colonne du Tableau 3 donne le nombre d'entrées de chaque ressource pour lesquelles l'évaluation a été réalisée, tous les noms du vocabulaire du corpus AQUAINT-2 de fréquence supérieure à 10 n'apparaissant pas dans nos ressources de référence. La $4^{ème}$ colonne de ce même tableau correspond au nombre de synonymes à trouver dans chaque ressource pour l'ensemble des entrées faisant partie du vocabulaire AQUAINT-2 tandis que sa $5^{ème}$ colonne fournit le pourcentage de synonymes effectivement trouvés parmi les 100 voisins sémantiques de chaque entrée de notre base distributionnelle. Ces voisins étant ordonnés, il est possible de faire le parallèle entre la recherche de synonymes et la recherche de documents en recherche d'information et de réutiliser ainsi les métriques d'évaluation classiquement utilisées pour cette dernière. C'est l'objet des dernières colonnes du Tableau 3 : 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 avoir examiné les 1, 5, 10 et 100 premiers voisins).

Une première vue d'ensemble du Tableau 3 laisse apparaître que malgré ses performances intéressantes sur des tests de similarité sémantique, la mesure que nous avons sélectionnée n'obtient dans l'absolu que des résultats assez modestes lorsqu'elle est appliquée au problème de l'extraction de synonymes. Cette faiblesse est observable aussi bien au niveau du taux de rappel des synonymes (environ 25% pour WordNet et 10% pour le thésaurus Moby) qu'au niveau de leur rang parmi les voisins sémantiques (cf. R-précision, MAP et P@{1,5,10,100}). Ce constat a une portée plus générale que notre travail spécifique dans la mesure où la mesure sémantique que nous avons utilisée peut être considérée comme classique. Cette faiblesse générale cache néanmoins des différences importantes suivant la fréquence des mots. On observe ainsi une corrélation claire entre le niveau des résultats et la fréquence des mots dans le corpus de constitution des données distributionnelles : plus cette fréquence est élevée, plus la mesure de similarité est efficace du point de vue de l'extraction des synonymes et plus son caractère sémantique semble s'affirmer si l'on considère cette tâche représentative d'un tel caractère. Même si cette constatation semble plaider en faveur de l'accroissement de la taille des corpus, elle n'écarte pas l'idée d'un comportement distributionnel différent des mots plus rares à prendre en compte de façon spécifique.

Sur un autre plan, le Tableau 3 montre que le profil des ressources de référence considérées a aussi son importance quant aux résultats obtenus. WordNet fournit un nombre restreint de synonymes stricts pour chaque nom (2,8 en moyenne) tandis que le thésaurus Moby contient pour chaque entrée un nombre beaucoup plus important de synonymes et de mots liés (50 en moyenne). Cette différence explique que la même mesure obtient, pour des mots de fréquence supérieure à 1 000, une précision au rang 1 égale à 0,413 pour le thésaurus Moby et de seulement 0,171 pour WordNet.

En l'absence d'un cadre d'évaluation clairement reconnu pour ce type de tâches, la comparaison avec d'autres travaux n'est pas aisée. Un certain nombre d'entre eux utilisent en effet pour leurs évaluations un ensemble de relations sémantiques de référence plus large que la synonymie. C'est le cas de (van der Plas & Bouma, 2004), qui adopte la version néerlandaise d'EuroWordNet comme référence mais en s'appuyant sur une distance intégrant les relations d'hyperonymie. (Pantel *et al.*, 2009) s'intéresse pour sa part à la notion d'ensemble d'entités (*Entity Sets*), sous-tendue par une gamme de relations très étendue. (Curran & Moens, 2002) est en revanche un travail le plus directement comparable au nôtre. Il met en œuvre diverses mesures de similarité fondées sur des cooccurrences syntaxiques qui sont ensuite évaluées du point de vue de l'extraction de voisins sémantiques en adoptant comme référence la fusion des thésaurus Roget, Moby et Macquarie. Cette évaluation porte sur 70 noms choisis au hasard dans WordNet en respectant

une diversité de fréquences et de degrés de spécificité. Parmi les différentes mesures testées, la meilleure performance (Dice† + T-test) obtenue est une précision au rang 1 de 0,76, au rang 5 de 0,52 et au rang 10 de 0,45 pour 70 noms, à comparer avec 0,413, 0,280 et 0,219 dans notre cas pour 3 732 noms. Il faut souligner néanmoins qu'outre la différence de taille du jeu de test, les références utilisées sont différentes, ce qui a une grande influence sur le niveau des résultats ainsi que nous l'avons illustré ci-dessus. Pour nos 3 732 noms, le thésaurus Moby fournit en moyenne 69 synonymes tandis que pour les 70 noms de (Curran & Moens, 2002), ce nombre monte à 331. On constate en outre que le taux de rappel est différent dans les deux évaluations : il est de 8,3% pour (Curran & Moens, 2002) tandis qu'il est de 11,4% dans notre cas. Même s'il est difficile d'estimer le niveau exact d'influence des différences de richesse des ressources utilisées, cette comparaison suggère que l'utilisation de cooccurrences syntaxiques permet d'obtenir une meilleure précision dans l'extraction de synonymes tandis que les cooccurrences graphiques tendraient à en favoriser le rappel.

4 Test d'amélioration d'une mesure de similarité sémantique

Le niveau des performances de notre mesure de similarité sémantique pour l'extraction de synonymes nous a conduit à examiner si des méthodes de repondération des vecteurs de cooccurrents telles que celles présentées dans (Broda *et al.*, 2009) ou (Zhitomirsky-Geffet & Dagan, 2009) pouvaient améliorer nos résultats. Nous nous sommes focalisés sur (Zhitomirsky-Geffet & Dagan, 2009) dans la mesure où les améliorations rapportées par (Broda *et al.*, 2009) sont modestes, même si l'évaluation dans (Zhitomirsky-Geffet & Dagan, 2009) portait sur des mots en relation d'implication (*entailment*) et non de synonymie.

fréq.	réf.	% syn. trouvés	R-préc.	MAP	P@1	P@5	P@10	P@100
> 10	W	21,5	0,060	0,074	0,087	0,039	0,026	0,006
(tous)	M	9,0	–	0,030	0,211	0,144	0,114	0,0445
> 1000	W	25,1	0,088	0,099	0,137	0,061	0,040	0,009
	M	10,5	–	0,045	0,360	0,245	0,192	0,073
$100 <$	W	23,4	0,069	0,087	0,092	0,040	0,025	0,006
≤ 1000	M	8,7	0,055	0,028	0,163	0,113	0,091	0,036
≤ 100	W	11,6	0,017	0,028	0,022	0,011	0,008	0,002
	M	**4,0**	0,015	0,006	0,034	0,022	0,019	0,012

TAB. 4 – Évaluation de l'extraction des synonymes après repondération des cooccurrents

La méthode proposée par (Zhitomirsky-Geffet & Dagan, 2009) est fondée sur un principe d'amorçage. Son idée générale est d'exploiter les valeurs de similarité calculées entre le vecteur de contexte VC_t d'un mot cible t et les vecteurs de contexte VC_i des autres mots i de la base distributionnelle considérée pour favoriser les cooccurrents de VC_t présents dans les vecteurs de contexte des mots les plus similaires à t. Plus formellement, la méthode prend la forme d'une repondération des cooccurrents c_j de t telle que :

$$VC_t(c_j) = \sum_{(i \neq t) \wedge (c_j \neq 0)} sim(t, i) \quad (1)$$

où $VC_t(c_j)$ est le poids du cooccurrent c_j du vecteur de contexte de t.

En outre, un seuil de similarité minimale est appliqué aux voisins sémantiques i de t et un second seuil imposant un poids minimal pour la prise en compte des cooccurrents d'un voisin sémantique i est également fixé. Nous avons établi la valeur de ces seuils dans notre cas par le biais d'un processus d'optimisation sur 60 entrées de notre base distributionnelle, choisies de manière équilibrée sur le plan fréquentiel.

Le Tableau 4 montre l'impact de cette procédure de repondération des cooccurrents des vecteurs de contexte sur l'évaluation présentée à la Section 3. Il apparaît clairement que si cette procédure s'est avérée très intéressante dans le cas des relations d'implication, ses résultats sont décevants concernant les synonymes puisqu'elle entraîne une nette chute de tous les résultats (à l'exception d'une valeur). Plusieurs explications peuvent être avancées. L'évaluation présentée dans (Zhitomirsky-Geffet & Dagan, 2009) concernait la capacité à reproduire des jugements humains sur la présence de relations d'implication entre deux mots. Outre le nombre réduit de mots sources (30 mots de fréquence supérieure à 500 pour 3 200 relations), il faut noter que cette tâche est plus proche des tests du TOEFL que de l'extraction de synonymes. Par ailleurs, les données distributionnelles étaient constituées dans ce cas de cooccurrents syntaxiques. Enfin, il est probable que cette procédure d'amorçage, que l'on peut voir comme une forme d'amplificateur, n'est effective que si les similarités initiales entre les mots sont suffisamment significatives, ce qui n'est sans doute pas le cas ici pour les mots de faible fréquence.

5 Conclusion et perspectives

Dans cet article, nous avons présenté dans un premier temps les expérimentations que nous avons réalisées afin de sélectionner la mesure de similarité sémantique reposant sur l'hypothèse distributionnelle qui soit la plus à même de rendre compte des relations de proximité sémantique entre mots. Cette sélection s'est appuyée sur un test de type TOEFL étendu. Nous avons ensuite appliqué cette mesure au problème de l'extraction automatique de synonymes en prenant comme référence deux ressources complémentaires : WordNet et le thésaurus Moby. Les résultats de cette application, compatibles avec l'état de l'art dans ce domaine, montrent que les tests de similarité sémantique utilisés habituellement ne garantissent pas nécessairement de bonnes performances pour une tâche comme l'extraction de synonymes. Enfin, nous avons montré que la méthode proposée dans (Zhitomirsky-Geffet & Dagan, 2009) pour améliorer la qualité des données distributionnelles n'est pas opérante pour une telle tâche.

Le prolongement le plus direct de ce travail est l'utilisation de cooccurrences syntaxiques en lieu et place des cooccurrences graphiques afin de déterminer si les premières apportent véritablement le surcroît de précision que notre analyse des résultats de (Curran & Moens, 2002) semble suggérer. Si une telle amélioration est constatée, nous envisageons de mener des travaux complémentaires concernant l'amélioration des données distributionnelles, en commençant par un test de la méthode de (Zhitomirsky-Geffet & Dagan, 2009) avec ces cooccurrences syntaxiques et au delà, la prise en compte de nouveaux critères comme l'utilisation de sens de mots discriminés automatiquement.

Références

BARONI M. & LENCI A. (2009). One distributional memory, many semantic spaces. In *EACL 2009 Workshop on Geometrical Models of Natural Language Semantics*, p. 1–8, Athens, Greece.

BAYARDO R. J., MA Y. & SRIKANT R. (2007). Scaling up all pairs similarity search. In *16th international conference on World Wide Web (WWW'07)*, p. 131–140, Banff, Alberta, Canada.

BRODA B., PIASECKI M. & SZPAKOWICZ S. (2009). Rank-Based Transformation in Measuring Semantic Relatedness. In *22nd Canadian Conference on Artificial Intelligence*, p. 187–190.

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

EHLERT B. (2003). Making Accurate Lexical Semantic Similarity Judgments Using Wordcontext Co-occurrence Statistics. Master's thesis, University of California, San Diego, USA.

FIRTH J. (1957). *Studies in Linguistic Analysis*, chapter A synopsis of linguistic theory 1930-1955, p. 1–32. Blackwell : Oxford.

FREITAG D., BLUME M., BYRNES J., CHOW E., KAPADIA S., ROHWER R. & WANG Z. (2005). New experiments in distributional representations of synonymy. In *Ninth Conference on Computational Natural Language Learning (CoNLL 2005)*, p. 25–32, Ann Arbor, Michigan, USA.

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

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

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

MORALIYSKI R. & DIAS G. (2007). One Sense Per Discourse for Synonym Detection. In *5th International Conference Recent Advances in Natural Language Processing (RANLP 2007)*, Borovets, Bulgaria.

PANTEL P., CRESTAN E., BORKOVSKY A., POPESCU A.-M. & VYAS V. (2009). Web-Scale Distributional Similarity and Entity Set Expansion. In *2009 Conference on Empirical Methods in Natural Language Processing*, p. 938–947, Singapore.

PIASECKI M., SZPAKOWICZ S. & BRODA B. (2007). Extended Similarity Test for the Evaluation of Semantic Similarity Functions. In *Language Technology Conference (LTC)*, p. 104–108, Poznań, Poland.

SALGREN M. (2006). *The Word-space model*. PhD thesis, Stockholm University.

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

SCHWAB D., TZE L. L. & LAFOURCADE M. (2007). Les vecteurs conceptuels, un outil complémentaire aux réseaux lexicaux. In *14ème Conférence sur le Traitement automatique des langues naturelles (TALN 2007)*, p. 293–302, Toulouse.

TURNEY P., LITTMAN M., BIGHAM J. & SHNAYDER V. (2003). Combining independent modules to solve multiple-choice synonym and analogy problems. In *4th International Conference on Recent Advances in Natural Language Processing (RANLP-03)*, p. 482–489, Borovets, Bulgaria.

TURNEY P. D. (2008). A Uniform Approach to Analogies, Synonyms, Antonyms, and Association. In *COLING 2008*, p. 905–912, Manchester, UK.

VAN DER PLAS L. & BOUMA G. (2004). Syntactic contexts for finding semantically related words. In *Fifteenth Computational Linguistics in the Netherlands Meeting (CLIN 2004)*, Leiden, Netherlands.

WEEDS J. (2003). *Measures and Applications of Lexical Distributional Similarity*. PhD thesis, Department of Informatics, University of Sussex.

ZESCH T. & GUREVYCH I. (2010). Wisdom of crowds versus wisdom of linguists - measuring the semantic relatdness of words. *Natural Language Engineering*, **16**(1), 25–59.

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