TALN 2011, Montpellier, 27 juin – 1er juillet 2011

Utiliser l'amorçage pour améliorer une mesure de similarité sémantique

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

Résumé. Les travaux sur les mesures de similarité sémantique de nature distributionnelle ont abouti à un certain consensus quant à leurs performances et ont montré notamment que leurs résultats sont surtout intéressants pour des mots de forte fréquence et une similarité sémantique étendue, non restreinte aux seuls synonymes. Dans cet article, nous proposons une méthode d'amélioration d'une mesure de similarité classique permettant de rééquilibrer ses résultats pour les mots de plus faible fréquence. Cette méthode est fondée sur un mécanisme d'amorçage : un ensemble d'exemples et de contre-exemples de mots sémantiquement liés sont sélectionnés de façon non supervisée à partir des résultats de la mesure initiale et servent à l'entraînement d'un classifieur supervisé. Celui-ci est ensuite utilisé pour réordonner les voisins sémantiques initiaux. Nous évaluons l'intérêt de ce réordonnancement pour un large ensemble de noms anglais couvrant différents domaines fréquentiels.

Abstract. Work about distributional semantic similarity measures has now widely shown that such measures are mainly reliable for high frequency words and for capturing semantic relatedness rather than strict semantic similarity. In this article, we propose a method for improving such a measure for middle and low frequency words. This method is based on a bootstrapping mechanism : a set of examples and counter-examples of semantically related words are selected in an unsupervised way from the results of the initial measure and used for training a supervised classifier. This classifier is then applied for reranking the initial semantic neighbors. We evaluate the interest of this reranking for a large set of english nouns with various frequencies.

Mots-clés : Extraction de voisins sémantiques, similarité sémantique, méthodes distributionnelles.

Keywords: Semantic neighbor extraction, semantic similarity, distributional methods.

1 Introduction

Le travail présenté ici prend place dans le domaine de la sémantique lexicale et plus particulièrement de la similarité sémantique au niveau lexical. La notion de *similarité sémantique* couvre, aussi bien du point de vue de sa définition que de sa caractérisation, une pluralité d'approches. Concernant sa définition, la dichotomie principale se fait entre une similarité reposant sur des relations sémantiques de nature paradigmatique (hyperonymie, synonymie, etc) et une similarité reposant sur des relations sémantiques de nature syntagmatique (relations de cohésion lexicale au statut théorique plus flou). Cette dichotomie recouvre celle faite entre les notions de *semantic similarity* et de *semantic relatedness*. Bien que justifiée par la différence de nature des relations impliquées, cette différenciation n'est pas en pratique toujours très nette, en particulier au niveau de l'évaluation. Dans le cadre du travail présenté ici, nous nous focalisons plus spécifiquement sur une caractérisation distributionnelle de la similarité sémantique. Les recherches la concernant ont montré que les relations sémantiques couvertes par une telle approche relèvent à la fois de l'axe paradigmatique et de l'axe syntagmatique. À défaut donc de nous restreindre à un seul type de relations, nous nous efforcerons de distinguer au niveau des évaluations les proximités sémantiques relevant de relations comme la synonymie de celles impliquant un ensemble plus large de relations sémantiques.

Au-delà d'une mise en œuvre « classique » de l'approche distributionnelle telle qu'elle est incarnée par (Curran & Moens, 2002), un certain nombre de propositions ont été faites pour améliorer les résultats dans le cadre de ce paradigme. Une part significative de ces propositions portent sur la pondération des éléments constitutifs des contextes associés aux mots mais un certain nombre impliquent des changements plus profonds. L'utilisation de techniques de réduction de dimensions, en l'occurrence l'analyse sémantique latente dans (Padó & Lapata, 2007), ou la redéfinition de l'approche distributionnelle dans le cadre bayésien dans (Kazama *et al.*, 2010), se classent

dans cette seconde catégorie. La première est quant à elle représentée par (Broda *et al.*, 2009) au travers du passage de valeurs de poids à des valeurs de rang ou par (Zhitomirsky-Geffet & Dagan, 2009), repris et étendu par (Yamamoto & Asakura, 2010), qui utilise une technique d'amorçage pour modifier les poids des éléments des contextes distributionnels en fonction des voisinages sémantiques calculés.

Dans sa perspective générale, le travail présenté ici se rapproche de (Zhitomirsky-Geffet & Dagan, 2009) du point de vue de l'utilisation de l'amorçage mais adopte une méthode différente : les « meilleurs » voisins sémantiques ne servent plus en effet à modifier directement les poids des éléments des contextes distributionnelles mais plus indirectement à entraîner un classifieur supervisé, à la manière de (Hagiwara *et al.*, 2009), pour apprendre une mesure de similarité améliorant certaines déficiences de la mesure initiale.

2 Une mesure de similarité sémantique distributionnelle

2.1 Définition

Pour qu'un mécanisme d'amorçage puisse être mis en œuvre, il est nécessaire de s'appuyer sur un processus initial produisant des résultats d'un niveau suffisamment élevé, au moins dans un certain périmètre. Dans le cas présent, cette exigence implique de disposer d'une mesure de similarité sémantique montrant un bon niveau de résultat dans les évaluations permettant classiquement de juger de telles mesures tels que des tests de type TOEFL ou l'extraction de voisins sémantiques pour la construction de thésaurus. Dans (Ferret, 2010), nous avons défini, au terme d'une sélection effectuée grâce un test de type TOEFL, une mesure de similarité sémantique distributionnelle présentant des résultats au moins comparables aux résultats de l'état de l'art pour des mesures de même nature. Cette mesure, définie pour l'anglais à partir du corpus AQUAINT-2, un corpus journalistique d'une taille de 380 millions de mots, présente les caractéristiques suivantes :

– contextes distributionnels formés de cooccurrents graphiques, c'est-à-dire des mots capturés dans une fenêtre de taille fixe centrée sur toutes les occurrences du mot cible. Ces cooccurrents sont plus précisément restreints aux mots pleins des textes, autrement dit les noms, verbes et adjectifs ;
– taille de fenêtre = 1, *i.e.* cooccurrents à très faible portée ;
– filtrage très conservateur des contextes : suppression des cooccurrents n'apparaissant qu'une seule fois ;
– utilisation de l'*information mutuelle* pour la pondération des cooccurrents formant les contextes ;
– mesure de similarité entre les contextes, pour évaluer la similarité sémantique des mots = *cosinus*.

Ces données distributionnelles n'ont été constituées que pour des noms de fréquence strictement supérieure à 10.

2.2 Application et évaluation

Une des applications des mesures de similarité telle que celle de la section précédente, application qui permet aussi de les évaluer, est l'extraction de voisins sémantiques. Dans (Ferret, 2010) comme dans la plupart des travaux similaires, cette extraction est réalisée de façon très simple : la mesure de similarité retenue est calculée entre chaque mot cible et chacun de ses voisins potentiels. Ces voisins sont ensuite triés dans l'ordre décroissant des valeurs de cette mesure et les N (ici, $N = 100$) premiers sont conservés comme voisins sémantiques du mot cible.

Le tableau 1 montre les résultats de l'application de la mesure de similarité sémantique de la section précédente à l'extraction de voisins sémantiques. Deux ressources de référence complémentaires sont considérées : WordNet (W), dans sa version 3.0, permet de caractériser la similarité fondée sur des relations paradigmatiques tandis que le thésaurus Moby (M) 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 trois tranches fréquentielles à peu près équilibrées en termes d'effectifs : les mots très fréquents (fréquence > 1000), moyennement fréquents ($100 <$ fréquence ≤ 1000) et faiblement fréquents ($10 <$ fréquence ≤ 100). Ces résultats se déclinent sous la forme de différentes mesures. La $5^{ème}$ colonne donne ainsi 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
> 10	W	10 473	2,9	24,6	8,2	9,8	11,7	5,1	3,4	0,7
(tous)	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
> 1000	W	3 690	3,7	28,3	11,1	12,5	17,2	7,7	5,1	1,0
# 4 378	M	3 732	69,4	11,4	10,2	4,9	41,3	28,0	21,9	7,9
	WM	4 164	63,2	11,5	11,0	6,5	41,3	26,8	20,8	7,3
$100 <$	W	3 732	2,6	28,6	10,4	12,5	13,6	5,8	3,7	0,7
≤ 1000	M	3 306	41,3	9,3	6,5	3,1	18,7	13,1	10,4	3,8
# 5 175	WM	4 392	32,0	9,8	9,3	7,4	20,9	12,3	9,3	3,2
≤ 100	W	3 051	2,3	11,9	2,1	3,3	2,6	1,2	0,9	0,3
# 5 117	M	2 178	30,1	2,8	1,2	0,5	2,5	1,5	1,5	0,9
	WM	3 687	18,9	3,5	2,1	2,4	3,3	1,7	1,5	0,7

TAB. 1 – Évaluation de l'extraction de voisins sémantiques

étant ordonnés, il est en outre possible de réutiliser les métriques d'évaluation classiquement utilisées en recherche d'information en faisant jouer aux mots cibles le rôle de requêtes et aux voisins celui des documents. C'est l'objet des dernières colonnes du tableau 1 : 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). Toutes ces valeurs sont données en pourcentage.

3 Améliorer une mesure de similarité sémantique

Pour reprendre rapidement l'analyse du tableau 1 faite dans (Ferret, 2010), le constat d'une faiblesse d'ensemble des résultats s'impose, en particulier pour les mots peu fréquents, et la nécessité de les améliorer apparaît clairement. (Ferret, 2010) rapporte ainsi une tentative dans ce sens transposant au problème de l'extraction de voisins sémantiques la méthode d'amorçage présentée dans (Zhitomirsky-Geffet & Dagan, 2009) et appliquée initialement à l'extraction de mots en relation d'implication textuelle. Cette tentative ne fut néanmoins pas concluante, aboutissant à une dégradation globale des résultats plutôt qu'à leur amélioration.

La méthode d'amélioration présentée ici reprend l'idée d'amorçage de (Zhitomirsky-Geffet & Dagan, 2009) mais l'applique différemment. (Hagiwara *et al.*, 2009) a montré qu'il est possible d'entraîner et d'appliquer avec un bon niveau de performance un classifieur de type Machine à Vecteurs de Support (SVM) pour décider si deux mots sont ou ne sont pas synonymes. La notion de synonymie est à prendre ici au sens large compte tenu des ressources utilisées pour l'évaluation. 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 *et al.*, 2009), 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 cet ensemble. Cette mesure n'offre pas de critère évident pour discriminer les mots sémantiquement liés mais le tableau 1 fournit des informations pour sélectionner un ensemble d'exemples et de contre-exemples 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. Le point clé de l'amélioration des résultats par ce moyen est de sélectionner de façon non supervisée, les ressources de référence ne servant que pour l'évaluation, un nombre suffisamment élevé de bons exemples et contre-exemples pour compenser les erreurs inhérentes à une telle sélection.

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 *et al.*, 2009) : 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.

Concernant la sélection des exemples et des contre-exemples, le tableau 1 montre clairement que le cas des exemples est beaucoup plus problématique que celui des contre-exemples dans la mesure où le nombre de mots sémantiquement liés diminue très fortement dès que l'on considère des voisins de rang un peu élevé. Dans les expérimentations de la section 4, nous avons ainsi pris comme exemples négatifs des couples {mot cible, voisin du mot cible de rang 10}. Le choix d'un rang supérieur garantirait un nombre plus faible de faux contre-exemples (*i.e.* couples de mots en fait 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.

Pour la sélection des exemples, le tableau 1 impose un double constat : les chances de trouver un voisin sémantiquement proche sont d'autant plus importantes que la fréquence du mot cible est élevée et que le rang du voisin est faible. Suivant cette logique, nous avons retenu comme premier ensemble d'exemples tous les couples {mot cible de fréquence > 1000, voisin de rang 1 du mot cible}, soit un total de 4 378 exemples pour lesquels 4 378 contre-exemples définis selon le principe décrit ci-dessus ont été également construits. Dans (Hagiwara *et al.*, 2009), le rapport nombre d'exemples / nombre de contre-exemples est égal à 6,5 environ mais il n'est pas apparu dans notre cas qu'un tel déséquilibre permettait d'obtenir des résultats significativement meilleurs. Compte tenu de notre nombre important d'exemples et de caractéristiques associées à chacun d'eux, nous avons donc opté pour un nombre identique d'exemples et de contre-exemples.

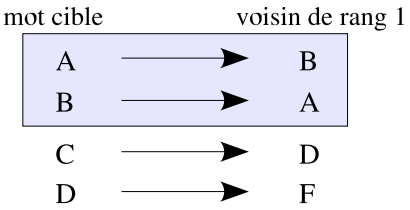

FIG. 1 – Principe de la restriction pour la sélection des exemples

En se fondant sur les résultats obtenus avec notre référence la plus riche (WM) pour 4 164 des 4 378 mots cibles impliqués ci-dessus, le taux d'erreur de la sélection des contre-exemples est, avec la méthode précédente, de 26,8% tandis que celui de la sélection des exemples est de 58,7%. Ce schéma de sélection simple présente donc l'inconvénient d'impliquer un nombre important d'exemples et de contre-exemples avec des taux d'erreur importants. Nous avons donc également testé une méthode de sélection plus restrictive au sein des mots cibles de fréquence supérieure à 1000, méthode illustrée par la figure 1. Les voisins sémantiques pour les mots de fréquence supérieure à 1000 se trouvant eux aussi majoritairement dans cette même tranche fréquentielle, nous faisons l'hypothèse que si un mot cible A ayant comme voisin de rang 1 un mot B, ce voisin a d'autant plus de chances d'être un mot sémantiquement lié à A que A est lui-même le voisin de rang 1 de B en tant que mot cible. En pratique, ces cas de symétrie entre mot cible et voisin de rang 1 se produisent pour 1052 mots cibles, ce qui permet de construire 526 exemples puisque de ce point de vue, les couples {A,B} et {B,A} sont équivalents. En revanche, nous avons retenu les 1052 contre-exemples correspondant à tous les mots cibles concernés suivant le principe (mot cible, voisin de rang 10). Notre hypothèse se trouve par ailleurs confirmée : parmi les 526 exemples ainsi sélectionnés, le taux d'erreur n'est en effet plus que de 43% par rapport à la référence WM.

4 Expérimentations et évaluation

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 *et al.*, 2009), nous avons adopté un

noyau RBF et une stratégie de type *grid search* pour l'optimisation du paramètre γ 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 chacun des deux ensembles d'apprentissage décrits à la section précédente en se fondant sur la mesure de précision calculée dans le cadre d'une validation croisée divisant ces ensembles en 5 parties. Chacun des deux modèles ainsi construits en utilisant l'outil LIBSVM a ensuite été appliqué à la totalité des 14 670 noms cibles de notre évaluation initiale. Plus précisément, pour chaque nom cible, une représentation d'exemple a été construite pour chaque couple {nom cible, voisin} 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. Le tableau 2 donne les résultats de ce réordonnancement pour le

fréq.	**réf.**	**R-préc.**		**MAP**		**P@1**		**P@5**		**P@10**	
$f > 10$	W	-0,8	(-10%)	-0,8	(-8%)	-0,9	(-8%)	-0,4	(-8%)	-0,3	(-9%)
	M	0,4	(6%)	0,2	(6%)	3,4	(14%)	1,1	(7%)	0,7	(5%)
	WM	0,1	(1%)	0,0	(0%)	2,1	(9%)	0,6	(4%)	0,5	(5%)
$f > 1000$	W	-2,1	(-19%)	-2,1	(-17%)	-2,4	(-14%)	-1,3	(-17%)	-0,8	(-16%)
	M	-0,5	(-5%)	-0,4	(-8%)	-1,0	(-2%) ‡	-2,2	(-8%)	-1,6	(-7%)
	WM	-0,9	(-8%)	-0,8	(-12%)	-2,1	(-5%)	-2,4	(-9%)	-1,7	(-8%)
$100 < f \leq 1000$	W	-1,7	(-16%)	-1,8	(-14%)	-2,1	(-15%)	-0,7	(-12%)	-0,3	(-8%)
	M	0,8	(12%)	0,5	(16%)	7,2	(39%)	3,3	(25%)	2,3	(22%)
	WM	-0,2	(-2%)	-0,4	(-5%)	3,8	(18%)	2,1	(17%)	1,6	(17%)
$f \leq 100$	W	1,9	(90%)	2,0	(61%)	2,5	(96%)	1,0	(83%)	0,5	(56%)
	M	1,0	(83%)	0,6	(120%)	5,6	(224%)	3,4	(227%)	2,3	(153%)
	WM	1,6	(76%)	1,5	(62%)	4,6	(139%)	2,5	(147%)	1,6	(107%)

TAB. 2 – Impact du réordonnancement des voisins avec le modèle à 526 exemples

modèle fondé sur 526 exemples et le tableau 3 pour celui fondé sur 4 378 exemples. Compte tenu du nombre de mesures, nous avons choisi de ne faire apparaître que les différences par rapport aux résultats du tableau 1, à la fois en termes de valeur et de pourcentage (outre le signe indiquant le sens de la différence, la couleur verte indique une variation positive et la couleur rouge, une variation négative). L'évaluation portant sur un processus de réordonnancement des voisins, les valeurs de rappel et de précision au rang maximum restent identiques par rapport à celles du tableau 1 et ne sont pas rappelées.

fréq.	**réf.**	**R-préc.**		**MAP**		**P@1**		**P@5**		**P@10**	
> 10	W	-0,4	(-5%) ‡	-0,5	(-5%)	-0,5	(-4%) ‡	-0,1	(-2%) ‡	-0,1	(-3%) ‡
	M	0,4	(6%)	0,2	(6%)	3,4	(14%)	1,5	(9%)	0,7	(5%)
	WM	0,1	(1%)	0,0	(0%)	2,3	(10%)	1,0	(7%)	0,6	(6%)
> 1000	W	-1,2	(-11%)	-1,1	(-9%)	-1,4	(-8%)	-0,3	(-4%) ‡	-0,2	(-4%)
	M	-0,2	(-2%)	-0,2	(-4%)	-0,2	(-0%) ‡	-0,6	(-2%) ‡	-0,8	(-4%)
	WM	-0,6	(-5%)	-0,5	(-8%)	-0,8	(-2%) ‡	-0,7	(-3%)	-0,9	(-4%)
$100 < f \leq 1000$	W	-1,5	(-14%)	-1,7	(-14%)	-1,6	(-12%)	-0,7	(-12%)	-0,4	(-11%)
	M	0,6	(9%)	0,4	(13%)	6,8	(36%)	2,7	(21%)	1,8	(17%)
	WM	-0,3	(-3%)	-0,4	(-5%)	3,6	(17%)	1,7	(14%)	1,2	(13%)
$f \leq 100$	W	1,7	(81%)	1,6	(48%)	2,1	(81%)	0,8	(67%)	0,5	(56%)
	M	0,9	(75%)	0,5	(100%)	5,0	(200%)	3,1	(207%)	2,0	(133%)
	WM	1,4	(67%)	1,1	(46%)	4,0	(121%)	2,2	(129%)	1,3	(87%)

TAB. 3 – Impact du réordonnancement des voisins avec le modèle à 4 378 exemples

Très clairement, la tendance générale pour les deux modèles considérés est la même : le processus de réordonnancement proposé induit une augmentation significative de toutes les mesures au niveau global avec M et WM comme références[1]. En revanche, une diminution de ces mêmes mesures est observée avec W comme référence.

[1]La significativité statistique des résultats a été évaluée grâce à un test de Wilcoxon avec un seuil de 0,01, les échantillons étant appariés. Seuls les résultats suivis du signe ‡ sont considérés comme non significatifs, ce qui ne concerne que des dégradations de performance.

Autrement dit, par rapport à la mesure de similarité initiale, ce réordonnancement favorise les mots sémantiquement liés mais le fait partiellement au détriment des synonymes. Cette tendance n'est pas surprenante de par le mécanisme d'amorçage utilisé : les premiers sont en effet largement 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.

L'analyse plus fine de ces résultats selon le domaine fréquentiel des noms considérés met en évidence une deuxième grande tendance : l'amélioration des résultats produite par le réordonnancement est d'autant plus sensible que la fréquence du nom est faible. Ainsi, pour les noms de fréquence inférieure à 100, cette amélioration s'observe quelle que soit la référence prise ; pour les noms de fréquence comprise entre 100 et 1000, elle s'identifie à la tendance générale tandis que pour les noms de fréquence supérieure à 1000, la variation correspond à une dégradation par rapport à toutes les références. D'une certaine façon, on peut donc dire que ce processus de réordonnancement rééquilibre la mesure de similarité initiale, très fortement biaisée vers les fortes fréquences. La comparaison des tableaux 2 et 3 ne fait quant à elle apparaître que des différences assez faibles entre les deux modèles SVM. L'utilisation des 4 378 exemples permet d'obtenir des résultats globaux un peu meilleurs mais cette supériorité n'est véritablement notable qu'avec W comme référence. Par ailleurs, elle s'inverse pour les mots de fréquence inférieure à 100 par rapport à toutes les références et pour la tranche fréquentielle intermédiaire, par rapport à M et à WM. On peut à cet égard observer un certain parallélisme en termes de tendances entre la comparaison des deux modèles SVM et la comparaison de la mesure initiale et de ces modèles SVM. En final, le choix parmi ces deux modèles peut aussi être motivé par le fait que le modèle fondé sur 526 exemples est beaucoup plus petit (nombre inférieur de vecteurs de support) et donc, plus efficace que son alter ego.

5 Conclusion et perspectives

Dans cet article, nous avons présenté une méthode d'amélioration d'une mesure de similarité sémantique de nature distributionnelle exploitant l'amorçage. Plus précisément, cette méthode est fondée sur le réordonnancement des voisins sémantiques obtenus par la mesure initiale grâce à un classifieur de type SVM. Ce classifieur est entraîné sur la base d'exemples et de contre-exemples sélectionnés de façon non supervisée à partir des résultats de la mesure de similarité initiale. Cette méthode a montré plus particulièrement son intérêt pour les mots de faible fréquence et pour une similarité sémantique dépassant la stricte synonymie. Dans la perspective de minimiser la taille des modèles construits, déjà explorée ici au travers du nombre d'exemples, nous envisageons de prolonger ce travail en y intégrant la problématique de la sélection de caractéristiques.

Références

BRODA B., PIASECKI M. & SZPAKOWICZ S. (2009). Rank-Based Transformation in Measuring Semantic Relatedness. In *22^{nd} 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.

FERRET O. (2010). Similarité sémantique et extraction de synonymes à partir de corpus. In *$17^{ème}$ Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2010)*.

HAGIWARA M., OGAWA Y. & TOYAMA K. (2009). Supervised synonym acquisition using distributional features and syntactic patterns. *Information and Media Technologies*, **4**(2), 59–83.

KAZAMA J., DE SAEGER S., KURODA K., MURATA M. & TORISAWA K. (2010). A bayesian method for robust estimation of distributional similarities. In *48^{th} Annual Meeting of the Association for Computational Linguistics*, p. 247–256.

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

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

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