Réordonnancer des thésaurus distributionnels en combinant différents critères

Olivier Ferret

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

RÉSUMÉ. *Dans cet article, nous proposons une méthode pour améliorer les thésaurus distributionnels grâce à un mécanisme d'amorçage : un ensemble d'exemples positifs et négatifs 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 la sélection des 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, soit de façon précoce, soit de façon tardive, à un critère déjà expérimenté touchant à la symétrie des relations sémantiques. Nous évaluons l'intérêt de ces propositions sur un large ensemble de noms en anglais couvrant un vaste spectre de fréquences. Cet article est une version étendue de (Ferret, 2013 ; Ferret, 2015a).*

ABSTRACT. *In this article, we propose a method for improving distributional thesauri 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, either by early fusion or late fusion, with an already tested criterion based on the symmetry of semantic relations. We evaluate the interest of the proposed procedure for a large set of English nouns with various frequencies. This article is an extended version of (Ferret, 2013 ; Ferret, 2015a).*

MOTS-CLÉS : *sémantique lexicale, similarité sémantique, thésaurus distributionnel.*

KEYWORDS: *lexical semantics semantic similarity distributional thesaurus.*

TAL. Volume 56 – n° 2/2015, pages 21 à 49

1. Introduction

Les ressources de nature distributionnelle sont utilisées dans un ensemble de tâches de plus en plus important, allant de l'analyse syntaxique (Henestroza Anguiano et Candito, 2012) à l'extraction de relations (Min *et al.*, 2012). Le travail sur lequel se focalise cet article concerne plus spécifiquement les thésaurus distributionnels, qui associent à un mot un ensemble de voisins dits sémantiques, généralement ordonnés selon l'ordre décroissant de leur similarité avec ce mot, à l'image des exemples donnés par le tableau 1. À la suite de Grefenstette (1994), la façon la plus répandue de construire de tels thésaurus à partir d'un corpus est de caractériser chaque mot du corpus par l'ensemble de ses contextes d'occurrence et d'évaluer le niveau de similarité de deux mots en fonction d'une mesure de similarité reposant sur les contextes qu'ils partagent. Cette mesure permet alors de sélectionner les plus proches voisins d'un mot. Ce schéma général se retrouve sous diverses variantes dans des travaux comme (Lin, 1998), (Curran et Moens, 2002), (Weeds, 2003) ou (Heylen *et al.*, 2008).

Au-delà du problème spécifique de la construction de thésaurus, cette façon d'aborder le problème de la similarité sémantique des mots est caractéristique de la mise en œuvre traditionnelle de l'approche distributionnelle. Cette mise en œuvre a fait depuis quelque temps l'objet de nombreux développements. Une partie d'entre eux se sont attachés à améliorer l'approche de Grefenstette (1994), mais sans la changer en profondeur. Ces travaux se focalisent principalement sur la pondération des éléments constituant les contextes distributionnels, à l'instar de Broda *et al.* (2009), qui transforment les poids au sein des contextes en rangs, ou de Zhitomirsky-Geffet et Dagan (2009), repris et étendus par Yamamoto et Asakura (2010), qui proposent une méthode fondée sur l'amorçage pour modifier les poids des éléments des contextes. Kazama *et al.* (2010) ont pour leur part adopté un point de vue bayésien pour aborder la question. D'autres travaux ont envisagé des changements plus radicaux. Les modèles à base d'exemples (Erk et Padó, 2010) ou de prototypes multiples (Reisinger et Mooney, 2010), dans lesquels la représentation d'un mot est fondée sur un ensemble d'exemples caractéristiques au lieu d'une agrégation de contextes d'occurrence, en sont une manifestation. Les méthodes s'appuyant sur la construction de représentations lexicales distribuées en sont une autre, que ce soit par le biais de techniques de factorisation de matrice comme l'analyse sémantique latente (Landauer et Dumais, 1997) ou la factorisation de matrice non négative (Van de Cruys, 2010), de modèles probabilistes fondés sur la mise en évidence de facteurs latents, prenant la forme de sens en utilisant l'allocation de Dirichlet latente (Dinu et Lapata, 2010) ou de classes sémantiques avec un modèle de Markov caché (Grave *et al.*, 2014), de méthodes fondées sur la notion de hachage comme le Random Indexing (Kanerva *et al.*, 2000) ou plus récemment de celles issues du Deep Learning pour la construction de représentations de type *word embedding* (Huang *et al.*, 2012 ; Mikolov *et al.*, 2013) ou du modèle GloVe de Pennington *et al.* (2014).

En dehors des avancées réalisées globalement dans le champ de la sémantique distributionnelle, certains travaux se concentrent sur des voies d'amélioration plus spécifiques aux thésaurus distributionnels. Même s'ils ne traitent pas explicitement de

cette notion de thésaurus, Zhitomirsky-Geffet et Dagan (2009) et Yamamoto et Asakura (2010), déjà mentionnés ci-dessus, relèvent de cette problématique. Ils s'appuient en effet sur un mécanisme d'amorçage dont la première étape consiste à trouver des voisins sémantiques selon une approche comparable à Grefenstette (1994), le résultat ne constituant rien d'autre qu'un thésaurus distributionnel. Ces voisins sont utilisés dans un second temps pour repondérer les éléments constitutifs des contextes distributionnels et aboutir ainsi à une version améliorée du thésaurus initial. Une telle forme d'amorçage se retrouve également au niveau de Ferret (2012). Dans ce cas, le thésaurus initial est à la base de la sélection non supervisée d'exemples positifs et négatifs de mots sémantiquement liés, exemples servant ensuite à entraîner un classifieur permettant de réordonnancer le thésaurus initial. Dans le cas de Ferret (2012), cette sélection s'appuie sur l'exploitation des relations de symétrie existant au niveau du thésaurus initial. Claveau *et al.* (2014) proposent, quant à eux, plusieurs façons de généraliser cette idée d'exploitation des relations à l'échelle du thésaurus pour améliorer celui-ci.

Dans cet article, après avoir défini la méthode de construction et d'évaluation des thésaurus distributionnels que nous utilisons et analysé en détail les caractéristiques de ces thésaurus, nous examinerons comment les principes développés par Ferret (2012) peuvent être conjugués à la sélection non supervisée d'exemples de mots sémantiquement liés fondée sur les mots composés (Ferret, 2013 ; Ferret, 2015a) pour améliorer les thésaurus distributionnels. Nous présenterons, en outre, deux modes de conjugaison des informations apportées par les deux critères de sélection des exemples, l'un correspondant à une fusion dite précoce, réalisée au niveau des ensembles d'apprentissage, l'autre à une fusion dite tardive, opérant au niveau des thésaurus réordonnancés selon chacun des critères.

2. Construire et évaluer un thésaurus distributionnel

2.1. *Paramètres distributionnels*

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. Une telle construction dépend d'un ensemble suffisamment large de paramètres pour qu'il soit en pratique impossible de mener une optimisation globale pour fixer la valeur de ceux-ci. Même les travaux les plus récents menés pour explorer cet espace de paramètres (Bullinaria et Levy, 2012 ; Kiela et Clark, 2014 ; Lapesa et Evert, 2014) font un certain nombre d'impasses et surtout, utilisent pour faire leurs évaluations des jeux de test souvent de petite taille constitués majoritairement de mots ayant des fréquences assez élevées. Des jeux de test tels que WordSim 353 (Gabrilovich et Markovitch, 2007) ou les quatre-vingts questions du TOEFL (Landauer et Dumais, 1997) en sont des exemples typiques. Le fait d'en conjuguer plusieurs de différents types, comme le font les travaux cités, vise à pallier ces insuffisances, mais il n'est pas évident que cumuler des résultats biaisés permette d'aboutir à des conclusions plus solides.

Dans notre cas, nous avons fait le choix de mener des évaluations à large échelle pour éviter certains de ces biais, tout en restant bien sûr prisonnier de l'impossibilité de tester toutes les valeurs de paramètres. Compte tenu de notre objectif final – la construction de thésaurus – nous avons adopté un test de type TOEFL, option la plus proche de cette optique, mais en utilisant le jeu de test WordNet-based Synonymy Test (WBST) proposé par Freitag *et al.* (2005), comportant 9 887 questions pour les noms. Ferret (2010) s'est attaché à la sélection des meilleurs paramètres distributionnels dans ce cadre. Nous reprenons ici les conclusions de ce travail.

Bien que notre langue cible soit l'anglais, nous avons ainsi choisi de limiter le niveau des traitements linguistiques appliqués au corpus source de nos données distributionnelles à l'étiquetage morphosyntaxique 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 assez largement répandue consistant à utiliser un analyseur syntaxique, à l'instar de Curran et Moens (2002). Plus précisément, nous nous sommes appuyé sur l'outil *TreeTagger* (Schmid, 1994) pour assurer le prétraitement du corpus AQUAINT-2 (Voorhees et Graff, 2008) qui est à la base de ce travail. Ce corpus, que l'on peut qualifier de taille moyenne avec ses 380 millions de mots environ, est composé d'articles de journaux.

Les expérimentations menées par Ferret (2010) ont abouti par ailleurs aux choix suivants concernant les paramètres de construction des contextes distributionnels et d'évaluation de leur similarité :

– 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 plein à droite et un mot plein 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 ponctuelle* entre le mot cible et son cooccurrent, restreinte aux valeurs positives (*Positive Pointwise Mutual Information*) ;
– 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 : seuls les mots de fréquence supérieure à 10 sont considérés, limite un peu arbitraire en dessous de laquelle nous considérons la pauvreté des contextes distributionnels comme trop importante pour que leur comparaison soit significative.

Les paramètres que nous avons ainsi sélectionnés, en particulier pour ce qui est de la taille de la fenêtre, de la fonction de pondération des cooccurrents et de la mesure de similarité entre contextes, sont très directement en phase avec ceux faisant consensus parmi les travaux récents réalisés avec des corpus de taille importante et pour des éva-

abnormality	defect [0,30], disorder [0,23], deformity [0,22], mutation [0,21], prolapse [0,21], anomaly [0,21] ...
agreement	accord [0,44], deal [0,41], pact [0,38], treaty [0,36], negotiation [0,35], proposal [0,32], arrangement [0,30] ...
cabdriver	waterworks [0,23], toolmaker [0,22], weaponeer [0,17], valkyry [0,17], wang [0,17], amusement-park [0,17] ...
machination	hollowness [0,15], share-price [0,12], clockmaker [0,12], huguenot [0,12], wrangling [0,12], alternation [0,12] ...

Tableau 1. *Premiers voisins de quelques entrées du thésaurus distributionnel A2ST*

luations portant également sur la similarité sémantique au sens large du terme (Kiela et Clark, 2014 ; Baroni *et al.*, 2014 ; Levy *et al.*, 2015). Le jeu de test WBST que nous avons utilisé est très clairement centré sur la similarité sémantique, par opposition à la proximité sémantique, ce qui explique en particulier le choix d'une fenêtre de petite taille. Il faut néanmoins remarquer que l'idée généralement acceptée d'une association des petites fenêtres à la similarité sémantique et des plus larges fenêtres à la proximité sémantique semble à relativiser dans le cas des grands corpus : les jeux de test utilisés dans bon nombre des travaux évoqués, qui mettent tous en avant les meilleures performances des fenêtres de petite taille, incluent aussi bien des relations de similarité sémantique que de proximité sémantique, les secondes étant même parfois plus nombreuses que les premières. Nous avons d'ailleurs fait le même constat pour les thésaurus distributionnels dans le cadre d'expérimentations non présentées ici.

2.2. *Construction et évaluation du thésaurus*

Disposant d'une mesure permettant d'évaluer la similarité sémantique d'un couple de mots, la procédure de construction d'un thésaurus est simple : les voisins sémantiques d'un mot sont trouvés en recherchant les N plus proches voisins de ce mot selon la mesure de similarité considérée. Plus précisément, cette recherche consiste à appliquer cette mesure de similarité entre le mot cible et tous les autres mots du vocabulaire considéré ayant la même catégorie morphosyntaxique (ici, les noms). Finalement, tous ces mots sont triés suivant leur valeur de similarité et seuls les N plus proches voisins, N étant égal à 100 dans nos expérimentations, sont conservés en tant que voisins sémantiques.

Nous avons appliqué la procédure de construction décrite à l'ensemble des noms du corpus AQUAINT-2 de fréquence strictement supérieure à 10, soit 26 210 noms. Au final, le thésaurus résultat, appelé A2ST, est constitué de 25 988 entrées, la différence s'expliquant par les cas où le contexte distributionnel d'un nom ne comporte aucun élément commun avec celui d'un autre nom. La limite fixée des 100 voisins est atteinte par 99,8 % d'entre elles. Le tableau 1 donne les premiers voisins associés à

type de relation	% des relations	
c	9,0	c : co-hyponymie (= H + h)
c + h	6,5	h : hyponymie
H + c	5,8	H : hyperonymie
H + c + h	3,8	s : synonymie
s	3,4	r1 + r2 : composition des relations r1 et r2
h	3,0	
H	2,8	

Tableau 2. *Relations sémantiques les plus fréquentes au sein du thésaurus Moby caractérisées en fonction de WordNet*

quatre entrées en illustrant le fait que ces voisins peuvent être très pertinents, au regard de la notion de similarité sémantique, pour certaines entrées, comme *abnormality* ou *agreement* ici, et beaucoup moins pour d'autres, comme dans le cas de *cabdriver* ou *machination*.

L'évaluation de l'ensemble du thésaurus demande néanmoins une procédure plus formelle et plus automatique. Comme dans le cas des paramètres distributionnels, cette évaluation est de nature intrinsèque et à large échelle : les voisins du thésaurus sont comparés à des ressources de référence. Nous avons plus spécifiquement adopté deux ressources complémentaires de large couverture : les synonymes de WordNet [W] (Miller, 1990), dans sa version 3.0, et le thésaurus Moby [M] (Ward, 1996). Les premiers sont représentatifs de la similarité sémantique tandis que le second recouvre un spectre plus large de relations sémantiques que l'on peut en première analyse regrouper sous la notion de proximité sémantique. Nous avons également créé une ressource fusionnant WordNet et le thésaurus Moby [WM]. En reprenant Ferret (2015b), le tableau 2 donne un aperçu des relations présentes dans Moby en les caractérisant en fonction de WordNet, soit comme des relations élémentaires de WordNet (synonymie, hyponymie, etc.), soit comme des relations composées d'une suite de ces relations élémentaires. Il montre ainsi que les relations les plus fréquentes sont composées mais également que Moby recèle une grande diversité de types de relations puisque les sept types de relations les plus fréquents ne couvrent que 34,3 % de ses relations.

Notre but étant d'abord d'évaluer le thésaurus construit et non la capacité de celui-ci à reconstituer les ressources de référence, nous avons filtré ces ressources en éliminant en leur sein, aussi bien au niveau des entrées que des mots qui leur sont liés, les termes ne faisant pas partie du vocabulaire des noms simples retenus pour construire nos données distributionnelles. Au final, seules 14 670 entrées du thésaurus, intersection entre les noms de ce dernier et ceux de WordNet[1], ont été utilisées

1. Il est à noter que tous les noms présents dans WordNet ne sont pas nécessairement associés à un synset, ce qui explique la différence entre le nombre d'entrées du thésaurus considéré et le nombre d'entrées évaluées par rapport à WordNet.

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	5 889	3,3	29,4	11,8	13,5	17,4	7,5	4,9	1,0
hautes	M	5 751	60,5	11,2	9,4	4,6	35,9	24,2	18,9	6,8
7 335	WM	6 754	52,6	11,4	11,1	7,4	36,4	22,8	17,5	6,0
	W	4 584	2,3	16,0	3,7	5,1	4,2	2,0	1,4	0,4
basses	M	3 465	32,5	4,4	2,3	0,9	4,4	3,4	3,1	1,4
7 335	WM	5 489	21,6	5,1	3,6	3,4	5,5	3,3	2,7	1,1

Tableau 3. *Évaluation du thésaurus distributionnel A2ST*

pour l'évaluation présentée dans le tableau 3. La troisième colonne de ce même tableau donne le nombre effectif de noms pour lesquels l'évaluation a été réalisée pour chaque ressource, chaque entrée retenue pour l'évaluation n'apparaissant pas dans chaque ressource de référence. La quatrième colonne correspond pour sa part au nombre moyen de voisins de référence (appelés synonymes par facilité de langage) à trouver dans chaque ressource pour chacune de leurs entrées faisant partie du vocabulaire AQUAINT-2.

Les voisins étant ordonnés pour chaque entrée du thésaurus, il est possible de faire le parallèle entre la recherche de voisins sémantiques et la recherche de documents en recherche d'information et de réutiliser ainsi les métriques d'évaluation classiquement utilisées pour cette dernière en faisant jouer aux entrées du thésaurus le rôle de requêtes et aux autres noms celui de documents. Les six dernières colonnes du tableau 3 donnent ainsi les résultats en pourcentage pour les métriques suivantes : 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 pour les 1, 5, 10 et 100 premiers voisins). Ces métriques sont complétées par la donnée à la cinquième colonne du tableau du pourcentage des voisins de référence figurant parmi les 100 voisins sémantiques de chaque entrée de notre thésaurus distributionnel. La fréquence des mots, en relation avec la taille des corpus, 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.

L'analyse du tableau 3 conduit à faire trois grandes observations. Tout d'abord, malgré leurs performances intéressantes sur un test de similarité sémantique à large

couverture et adapté à l'application visée, les paramètres distributionnels sélectionnés n'obtiennent dans l'absolu que des résultats assez modestes lorsqu'ils sont appliqués au problème de la construction d'un thésaurus distributionnel. Cette faiblesse est observable aussi bien au niveau du taux de rappel des voisins de référence – environ 25 % pour WordNet et 10 % pour le thésaurus Moby – qu'au niveau de leur rang parmi les voisins retenus : la R-précision générale dépasse à peine 8 % dans le meilleur des cas, en l'occurrence WordNet. Ce constat a une portée plus générale que notre travail spécifique dans la mesure où ces paramètres peuvent être considérés comme classiques.

La deuxième observation est que cette faiblesse générale recouvre 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 qualité des voisins sémantiques est élevée, à la fois en termes de quantité et de rang. Le phénomène conduit d'ailleurs à considérer que pour les basses fréquences, les résultats obtenus sont difficilement exploitables d'un point de vue applicatif. Ce niveau de résultats est assez aisément compréhensible : un mot cible avec peu d'occurrences ne peut avoir qu'un contexte distributionnel très pauvre. Ainsi, deux mots de faible fréquence risquent fort de n'avoir aucun élément en commun au niveau de leurs contextes et, dans le cas de la comparaison d'un mot de faible fréquence avec un mot de fréquence plus élevée, la taille très réduite de l'intersection de leurs contextes rend cette comparaison très sensible à des cooccurrences non significatives sur le plan sémantique. Même si cette constatation plaide en faveur de l'accroissement de la taille des corpus, ce que l'on observe de fait actuellement, des mots de faible fréquence existeront toujours dans ces corpus en vertu de la loi de Zipf. Par ailleurs, il existe de nombreuses situations où de très larges corpus ne sont pas disponibles.

La dernière observation suscitée par le tableau 3 est 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 (2,9 en moyenne) tandis que le thésaurus Moby contient pour chaque entrée un nombre beaucoup plus important de mots sémantiquement proches (50 en moyenne). Cette différence de profil explique en particulier que si les valeurs pour Moby sont assez élevées par rapport à celles pour WordNet pour des précisions à un rang faible – P@1 = 35,9 pour les hautes fréquences par exemple, à comparer à 17,4 – le rapport s'inverse dans les mêmes conditions pour la MAP : 13,5 pour WordNet et 4,6 pour Moby. Intuitivement, les premiers voisins du thésaurus ont dans le cas de Moby plus de chances de figurer dans un ensemble de voisins de référence plus large et couvrant un plus grand nombre de types de relations sémantiques mais, en contrepartie, il est plus difficile pour le thésaurus de couvrir ce large ensemble de voisins de façon significative.

2.3. *Mise en perspective*

Aborder la problématique de la similarité sémantique fondée sur des bases distributionnelles par le biais des thésaurus n'est pas la façon de faire la plus répandue

jeu de test	réf.	#mots éval	#syn./ mot	rappel	R-préc.	MAP	P@1	P@5	P@10	P@100
	W	59	5,2	23,6	9,3	9,6	20,3	8,1	5,6	1,2
70	M	64	103,2	11,2	11,4	5,1	54,7	43,1	33,3	11,6
(69)	WM	65	103,1	11,2	12,0	5,8	55,4	43,4	33,5	11,6
	W	247	4,7	27,2	12,3	12,7	19,8	8,4	5,8	1,3
300	M	253	97,6	11,1	11,5	5,6	53,0	37,0	29,0	10,8
(296)	WM	269	93,0	11,2	12,2	6,7	52,0	36,0	28,2	10,4

Tableau 4. *Évaluation des voisins du thésaurus A2ST pour le jeu de test de 70 mots de Curran et Moens (2002) et de 300 mots de Curran (2003)*

à l'heure actuelle. En conséquence, les méthodologies d'évaluation dans cette sphère sont moins uniformes que pour l'évaluation de la stricte similarité sémantique, réalisée grâce à des jeux de test de type WordSim 353. Les comparaisons sont donc aussi plus difficiles. Dans le cas du néerlandais par exemple, Van der Plas et Bouma (2004) et Van de Cruys (2010) ont ainsi adopté la version néerlandaise d'EuroWordNet comme référence, assez comparable à WordNet, mais en s'appuyant sur sa structure hiérarchique : au lieu de simplement se fonder sur l'appartenance à une liste de synonymes présents dans EuroWordNet, la pertinence sémantique d'un voisin par rapport à une entrée est définie en calculant la mesure de Wu et Palmer (1994) entre cette entrée et le voisin. Cette méthode présente l'avantage d'intégrer de façon cohérente des types de relations différents dans une même mesure. Elle est cependant moins directement interprétable que l'option que nous avons adoptée. Pantel *et al.* (2009) s'inscrivent, pour leur part, dans un cadre plus applicatif en s'intéressant à la notion d'ensemble d'entités (*Entity Sets*), sous-tendue par une gamme de relations très étendue et se focalisant beaucoup sur des entités nommées.

Le travail de Curran et Moens (2002) est en revanche 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 obtenue (Dice† + T-test) est une précision au rang 1 de 76 %, au rang 5 de 52 % et au rang 10 de 45 % pour 70 noms, à comparer avec 41,3 %, 28,0 % et 21,9 % dans notre cas en se restreignant aux 3 732 noms de fréquence $>$ à 1 000[2].

2. Nous reprenons ici les chiffres de Ferret (2010), qui effectue un découpage plus fin en trois tranches fréquentielles à peu près de même taille, les fréquences $>$ à 1 000 constituant la tranche supérieure et les fréquences $\leq$ 100, la tranche inférieure.

méthode	#mots éval	#syn./ mot	rappel	R-préc.	MAP	P@1	P@5	P@10	P@100
A2ST	12 243	38,7	9,8	7,7	5,6	22,5	14,1	10,8	3,8
A2ST-SYNT	11 887	39,4	13,2	10,7	7,9	29,4	18,9	14,6	5,2
[Lin, 98]	9 823	44,5	12,7	11,6	8,1	36,1	23,7	18,2	5,6
[Huang *et al.*, 12]	10 537	42,6	3,8	1,9	0,8	7,1	5,0	4,0	1,6
[Mikolov *et al.*, 13]	12 326	38,6	6,2	5,5	4,2	16,3	9,5	7,0	2,4
ESA	7 756	44,3	7,0	6,9	5,1	13,2	9,1	7,3	3,1
[Baroni *et al.*, 14]-C	12 052	39,3	13,6	12,5	9,8	31,9	19,6	15,2	5,3
[Baroni *et al.*, 14]-P	12 052	39,3	11,3	10,9	8,5	30,3	18,4	13,8	4,4

Tableau 5. *Évaluation de thésaurus construits selon des méthodes différentes*

Une explication de cette différence pourrait être l'utilisation de cooccurrents syntaxiques par Curran et Moens (2002) alors que nous nous contentons de cooccurrents graphiques. Néanmoins, comme le montre la ligne A2ST-SYNT du tableau 5, la référence étant [WM], l'utilisation de cooccurrents syntaxiques sur le corpus AQUAINT-2 ne suffit pas à expliquer la différence avec Curran et Moens (2002). Même les meilleurs résultats pour ce thésaurus A2ST – P@1 = 44,0 %, P@5 = 31,5 % et P@10 = 25,2 % pour 3 727 noms de fréquence $>$ à 1 000 avec [M] comme référence – sont encore assez éloignés des chiffres de Curran et Moens (2002). Deux autres facteurs sont aussi à considérer. Tout d'abord, le niveau de richesse des références utilisées est très différent. Pour 3 732 noms de fréquence $>$ à 1 000, le thésaurus Moby fournit en moyenne 69 mots sémantiquement liés dans notre cas tandis que pour les 70 noms de Curran et Moens (2002), ce nombre monte à 331. Or, ce facteur a une grande influence sur les résultats ainsi que nous l'avons illustré ci-dessus où le passage d'une moyenne de 2,9 synonymes par entrée pour WordNet à 50 mots liés pour Moby s'accompagne d'une montée de la précision au rang 5 de 5,1 % à 16,4 %. Le même phénomène, observé aussi entre WordNet et Moby, explique que le rappel soit dans notre cas supérieur, avec 11,4 %, à celui de Curran et Moens (2002), égal à 8,3 %.

Le second facteur est néanmoins aussi important. Bien que les 70 noms du jeu de test aient été en principe sélectionnés de manière équilibrée en termes notamment de fréquence, on constate en pratique que sur les 69 présents dans notre thésaurus, 65 figurent parmi les entrées de fréquence $>$ à 1 000 et aucun parmi les mots de fréquence $\leq$ à 100. La première ligne du tableau 4 montre par ailleurs que les performances de notre thésaurus obtenus pour ces 69 entrées sont bien supérieures aux performances de l'ensemble de nos entrées de fréquence $>$ à 1 000. Ce constat s'applique également à un jeu de test plus large constitué de 300 noms et utilisé par Curran (2003). Parmi les 296 noms faisant partie de notre thésaurus, 244 figurent parmi les entrées de fréquence $>$ à 1 000 tandis que 3 seulement ont une fréquence $\leq$ à 100. Encore une fois, les

performances pour ce jeu de test étendu sont plus élevées que celles obtenues pour l'ensemble de nos entrées de fréquence $>$ à 1 000.

Pour achever cette mise en perspective, le tableau 5 compare les résultats de notre thésaurus (A2ST) avec ceux de plusieurs autres thésaurus en utilisant [WM] comme référence et les mêmes entrées que pour l'évaluation d'A2ST. A2ST-SYNT est le thésaurus que nous avons produit dans les mêmes conditions qu'A2ST en nous contentant de remplacer les cooccurrents graphiques par des cooccurrents syntaxiques obtenus grâce à l'analyseur MINIPAR (Lin, 1994). Comme nous l'avons indiqué précédemment, et en cohérence avec Curran et Moens (2002) et Heylen *et al.* (2008), cette substitution a un effet très clairement positif sur les résultats et ce, pour toute la gamme des fréquences. La seule restriction à noter est un petit rétrécissement du nombre des entrées pour lesquelles des voisins sont trouvés. [Lin, 98] est le thésaurus mis à disposition par Lin[3], construit comme A2ST-SYNT grâce à des cooccurrents syntaxiques obtenus par l'analyseur MINIPAR. L'évaluation de ce thésaurus donne de meilleurs résultats que pour A2ST-SYNT, ce qui peut s'expliquer par deux facteurs : d'une part, le corpus utilisé par Lin, d'une taille de 1,5 milliard de mots, est beaucoup plus important que le corpus AQUAINT-2, d'autre part, du fait des entrées disponibles, l'évaluation du corpus de Lin a été réalisée sur un plus petit ensemble d'entrées (seulement 1 510 pour les fréquences $\leq$ à 100 à comparer à 3 687 pour A2ST), en moyenne de plus forte fréquence comme le montre le nombre plus élevé de synonymes par entrée.

Les lignes restantes du tableau 5 correspondent à des thésaurus que nous avons construits en suivant le même processus que pour A2ST mais en utilisant des représentations différentes, toujours exprimées sous la forme vectorielle, en lieu et place des contextes distributionnels classiques (l'exception étant [Baroni *et al.*, 14]-C, fondé sur des contextes distributionnels classiques mais non construits par nos soins). Dans chacun des cas, nous appliquons la même mesure, en l'occurrence la mesure *cosinus*, aux représentations associées aux mots afin d'évaluer la similarité de ceux-ci. Dans ce cadre, [Huang *et al.*, 12] et [Mikolov *et al.*, 13] renvoient à deux approches récentes évoquées en introduction et fondées sur la construction de représentations distribuées de mots par des réseaux de neurones. Dans le cas de [Huang *et al.*, 12], nous avons utilisé les représentations construites à partir de Wikipédia[4] fournies par les auteurs tandis que dans le cas de [Mikolov *et al.*, 13], nous avons calculé ces représentations à partir du corpus AQUAINT-2 grâce au logiciel *word2vec*[5] en utilisant les meilleurs paramètres sélectionnés par Mikolov *et al.* (2013)[6]. Dans les deux cas, les résultats sont significativement inférieurs à ceux d'A2ST, avec un niveau particulièrement bas pour [Huang *et al.*, 12] qui peut s'expliquer au moins en partie par la différence de corpus. Ces résultats suggèrent néanmoins que l'utilisation de ce type de représentations distribuées n'est pas encore une option intéressante pour la construction de thésaurus distributionnels, un peu à contresens de Baroni *et al.* (2014) mais plus compatible avec

3. `http://webdocs.cs.ualberta.ca/~lindek/Downloads/sim.tgz`
4. `http://nlp.stanford.edu/~socherr/ACL2012_wordVectorsTextFile.zip`
5. `http://code.google.com/p/word2vec`
6. word2vec -cbow 0 -size 600 -window 10 -negative 0 -hs 0 -sample 1e-5

Hill *et al.* (2014). Ce constat est renforcé par les deux dernières lignes du tableau 5, qui donnent les résultats des thésaurus construits avec les vecteurs de contexte mis à disposition par Baroni *et al.* (2014)[7] : [Baroni *et al.*, 14]-P correspond à des vecteurs construits grâce au modèle CBOW de Mikolov *et al.* (2013) tandis que [Baroni *et al.*, 14]-C correspond à des vecteurs de cooccurrents obtenus de façon classique par une fenêtre graphique. Le niveau des résultats obtenus, rivalisant et même dépassant pour bon nombre de mesures les résultats d'A2ST-SYNT et ceux du thésaurus de Lin, confirme l'observation faite à propos du thésaurus de Lin de la grande importance de la taille du corpus initial sur les résultats, égale à 2,8 milliards de mots dans le cas de Baroni *et al.* (2014). Mais l'observation la plus importante est ici la supériorité de [Baroni *et al.*, 14]-C par rapport à [Baroni *et al.*, 14]-P, ce qui vient limiter le constat général fait par Baroni *et al.* (2014) de la supériorité des modèles neuronaux par rapport aux approches traditionnelles. Ce constat n'est visiblement pas vérifié dans le cas des thésaurus distributionnels.

Enfin, nous donnons également l'évaluation d'un thésaurus fondé sur l'approche ESA proposée par Gabrilovich et Markovitch (2007). Dans ce cas, les traits sur lesquels s'appuie le calcul de la similarité entre deux mots sont constitués de concepts Wikipédia, chaque concept correspondant en pratique à un article de cette encyclopédie. Pour construire notre thésaurus, nous avons exploité les données constituées par Popescu et Grefenstette (2011)[8]. Bien que l'ensemble des entrées soit ici plus limité que pour les autres thésaurus, il est suffisamment important pour se rendre compte qu'il existe une certaine variabilité de la performance de l'approche ESA, qui est plutôt bonne sur le test WordSim 353 et moins intéressante pour la construction d'un thésaurus, se situant assez proche de celle de Mikolov *et al.* (2013).

3. Utiliser l'amorçage pour améliorer un thésaurus distributionnel

Les analyses faites ci-dessus ont montré que les performances de l'approche distributionnelle restent globalement modestes quant à la qualité des thésaurus qu'elle produit et que ce constat n'est pas propre au cadre que nous avons adopté, même si certains facteurs, comme l'utilisation de cooccurrents syntaxiques ou l'augmentation de la taille des corpus, permettent d'améliorer en partie la situation. Il est, dès lors, légitime de chercher à améliorer ces thésaurus. Nous nous sommes plus spécifiquement concentré sur des méthodes non supervisées ou très faiblement supervisées. Dans cette section, nous proposons un cadre d'amélioration se fondant sur l'amorçage.

3.1. *Principes*

L'évaluation de notre thésaurus distributionnel initial, A2ST, montre que les voisins sémantiques obtenus sont significativement meilleurs pour certaines entrées que

7. `http://clic.cimec.unitn.it/composes/semantic-vectors.html`
8. Nous remercions Adrian Popescu pour avoir mis à notre disposition ces données.

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) ont 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 par 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, exploitant les résultats de 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 2.1.

À la différence de Hagiwara (2008), nous ne faisons volontairement pas l'hypothèse de l'accès possible à un ensemble d'exemples et de contre-exemples étiquetés manuellement pour réaliser l'entraînement d'un tel classifieur. Le nombre de ces exemples dans (Hagiwara, 2008), 2 148 pour les positifs et 13 855 pour les négatifs, est en effet très important. En revanche, les voisins sémantiques de notre thésaurus initial peuvent être exploités pour construire un tel ensemble. La mesure de similarité à laquelle est adossé ce thésaurus n'offre pas de critère évident pour discriminer les mots sémantiquement liés[9]. 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, que l'on peut rapprocher dans une certaine mesure de la notion d'auto-apprentissage (*self-training*) :

– construction d'un thésaurus distributionnel ;
– sélection non supervisée d'un ensemble d'exemples et de contre-exemples de mots sémantiquement similaires au sein de ce thésaurus au moyen d'heuristiques ;
– entraînement d'un classifieur statistique à partir des exemples sélectionnés ;

9. 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.

– 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 3.3, 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 par 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.

3.3. *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 3 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 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 4, nous avons ainsi construit 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 à 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 3 impose un double constat : trouver un voisin sémantiquement proche est d'autant plus probable que la fréquence de l'entrée

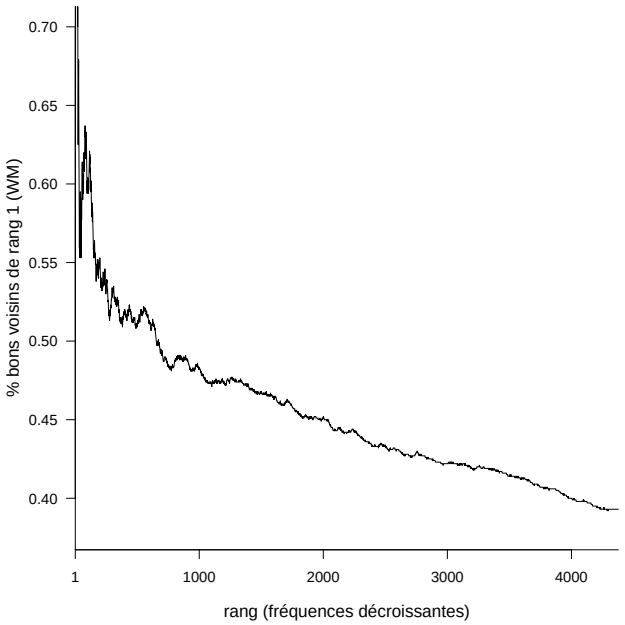

Figure 1. *Proportion de bon voisins de rang 1 selon les fréquentes décroissantes de leur entrée*

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). Comme le montre la figure 1, qui donne la proportion de bons voisins au rang 1 selon les fréquences décroissantes des entrées de forte fréquence, il n'existe pas de critère évident permettant de fixer un seuil. À titre indicatif, prendre les 2 148 premières entrées en termes de fréquence (nombre d'exemples positifs de Hagiwara (2008)) conduit à un taux d'erreur encore égal à 55,7 % et ce taux ne descend en dessous de 50 % qu'avec un nombre d'entrées égal à 654.

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.

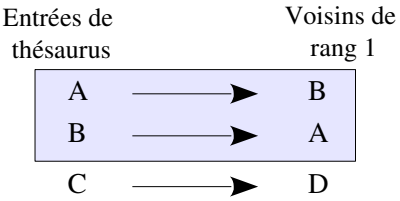

Figure 2. *Sélection d'exemples reposant sur la symétrie des relations sémantiques*

3.4. *Sélection des exemples fondée sur les relations de symétrie dans le thésaurus*

Notre première méthode de sélection d'exemples de mots sémantiquement similaires a été introduite par 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, situation qu'illustre la figure 2. Plus précisément, les résultats du tableau 3 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 entrées et voisins de rang 1 pour 1 592 entrées. Un ensemble de 796 exemples de mots sémantiquement similaires ont finalement été produits puisque les couples (A, B) et (B, A) représentent un même exemple.

3.5. *Sélection des exemples fondée sur les mots 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 composés le même rôle syntaxique sont eux-mêmes susceptibles d'être sémantiquement similaires. Par exemple, le fait que les noms composés *movie_director* et *film_director* soient considérés comme similaires et que les têtes syntaxiques de ces deux mots composés soient identiques conduit à valider la similarité sémantique observée entre *film* et *movie* dans le thésaurus initial.

3.5.1. *Construction d'un thésaurus distributionnel de noms composés*

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 morphosyntaxiques ; 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 morphosyntaxique, l'étiqueteur *TreeTagger* dans notre cas, en s'appuyant sur un ensemble de patrons morphosyntaxiques. Nous nous sommes limité aux trois patrons de noms composés suivants[10] :

Patrons	**Exemples**
NN NN	village chief, league team, cruise ship, oil producer, movie director
JJ NN	medical information, commercial right, educational program
NN IN NN	sense of duty, director of photography, fall in oil[11]

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

Tableau 6. *Évaluation du thésaurus distributionnel de noms composés*

Le tableau 6 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 en-

10. Avec NN : nom (y compris NNS pour les formes plurielles), JJ : adjectif et IN : préposition.
11. Comme on peut le constater avec ce dernier exemple, l'extraction des termes n'est pas parfaite. Même si *fall in oil* correspond à un syntagme nominal correct, il est probable que le terme à extraire soit dans ce cas plus long, comme dans *fall in oil prices*.

trées, à comparer à 83,5 % des entrées pour les mono-termes[12]. 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 beaucoup d'évaluations standard : 70 pour Curran et Moens (2002) ou 353 pour Gabrilovich et Markovitch (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 tend à être confirmé par l'évaluation de plus large ampleur de Ferret (2014).

3.5.2. *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 E_i, son expansion, au sens des grammaires de dépendance. Conformément au principe sous-tendant notre méthode de 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[13]. Comme le montre le tableau 6, 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 après $\rightarrow$) en appliquant les trois règles suivantes,

12. Il faut néanmoins noter que dans le cas des mono-termes, les entrées considérées pour l'évaluation sont des noms présents dans WordNet. Si l'on considère toutes les entrées du thésaurus, la couverture n'est plus que de 47 %, ce qui reste toutefois très supérieur à 2,8 %.

13. La similarité des expansions ne nous intéresse pas ici lorsque ce sont des adjectifs.

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*

(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*

4. Expérimentations et évaluation

4.1. *Sélection des exemples de mots sémantiquement similaires*

Le tableau 7 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 règles de sélection. Les chiffres donnés entre crochets représentent, quant à eux, les pourcentages d'erreurs parmi les contre-exemples. 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 particulière mais répondent plutôt à 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.

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 ensuite été utilisée. 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 égales, que n'ont tendance à être similaires des expansions de deux noms composés dont les têtes sont égales. Ce résultat peut se comprendre si l'on fait l'hypothèse, *a priori* fondée, que la tête d'un composé est plus représentative du sens de ce composé que son expansion. Plus globalement, le tableau 7 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 –

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

Tableau 7. *Évaluation de la qualité des exemples sélectionnés par rapport aux ressources de référence (% bons exemples [% mauvais contre-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.

4.2. *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 recherche en grille (*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 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 cinq parties. Chaque modèle SVM correspondant a été construit en utilisant l'outil LIBSVM (Chang et Lin, 2001) 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.

4.3. *Évaluation*

Le tableau 8 donne les résultats globaux du réordonnancement réalisé sur la base des exemples sélectionnés par chacune des deux méthodes présentées (*symétrie* et *composés*) et leur combinaison (*sym.+comp.*) tandis que la figure 3 en donne une vision plus détaillée en fonction des tranches fréquentielles pour la seule combinaison *sym.+comp.* Cette dernière correspond à ce que nous avons appelé dans l'introduction *fusion précoce*. Chacun de 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 3. 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 $(+0,4)$
sym.+comp.	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 $(+0,9)$

Tableau 8. *Réordonnancement des voisins sémantiques de toutes les entrées du thésaurus initial pour chaque méthode de sélection d'exemples et leur combinaison*[14]

La tendance générale est claire : le processus de réordonnancement conduit à une amélioration significative des résultats globaux pour toutes les méthodes dans le cas des références M et WM, la seule exception étant une très légère diminution de la MAP pour la référence WM dans le cas de la méthode dite *composés*. Parallèlement, une diminution des résultats est observée pour la référence W, jugée statistiquement non significative. 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 8. Cependant, les résultats du tableau 8 montrent clairement l'intérêt de l'association des deux méthodes de sélection : d'un côté, la sélection fondée sur la symétrie des relations vient rééquilibrer ce biais au bénéfice des résultats globaux, de l'autre côté, les exemples apportés par la méthode fondée sur les mots composés élargissent l'ensemble d'apprentissage de celle fondée sur la symétrie dont la plus grande qualité des résultats ne suffit pas totalement à compenser la taille restreinte.

L'analyse des résultats donnés par la figure 3 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.

14. 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.

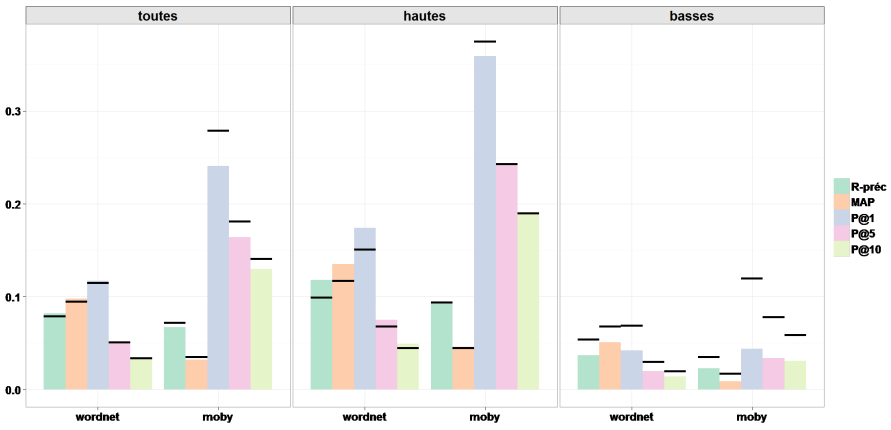

Figure 3. *Résultats du réordonnancement du thésaurus initial suivant les tranches fréquentielles. L'histogramme représente les valeurs pour le thésaurus initial et les barres* ***noires****, les valeurs pour le réordonnancement conjuguant l'utilisation de la symétrie des relations et celle des mots composés.*

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. L'évaluation de ces trois thésaurus confirme, par ailleurs, les résultats du tableau 7 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.

Enfin, le tableau 9 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 présents 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 troisième 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.

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 ...

Tableau 9. *Impact du réordonnancement pour l'entrée* esteem

4.4. *Fusion des thésaurus*

Les performances de l'approche *sym.+comp.* présentées à la section précédente illustrent l'intérêt de combiner des approches reposant sur des critères différents. Les travaux sur la fusion de données font généralement la distinction entre fusion précoce et fusion tardive (Atrey *et al.*, 2010), la première opérant au niveau des représentations, la seconde au niveau des résultats. L'approche précédente *sym.+comp.* peut être considérée comme relevant d'une fusion précoce dans la mesure où la fusion, ici des ensembles d'apprentissage, s'opère en amont du processus de réordonnancement. À l'instar de Curran (2002), nous avons également testé une fusion tardive des deux méthodes de réordonnancement de thésaurus. Chaque thésaurus résultat donnant pour chacune de ses entrées une liste de voisins ordonnés selon l'ordre décroissant de leur proximité avec leur entrée, la solution la plus évidente est de procéder pour chaque entrée à une fusion de la liste des voisins issue de chacun des thésaurus résultats en adoptant une méthode classique de vote. Le tableau 10 donne les résultats que nous avons obtenus avec quatre de ces méthodes. Trois d'entre elles, *Borda*[15], *Condorcet*[16] (Nuray et Can, 2006) et *Reciprocal Rank Fusion* (RRF, avec le paramètre $k = 60$ de (Cormack *et al.*, 2009))[17], s'appuient uniquement sur les rangs tandis que *CombSum*,

15. La méthode *Borda* attribue à chaque voisin de chaque liste à fusionner un poids égal à : taille de la liste − rang du voisin − 1. Le poids final de chaque voisin au sein de l'union des listes est donné par la somme de ses poids dans chaque liste.

16. La méthode *Condorcet* fusionne les listes de voisins en s'assurant que dans le résultat de la fusion, pour un voisin A de rang i et un voisin B de rang j avec $i < j$, le rang de A dans chacune des listes fusionnées est inférieur au rang de B pour une majorité de ces listes.

17. RRF classe les voisins selon l'ordre croissant du score $\sum_l \frac{1}{k+rang(v,l)}$ où $rang(v, l)$ est le rang du voisin v dans la liste l.

Thésaurus	**R-préc.**	**MAP**	**P@1**	**P@5**	**P@10**
initial	7,7	5,6	22,5	14,1	10,8
symétrie	+ 0,3	+ 0,1	+ 2,1	+ 0,8	+ 0,6
composés	+ 0,1	+ 0,0	+ 2,0	+ 0,9	+ 0,6
sym.+ comp.	+ 0,3	+ 0,2	+ 2,8	+ 1,2	+ 0,9
RRF	+ 0,7	+ 0,6	+ 3,7	+ 1,9	+ 1,4
Borda	+ 0,7	+ 0,5	+ 3,6	+ 1,7	+ 1,3
Condorcet	+ 0,5	+ 0,4	+ 3,4	+ 1,6	+ 1,2
CombSum	**+ 0,9**	**+ 0,8**	**+ 4,7**	**+ 2,2**	**+ 1,5**

Tableau 10. *Comparaison des différentes méthodes de fusion avec la référence [WM]*

utilisée ici avec une normalisation des valeurs selon Lee (1997)[18], exploite les valeurs de similarité. Trois thésaurus sont ainsi fusionnés : le thésaurus initial, le thésaurus réordonné grâce au critère de symétrie et celui réordonné grâce aux mots composés.

Outre les résultats pour ces quatre méthodes de fusion, donnés pour la seule référence [WM] pour des raisons de place, le tableau 10 rappelle les résultats pour les thésaurus fusionnés. Ces résultats, de même que ceux issus des fusions, sont donnés en différence de valeur par rapport au thésaurus initial. Un premier constat d'évidence s'impose : les méthodes de fusion permettent toutes de dépasser les résultats de chacun des quatre thésaurus fusionnés. Les gains en termes de R-précision et de MAP apparaissent modestes, mais la référence étant [WM], le nombre de voisins de référence est important, ce qui a un impact direct sur ces deux mesures. En revanche, les gains sont nettement plus substantiels concernant la précision aux rangs 1, 5 et 10. Dans une optique applicative, cette tendance est la plus importante : seuls les voisins des tout premiers rangs sont en effet utilisés dans un tel contexte. Parmi l'ensemble des méthodes de fusion, *CombSum* se détache clairement pour toutes les mesures, l'effet étant particulièrement notable pour la précision au rang 1. L'utilisation des valeurs de similarité, dont la normalisation est indispensable dans le cas présent, s'avère donc supérieure à celle des rangs. Parmi les méthodes exploitant les rangs, *RRF* est la meilleure option, de façon similaire aux constatations de Cormack *et al.* (2009) dans le domaine de la recherche d'information.

18. *CombSum* attribue un score à chaque voisin, après fusion des listes, égal à la somme des valeurs de similarité de ce voisin dans chacune des listes. La normalisation des valeurs de similarité $sim(v)$ est, quant à elle, donnée par $\frac{sim(v)-min_{sim}}{max_{sim}-min_{sim}}$.

5. Conclusions 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 appliquant 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. Nous avons plus particulièrement testé deux méthodes de fusion des résultats de ces deux critères. L'une, qualifiée de précoce, consiste à regrouper les ensembles d'apprentissage produits par les deux critères. Les améliorations apportées par ce biais sont plus particulièrement notables pour les noms de fréquence faible et pour des mots similaires plutôt que pour de stricts synonymes. L'autre méthode, dite tardive, fusionne les listes de voisins réordonnancés produits par chacun des critères et permet d'obtenir ainsi des résultats nettement supérieurs à la fusion précoce.

Au-delà de leur analyse précise, les résultats obtenus doivent être replacés dans un contexte plus large sur les possibilités offertes pour améliorer les thésaurus distributionnels. Comme nous l'avons vu à la section 2.3, certains paramètres de base intervenant dans la construction des thésaurus distributionnels ont une incidence sensible sur la qualité de ceux-ci, en particulier la taille du corpus utilisé et la nature des contextes distributionnels. Une façon évidente d'améliorer la qualité des thésaurus est ainsi d'augmenter la taille de leur corpus source, tendance que l'on observe clairement dans les travaux récents où les corpus sont rarement inférieurs au milliard de mots. Cette approche est concevable en domaine général où les corpus de grandes tailles ne sont pas rares. Elle est plus difficile à mettre en œuvre dans beaucoup de domaines spécialisés, même si ce n'est pas le cas de tous.

Concernant la nature des contextes distributionnels, l'utilisation de cooccurrents syntaxiques apporte également un plus indéniable. Elle se heurte à la disponibilité d'un analyseur syntaxique et aux temps de traitement qu'il engendre mais le tableau 5 illustre assez bien le fait que les cooccurrents syntaxiques permettent de compenser une taille de corpus plus faible, ce qui module un peu le problème du temps de traitement. Dans le registre des cooccurrents graphiques, Claveau *et al.* (2014) montrent que la directionnalité des cooccurrents, conjuguée dans le cas présent à une fonction de pondération et à une mesure de similarité issues de la recherche d'information[19], a une incidence importante sur les résultats, ce que Curran (2003) avait déjà noté à une moindre échelle. Enfin, des travaux récents suggèrent des gains intéressants en lien avec d'autres aspects des contextes distributionnels, comme la sélection des éléments constituant les contextes et la normalisation de leur poids (Polajnar et Clark, 2014) ou l'adoption d'une variante de la fonction de pondération PPMI compensant sa sensibilité aux faibles fréquences (Levy *et al.*, 2015).

19. Les expériences menées depuis confirment ce gain pour le couple *cosinus* – PPMI.

Comparés aux différences de performance imputables à tous ces facteurs, les gains obtenus par les méthodes d'amorçage proposées sont en apparence assez limités. Mais plusieurs points sont à prendre en considération pour juger de leur intérêt. Tout d'abord, les principes d'amorçage utilisés peuvent être appliqués indépendamment du mode de représentation et de constitution des données distributionnelles. Les améliorations issues de ces données et celles issues des méthodes proposées ici sont donc complémentaires, voire susceptibles de synergie : *a priori*, plus la qualité du thésaurus initial est élevée et plus l'effet d'amorçage doit être lui-même important, de meilleurs exemples devant conduire à un meilleur classifieur et donc *in fine*, à un meilleur réordonnancement des voisins. C'est un effet qu'il nous reste néanmoins à confirmer.

La seconde dimension à prendre en compte est la différenciation des résultats obtenus en fonction des tranches fréquentielles. Si le gain au niveau global reste modeste, il est en revanche beaucoup plus significatif au niveau des entrées de basse fréquence ainsi que l'illustre la figure 3. Or, ces entrées représentent la moitié du vocabulaire du corpus. L'effet plutôt négatif du réordonnancement au niveau des hautes fréquences, sauf dans certains cas avec Moby comme référence, se comprend d'ailleurs assez bien en considérant que les exemples sélectionnés en sont issus, ce qui rend la difficulté d'une amélioration possible beaucoup plus grande. Sur le plan applicatif, une façon directe d'élever la performance globale est donc de construire un thésaurus hybride composé des voisins du thésaurus initial pour les entrées de haute fréquence et des voisins du thésaurus réordonnancé pour les entrées de basse fréquence.

Nous envisageons, par ailleurs, plusieurs pistes d'extension de ce travail. En premier lieu, nous souhaitons élargir les critères de sélection non supervisée d'exemples. 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 ou l'exploitation des chaînes de coréférence, à l'instar de Adel et Schütze (2014) mais en exploitant un système de résolution des coréférences n'intégrant pas déjà la connaissance que l'on cherche à extraire. Au-delà d'une approche purement non supervisée de sélection d'exemples, il serait également intéressant d'étudier dans quelle mesure un ensemble très restreint d'exemples fournis manuellement peut aider ou non à améliorer significativement les résultats. Enfin, sur un autre plan, l'évaluation menée, fondée sur la comparaison avec des ressources de référence, pourrait être complétée avec profit par une évaluation extrinsèque permettant de juger de l'impact des améliorations du thésaurus distributionnel sur une tâche à laquelle il contribue, à l'image de Claveau et Kijak (2015) dans le champ de la recherche d'information pour l'expansion de requêtes. Nous serions, pour notre part, intéressé par une application à la segmentation thématique, dans le prolongement de Adam et Morlane-Hondère (2009).

6. Bibliographie

Adam C., Morlane-Hondère F., « Détection de la cohésion lexicale par voisinage distributionnel : application à la segmentation thématique », *RECITAL'09*, Senlis, France, 2009.

Adel H., Schütze H., « Using Mined Coreference Chains as a Resource for a Semantic Task », *EMNLP 2014*, Doha, Qatar, p. 1447-1452, 2014.

Atrey P. K., Hossain M. A., El Saddik A., Kankanhalli M. S., « Multimodal fusion for multimedia analysis : a survey », *Multimedia Systems*, vol. 16, n° 6, p. 345-379, 2010.

Baroni M., Dinu G., Kruszewski G., « Don't count, predict ! A systematic comparison of context-counting vs. context-predicting semantic vectors », *ACL 2014*, Baltimore, Maryland, USA, p. 238-247, 2014.

Broda B., Piasecki M., Szpakowicz S., « Rank-Based Transformation in Measuring Semantic Relatedness », *Canadian AI 2009*, p. 187-190, 2009.

Bullinaria J. A., Levy J. P., « Extracting semantic representations from word co-occurrence statistics : stop-lists, stemming, and SVD », *Behavior Research Methods*, vol. 44, n° 3, p. 890-907, 2012.

Chang C.-C., Lin C.-J., *LIBSVM : a library for support vector machines*, `http://www.csie.ntu.edu.tw/~cjlin/libsvm`. 2001.

Claveau V., Kijak E., « Thésaurus distributionnels pour la recherche d'information et vice-versa », *CORIA 2015*, Paris, France, 2015.

Claveau V., Kijak E., Ferret O., « Improving distributional thesauri by exploring the graph of neighbors », *COLING 2014*, Dublin, Ireland, p. 709-720, 2014.

Cormack G. V., Clarke C. L. A., Buettcher S., « Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods », *SIGIR'09*, p. 758-759, 2009.

Curran J., « Ensemble Methods for Automatic Thesaurus Extraction », *EMNLP 2002*, p. 222-229, 2002.

Curran J., Moens M., « Improvements in automatic thesaurus extraction », *Workshop of the ACL Special Interest Group on the Lexicon (SIGLEX)*, Philadelphia, USA, p. 59-66, 2002.

Curran J. R., From Distributional to Semantic Similarity, PhD thesis, University of Edinburgh, 2003.

Dinu G., Lapata M., « Measuring Distributional Similarity in Context », *EMNLP 2010*, Cambridge, MA, USA, p. 1162-1172, 2010.

Erk K., Padó S., « Exemplar-Based Models for Word Meaning in Context », *ACL 2010, short paper*, Uppsala, Sweden, p. 92-97, 2010.

Ferret O., « Similarité sémantique et extraction de synonymes à partir de corpus », *TALN 2010*, Montréal, Canada, 2010.

Ferret O., « Combining Bootstrapping and Feature Selection for Improving a Distributional Thesaurus », *ECAI 2012*, Montpellier, France, p. 336-341, 2012.

Ferret O., « Sélection non supervisée de relations sémantiques pour améliorer un thésaurus distributionnel », *TALN 2013*, Les Sables d'Olonne, France, p. 48-61, 2013.

Ferret O., « Compounds and distributional thesauri », *Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, p. 2979-2984, 2014.

Ferret O., « Early and Late Combinations of Criteria for Reranking Distributional Thesauri », *ACL-IJCNLP 2015, short paper session*, Beijing, China, p. 470-476, 2015a.

Ferret O., « Typing relations in distributional thesauri », *in* N. Gala, R. Rapp, G. Bel (eds), *Language Production, Cognition, and the Lexicon*, vol. 48 of *Text, Speech and Language Technology*, Springer, p. 113-134, 2015b.

Freitag D., Blume M., Byrnes J., Chow E., Kapadia S., Rohwer R., Wang Z., « New experiments in distributional representations of synonymy », *CoNLL 2005*, p. 25-32, 2005.

Gabrilovich E., Markovitch S., « Computing semantic relatedness using wikipedia-based explicit semantic analysis », *IJCAI 2007*, Hyderabad, India, p. 6-12, 2007.

Grave E., Obozinski G., Bach F., « A Markovian approach to distributional semantics with application to semantic compositionality », *COLING 2014*, p. 1447-1456, 2014.

Grefenstette G., *Explorations in automatic thesaurus discovery*, Kluwer, 1994.

Hagiwara M., « A Supervised Learning Approach to Automatic Synonym Identification Based on Distributional Features », *ACL-08 : HLT, student session*, p. 1-6, 2008.

Henestroza Anguiano E., Candito M., « Probabilistic Lexical Generalization for French Dependency Parsing », *SP-Sem-MRL 2012 workshop*, Jeju, Republic of Korea, p. 1-11, 2012.

Heylen K., Peirsman Y., Geeraerts D., Speelman D., « Modelling Word Similarity : An Evaluation of Automatic Synonymy Extraction Algorithms », *LREC'08*, 2008.

Hill F., Reichart R., Korhonen A., « SimLex-999 : Evaluating Semantic Models with (Genuine) Similarity Estimation », *CoRR*, 2014.

Huang E. H., Socher R., Manning C. D., Ng A. Y., « Improving word representations via global context and multiple word prototypes », *ACL'12*, p. 873-882, 2012.

Kanerva P., Kristoferson J., Holst A., « Random Indexing of Text Samples for Latent Semantic Analysis », *CogSci 2000*, Lawrence Erlbaum, p. 103-6, 2000.

Kazama J., De Saeger S., Kuroda K., Murata M., Torisawa K., « A Bayesian Method for Robust Estimation of Distributional Similarities », *ACL 2010*, p. 247-256, 2010.

Kiela D., Clark S., « A Systematic Study of Semantic Vector Space Model Parameters », 2^{nd} *CVSC workshop*, Gothenburg, Sweden, p. 21-30, 2014.

Landauer T. K., Dumais S. T., « A solution to Plato's problem : the latent semantic analysis theory of acquisition, induction, and representation of knowledge », *Psychological review*, vol. 104, n° 2, p. 211-240, 1997.

Lapesa G., Evert S., « A Large Scale Evaluation of Distributional Semantic Models : Parameters, Interactions and Model Selection », *TALC*, vol. 2, p. 531-545, 2014.

Lee J. H., « Analyses of Multiple Evidence Combination », *SIGIR'97*, ACM, p. 267-276, 1997.

Levy O., Goldberg Y., Dagan I., « Improving Distributional Similarity with Lessons Learned from Word Embeddings », *TALC*, vol. 3, p. 211-225, 2015.

Lin D., « PRINCIPAR : An efficient, broad-coverage, principle-based parser », *COLING'94*, Kyoto, Japan, p. 42-48, 1994.

Lin D., « Automatic retrieval and clustering of similar words », *ACL-COLING'98*, Montréal, Canada, p. 768-774, 1998.

Mikolov T., Yih W.-t., Zweig G., « Linguistic Regularities in Continuous Space Word Representations », *NAACL HLT 2013*, Atlanta, Georgia, p. 746-751, 2013.

Miller G. A., « WordNet : An On-Line Lexical Database », *International Journal of Lexicography*, 1990.

Min B., Shi S., Grishman R., Lin C.-Y., « Ensemble Semantics for Large-scale Unsupervised Relation Extraction », *EMNLP-CoNLL 2012*, Jeju Island, Korea, p. 1027-1037, July, 2012.

Nuray R., Can F., « Automatic Ranking of Information Retrieval Systems Using Data Fusion », *Information Processing and Management*, vol. 42, n° 3, p. 595-614, 2006.

Pantel P., Crestan E., Borkovsky A., Popescu A.-M., Vyas V., « Web-Scale Distributional Similarity and Entity Set Expansion », *EMNLP 2009*, Singapore, p. 938-947, 2009.

Pennington J., Socher R., Manning C., « Glove : Global Vectors for Word Representation », *EMNLP 2014*, Doha, Qatar, p. 1532-1543, 2014.

Polajnar T., Clark S., « Improving Distributional Semantic Vectors through Context Selection and Normalisation », *EACL 2014*, Gothenburg, Sweden, p. 230-238, 2014.

Popescu A., Grefenstette G., « Social Media Driven Image Retrieval », *1^{st} ACM International Conference on Multimedia Retrieval (ICMR'11)*, ACM, Trento, Italy, p. 1-8, 2011.

Ramisch C., Villavicencio A., Boitet C., « mwetoolkit : a Framework for Multiword Expression Identification », *LREC'10*, Valetta, Malta, p. 662-669, 2010.

Reisinger J., Mooney R. J., « Multi-Prototype Vector-Space Models of Word Meaning », *HLT-NAACL 2010*, Los Angeles, California, USA, p. 109-117, June, 2010.

Schmid H., « Probabilistic Part-of-Speech Tagging Using Decision Trees », *International Conference on New Methods in Language Processing*, Manchester, UK, p. 44-49, 1994.

Van de Cruys T., Mining for Meaning. The Extraction of Lexico-semantic Knowledge from Text, PhD thesis, University of Groningen, The Netherlands, 2010.

Van der Plas L., Bouma G., « Syntactic Contexts for Finding Semantically Related Words », *CLIN 2004*, Leiden, Netherlands, 2004.

Voorhees E., Graff D., *AQUAINT-2 Information-Retrieval Text Research Collection*, `https://catalog.ldc.upenn.edu/LDC2008T25`. 2008.

Ward G., *Moby Thesaurus*, Moby Project, 1996.

Weeds J., Measures and Applications of Lexical Distributional Similarity, PhD thesis, Department of Informatics, University of Sussex, 2003.

Wu Z., Palmer M., « Verbs semantics and lexical selection », *ACL'94*, Las Cruces, New Mexico, USA, p. 133-138, 1994.

Yamamoto K., Asakura T., « Even Unassociated Features Can Improve Lexical Distributional Similarity », *NLPIX 2010 workshop*, Beijing, China, p. 32-39, 2010.

Zhitomirsky-Geffet M., Dagan I., « Bootstrapping Distributional Feature Vector Quality », *Computational Linguistics*, vol. 35, n° 3, p. 435-461, 2009.