Déclasser les voisins non sémantiques pour améliorer les thésaurus distributionnels

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

Résumé. La plupart des méthodes d'amélioration des thésaurus distributionnels se focalisent sur les moyens – représentations ou mesures de similarité – de mieux détecter la similarité sémantique entre les mots. Dans cet article, nous proposons un point de vue inverse : nous cherchons à détecter les voisins sémantiques associés à une entrée les moins susceptibles d'être liés sémantiquement à elle et nous utilisons cette information pour réordonner ces voisins. Pour détecter les faux voisins sémantiques d'une entrée, nous adoptons une approche s'inspirant de la désambiguïsation sémantique en construisant un classifieur permettant de différencier en contexte cette entrée des autres mots. Ce classifieur est ensuite appliqué à un échantillon des occurrences des voisins de l'entrée pour repérer ceux les plus éloignés de l'entrée. Nous évaluons cette méthode pour des thésaurus construits à partir de cooccurrents syntaxiques et nous montrons l'intérêt de la combiner avec les méthodes décrites dans (Ferret, 2013b) selon une stratégie de type vote.

Abstract.
Downgrading non-semantic neighbors for improving distributional thesauri
Most of the methods for improving distributional thesauri focus on the means – representations or similarity measures – to detect better semantic similarity between words. In this article, we propose a more indirect approach focusing on the identification of the neighbors of a thesaurus entry that are not semantically linked to this entry. This identification relies on a discriminative classifier trained from unsupervised selected examples for building a distributional model of the entry in texts. Its bad neighbors are found by applying this classifier to a representative set of occurrences of each of these neighbors. We evaluate more particularly the interest of this method for thesauri built from syntactic co-occurrents and we show the interest of associating this method with those of (Ferret, 2013b) following an ensemble strategy.

Mots-clés : Sémantique lexicale, similarité sémantique, thésaurus distributionnels.

Keywords: Lexical semantics, semantic similarity, distributional thesauri.

1 Introduction

Les ressources distributionnelles sont utilisées dans un ensemble de tâches de plus en plus important, allant de l'extraction de relations (Min *et al.*, 2012) à l'analyse syntaxique (Henestroza Anguiano & Candito, 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 du 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 & 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 quelques 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, en s'attachant à la pondération des éléments constituant les contextes distributionnels, à l'instar de (Broda *et al.*, 2009), (Zhitomirsky-Geffet & Dagan, 2009) ou (Yamamoto & Asakura, 2010). (Kazama *et al.*, 2010) a pour sa part adopté un point de vue bayésien pour aborder la question.

advance	gain [0,13], surge [0,10], progress [0,09], improvement [0,09], increase [0,09], decline [0,09] ...
distress	pain [0,14], anguish [0,13], anxiety [0,13], discomfort [0,12], grief [0,12], hardship [0,11] ...
inquisitor	good-cop [0,06], uranium-enrichment [0,06], misanthrope [0,05], interviewer [0,05] ...
insomniac	angler [0,07], tonsillitis [0,07], procrastinator [0,06], shuffler [0,06], grenadian [0,06] ...

TABLE 1 – Premiers voisins de quelques entrées du thésaurus distributionnel de la section 2

D'autres travaux ont envisagé des changements plus radicaux. Les modèles à base d'exemples (Erk & Pado, 2010) ou de prototypes multiples (Reisinger & 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 méthodes de factorisation de matrice comme l'analyse sémantique latente (Padó & Lapata, 2007) ou la factorisation de matrice non négative (Van de Cruys, 2010), de méthodes fondées sur la notion de hachage comme le Random Indexing (Kanerva *et al.*, 2000) ou plus récemment des méthodes 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 & Dagan, 2009) et (Yamamoto & Asakura, 2010) relèvent de cette problématique. Ils s'appuient en effet sur un mécanisme d'amorçage dans lequel la première étape consiste à trouver des voisins sémantiques selon une approche du type (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) et de (Ferret, 2013b). 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 purement sur le thésaurus initial en exploitant ses relations de symétrie tandis que (Ferret, 2013b) utilise en complément un thésaurus distributionnel de mots composés. Claveau *et al.* (2014) proposent quant à eux plusieurs façons de généraliser l'idée avancée dans (Ferret, 2012) de l'exploitation des relations à l'échelle du thésaurus.

Le travail que nous présentons dans cet article reprend l'optique, développée dans les travaux du paragraphe précédent, d'une amélioration d'un thésaurus distributionnel fondé sur un processus de réordonnancement de ses voisins et s'inscrit de ce point de vue dans une nouvelle approche visant à identifier les voisins les moins susceptibles d'être en relation sémantique avec leur entrée afin de les déclasser. Plus précisément, nous verrons comment cette approche, appliquée dans le contexte de thésaurus construits sur la base de cooccurrents graphiques (Ferret, 2013a), c'est-à-dire extraits sur la base d'une fenêtre glissante de taille fixe, peut également être appliquée avec succès à des thésaurus fondés sur des cooccurrents syntaxiques. Nous montrerons également l'intérêt de combiner cette approche avec celle présentée dans (Ferret, 2013b).

2 Thésaurus initial

Avant de présenter plus avant la méthode d'amélioration des thésaurus distributionnels que nous proposons, nous présenterons en premier lieu la façon dont nous construisons de tels thésaurus et la façon dont nous les évaluons, en les comparant de ce point de vue aux travaux de référence et aux dernières avancées dans ce domaine.

2.1 Construction du thésaurus initial

Le processus de construction du thésaurus initial que nous avons suivi est très comparable à celui décrit dans (Grefenstette, 1994), (Lin, 1998) ou (Curran & Moens, 2002) pour la construction de thésaurus fondés sur des cooccurrents syntaxiques. Il reprend très concrètement le processus décrit dans (Ferret, 2010) pour les cooccurrences graphiques, avec quelques adaptations. Le point de départ est constitué comme dans (Ferret, 2010) par les 380 millions de mots du corpus AQUAINT-2 provenant d'articles de journaux écrits en anglais. Dans le cas présent, le prétraitement linguistique du corpus a été réalisé par l'analyse syntaxique MINIPAR (Lin, 1994) permettant de fournir trois types d'informations exploités pour la construction du thésaurus : la forme lemmatisée des mots, leur catégorie morphosyntaxique et plus spécifiquement

type cooc.	réf.	#mots éval.	#syn. / mot	rappel	R-préc.	MAP	P@1	P@5	P@10	P@100
	W	10 178	2,9	31,5	11,5	13,3	15,6	6,9	4,5	0,9
syntaxiques	M	9 060	50,4	12,8	9,4	4,8	30,6	21,7	17,3	6,5
14 117	WM	11 887	39,4	13,2	10,8	7,9	29,4	18,9	14,6	5,2
	W	10 473	2,9	24,6	8,2	9,8	11,7	5,1	3,4	0,7
graphiques	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

TABLE 2 – Évaluation du thésaurus initial fondé sur des cooccurrents syntaxiques et comparaison par rapport à l'usage de cooccurrents graphiques

les relations de dépendance syntaxique qui les unissent. Pour la constitution des données distributionnelles, la principale différence avec (Ferret, 2010) réside au niveau des éléments constitutifs des contextes distributionnels, prenant la forme de cooccurrents syntaxiques. Plus précisément, chaque cooccurrent est représenté sous la forme de la paire (*relation syntaxique*, *lemme cooccurrent*), en se limitant aux noms, verbes et adjectifs. Le filtrage fréquentiel des contextes, éliminant les cooccurrents de fréquence 1, la pondération des cooccurrents par l'information mutuelle ramenée aux valeurs positives (PPMI : *Positive Pointwise Mutual Information*) et l'adoption de la mesure *Cosinus* pour évaluer la similarité des contextes sont pour leur part similaires aux paramètres sélectionnés par (Ferret, 2010) ainsi qu'aux conclusions d'études récentes et plus exhaustives couvrant aussi les cooccurrents syntaxiques, comme (Kiela & Clark, 2014) [1]. Un filtre fréquentiel a en outre été appliqué aux mots cibles et à leurs cooccurrents et ne conserve que les mots de fréquence supérieure à 10.

La construction proprement dite du thésaurus a été réalisée de façon classique en sélectionnant les voisins sémantiques les plus proches de chaque entrée considérée selon la mesure de similarité entre contextes. Plus précisément, cette mesure a été calculée entre chaque entrée et l'ensemble de ses voisins possibles et ceux-ci ont été ordonnés selon l'ordre décroissant des valeurs calculées. Seuls les 100 premiers voisins ont alors été conservés pour chaque entrée du thésaurus. Entrées et voisins se limitent dans le cas présent à des noms [2].

2.2 Évaluation et mise en perspective du thésaurus initial

Le tableau 2 donne les résultats de l'évaluation du thésaurus distributionnel obtenu (lignes *syntaxiques*) et les met en perspective avec ceux du thésaurus construit dans (Ferret, 2010) sur la base du même corpus mais avec des cooccurrents graphiques (lignes *graphiques*). Cette évaluation est réalisée comme dans (Ferret, 2010) en comparant les voisins sémantiques extraits à deux ressources de référence complémentaires : les synonymes de WordNet [W] (Miller, 1990), dans sa version 3.0, qui permettent de caractériser une similarité fondée sur des relations paradigmatiques et le thésaurus Moby [M] (Ward, 1996), qui regroupe des mots liés par des relations plus diverses de proximité sémantique. 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]. Compte tenu de l'inévitable incomplétude de ces ressources de référence vis-à-vis des notions de similarité et de proximité sémantique, les chiffres du tableau 2 doivent être considérés comme des minima.

Ces résultats se déclinent sous la forme de différentes mesures, à commencer à la $5^{ème}$ colonne par le taux de rappel par rapport aux ressources considérées pour les 100 premiers voisins de chaque nom. Ces voisins étant ordonnés, il est en outre possible de réutiliser les métriques d'évaluation classiquement adoptées en recherche d'information en faisant jouer aux mots cibles le rôle de requêtes et aux voisins celui des documents. Les dernières colonnes du tableau 2 rendent compte de ces mesures : la R-précision (R-préc.) est la précision obtenue en se limitant aux R premiers voisins, R étant le nombre de synonymes dans la ressource de référence pour l'entrée considérée ; la MAP (Mean Average Precision) est la moyenne des précisions pour chacun des rangs auxquels un synonyme de référence a été identifié ; enfin, sont données les précisions pour différents seuils de nombre de voisins sémantiques examinés (précision après examen des 1, 5, 10 et 100 premiers voisins). Pour une meilleure lisibilité, toutes ces valeurs sont données sur une échelle de 0 à 100.

1. Claveau *et al.* (2014) ont montré qu'une version ajustée de la pondération Okapi-BM25 permet de dépasser les performances de PPMI pour des cooccurrents graphiques mais les différences par rapport au point de comparaison adopté ne permettent pas d'attribuer de façon sûre cette amélioration des performances au type de pondération.

2. Le thésaurus construit est disponible sur le site `https://github.com/osf9018/a2st`.

De façon non surprenante, le premier constat qu'impose le tableau 2 est la supériorité des cooccurrents syntaxiques sur les cooccurrents graphiques, même si cette supériorité s'accompagne d'un plus grand nombre d'entrées dépourvues de voisins. La différence de 553 entrées est toutefois faible au regard du nombre total d'entrées et s'explique *a priori* par la densité moindre des cooccurrents syntaxiques par rapport aux cooccurrents graphiques, aboutissant parfois à une intersection vide des contextes distributionnels. En revanche, les observations faites dans (Ferret, 2010) à propos des cooccurrents graphiques se confirment ici avec des cooccurrents syntaxiques. En particulier, les résultats obtenus sont fortement sensibles à la ressource de référence utilisée pour l'évaluation, à la fois du point de vue du nombre de voisins par entrée et du type de ces voisins. Ainsi, la précision à différents rangs est significativement plus forte avec Moby comme référence, qui fournit beaucoup de voisins de différents types pour chaque entrée, qu'avec WordNet, qui ne donne pour chaque entrée qu'un ensemble restreint de synonymes. La MAP et la R-précision montrent une tendance inverse pour les mêmes raisons. Il est à noter que la richesse des références utilisées explique au moins en partie pourquoi des travaux comme (Curran & Moens, 2002) ou plus récemment (Riedl & Biemann, 2013) affichent des valeurs élevées pour la précision au rang 1 par exemple.

méthode	**#mots éval.**	**#syn./ mot**	**rappel**	**R-préc.**	**MAP**	**P@1**	**P@5**	**P@10**	**P@100**
A2ST-SYNT	11 887	39,4	13,2	10,8	7,9	29,4	18,9	14,6	5,2
(Lin, 1998)	9 823	44,5	12,7	11,6	8,1	36,1	23,7	18,2	5,6
(Huang *et al.*, 2012)	10 537	42,6	3,8	1,9	0,8	7,1	5,0	4,0	1,6
(Mikolov *et al.*, 2013)	12 326	38,6	6,2	5,5	4,2	16,3	9,5	7,0	2,4
(Baroni *et al.*, 2014)-count	12 052	39,3	13,6	12,5	9,8	31,9	19,6	15,2	5,3
(Baroni *et al.*, 2014)-predict	12 052	39.3	11.3	10.9	8.5	30.3	18.4	13.8	4.4

TABLE 3 – Comparaison de plusieurs approches pour la construction de thésaurus distributionnels

Le tableau 3 permet de mettre en perspective les résultats de notre thésaurus (A2ST-SYNT) avec ceux de plusieurs autres thésaurus, en utilisant [WM] comme référence. (Lin, 1998) 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 milliards 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, en moyenne de plus forte fréquence comme le montre le nombre de synonymes par entrée. (Huang *et al.*, 2012) et (Mikolov *et al.*, 2013) correspondent quant à eux à deux approches récentes fondées sur la construction de représentations distribuées de mots, appelées *word embeddings*, par des réseaux de neurones. Ces représentations sont utilisées en lieu et place des contextes distributionnels classiques lors de la construction des thésaurus. Dans le cas de (Huang *et al.*, 2012), nous avons utilisé les représentations construites à partir de Wikipédia [4] fournies par les auteurs tandis que dans le cas de (Mikolov *et al.*, 2013), nous avons calculé ces représentations à partir du corpus AQUAINT-2 en utilisant les meilleurs paramètres du modèle *Skip-gram* sélectionnés par (Mikolov *et al.*, 2013). Dans les deux cas, les résultats sont significativement inférieurs à ceux de A2ST-SYNT, avec un niveau particulièrement bas pour (Huang *et al.*, 2012) 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 l'option la plus intéressante pour la construction de thésaurus distributionnels. Ce constat est renforcé par les deux dernières lignes du tableau 3, qui donnent les résultats des thésaurus construits avec les vecteurs de contexte mis à disposition par Baroni *et al.* (2014) [5] : *(Baroni* et al.*, 2014)-predict* correspond à des vecteurs construits grâce au modèle CBOW de (Mikolov *et al.*, 2013) tandis que *(Baroni* et al.*, 2014)-count* 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 de 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.*, 2014)-count* par rapport à *(Baroni* et al.*, 2014)-predict*, ce qui vient limiter le constat général fait par Baroni *et al.* (2014) de la supériorité de l'approche *predict* par rapport à l'approche *count*. Ce constat n'est visiblement pas vérifié dans le cas des thésaurus distributionnels.

3. http://webdocs.cs.ualberta.ca/lindek/Downloads/sim.tgz
4. http://nlp.stanford.edu/~socherr/ACL2012_wordVectorsTextFile.zip
5. http://clic.cimec.unitn.it/composes/semantic-vectors.html

3 Principes et vue d'ensemble

Dans ce travail, nous adoptons une approche indirecte de l'amélioration des thésaurus distributionnels en nous focalisant sur la détection des voisins les moins sémantiquement liés à leur entrée. En dehors du tableau 2, les exemples du tableau 1 nous permettent d'appréhender plus qualitativement ces voisins « pas très sémantiques ». Ainsi, le mot *misanthrope* pour l'entrée *inquisitor* ou le mot *procrastinator* pour l'entrée *insomniac* sont des exemples assez évidents que certains voisins sont à rejeter sans équivoque. En détectant ces voisins et en les repoussant en queue de liste, les mots véritablement liés à l'entrée considérée, lorsqu'ils sont présents dans la liste de ses voisins, voient leur rang augmenter de façon mécanique. Il en résulte un thésaurus plus intéressant à exploiter dans la mesure où en pratique, seuls les tout premiers voisins d'une entrée sont utilisables compte tenu de l'augmentation importante du taux d'erreur en fonction du rang.

L'approche que nous proposons pour identifier ces voisins pas très sémantiques repose, comme la construction du thésaurus initial, sur l'hypothèse distributionnelle. Sa mise en œuvre est en revanche différente. Cette hypothèse stipule dans les grandes lignes que la signification d'un mot peut être caractérisée par l'ensemble des contextes dans lesquels ce mot est rencontré. De ce fait, deux mots sont considérés comme sémantiquement similaires s'ils sont observés dans un ensemble suffisant large de contextes communs. Dans les travaux du type de (Curran & Moens, 2002), cette approche est implémentée en associant à chaque mot un contexte distributionnel rassemblant tous les mots avec lesquels il cooccurre dans un large corpus. Suivant que les cooccurrents sont graphiques ou syntaxiques, un tel contexte est une représentation peu structurée de type « sac de mots » ou « sac de paires (mot, relation syntaxique) ». (Kazama *et al.*, 2010) propose une variante de ce schéma en modélisant la représentation distributionnelle d'un mot sous la forme d'une distribution multinomiale, sans que cela n'induise de nouvelles possibilités de représentation.

Or, cette approche présente l'inconvénient de n'autoriser qu'une faible diversité et une faible structuration des éléments utilisés pour construire les modèles. Des traits comme des ngrammes de mots ou de catégories morphosyntaxiques ne sont ainsi pas utilisés pour la construction de thésaurus alors qu'ils le sont largement pour des tâches comme la désambiguïsation sémantique par exemple. Les inclure dans une représentation de type « sac de mots » conduirait en effet à une inflation très significative des modèles, sans compter que des mesures de similarité telles que la mesure *Cosinus* ne sont pas adaptées à des représentations hétérogènes (Alexandrescu & Kirchhoff, 2007). Pour dépasser cette limite, nous proposons donc de construire un modèle discriminant pour représenter les contextes d'un mot, ce type de modèle étant naturellement capable d'intégrer une large diversité et un grand nombre de traits. Un tel modèle a plus précisément pour objectif de permettre la discrimination sémantique d'un mot en contexte, c'est-à-dire une occurrence de ce mot dans un corpus, vis-à-vis de tous les autres mots, et plus particulièrement de ceux de ses voisins issus d'un thésaurus distributionnel les plus éloignés de lui sur le plan sémantique. L'hypothèse sous-jacente s'appuie sur l'idée distributionnelle : un mot et un synonyme de ce mot doivent apparaître dans les mêmes contextes et sont donc représentables par les mêmes traits puisque ceux-ci sont extraits de ces contextes. Un modèle reposant sur ces traits capable d'identifier l'occurrence d'un mot en contexte doit donc être à même d'identifier une occurrence d'un de ses synonymes et plus généralement une occurrence d'un mot qui peut se substituer à lui sur le plan paradigmatique. Comme nous le démontrerons dans ce qui suit, un tel modèle est en pratique particulièrement bien adapté à la détection des mots qui ne sont pas sémantiquement liés au mot cible de ce modèle.

Le fait de vouloir capturer les contextes d'un mot au travers d'un modèle discriminant pose néanmoins un problème spécifique. Dans le cas d'un modèle de type « sac de mots », les contextes de deux mots peuvent être directement comparés par le biais d'une mesure de similarité. Dans le cas d'un modèle discriminant, une telle comparaison directe n'est pas possible. Ils ne peuvent en fait être comparés que le biais d'une application à un ensemble d'exemples, c'est-à-dire d'occurrences de mots dans le cas présent. Ainsi, pour déterminer si un voisin d'une entrée d'un thésaurus est un « mauvais » voisin du point de vue sémantique, nous avons choisi d'appliquer un modèle discriminant appris pour cette entrée à un échantillon d'occurrences de ce voisin issu du corpus utilisé pour construire le thésaurus. D'un point de plus global, la méthode d'amélioration d'un thésaurus distributionnel proposée se caractérise donc par l'application des cinq étapes suivantes pour chaque entrée du thésaurus :

– construction d'un classifieur déterminant si une occurrence d'un mot s'identifie ou non à une occurrence de l'entrée considérée ;
– sélection d'un échantillon d'occurrences pour chacun des voisins de l'entrée dans le thésaurus. En pratique, chaque occurrence correspond à une phrase ;
– application du classifieur de la première étape à ces échantillons ;
– détection des mauvais voisins en fonction de ces résultats de classification ;
– réordonnancement des voisins de l'entrée considérée par déclassement des mauvais voisins détectés.

4 Améliorer un thésaurus distributionnel

4.1 Construction des modèles discriminants des mots en contexte

La première, et la plus importante étape de notre processus d'amélioration d'un thésaurus distributionnel est la définition d'un modèle déterminant dans quelle mesure l'occurrence O_i d'un mot V peut être une occurrence d'un mot de référence E, en faisant bien entendu abstraction du rattachement effectif de O_i à V. Cette tâche peut également être envisagée comme une tâche d'étiquetage dans laquelle les occurrences d'un mot cible V ont deux étiquettes possibles : E et $nonE$. Dans le contexte plus général qui nous est propre, l'intérêt de cette tâche n'est pas à prendre au premier degré mais vient plutôt du fait qu'un tel classifieur est susceptible de modéliser les contextes dans lesquels E est observé et donc, si l'on s'attache à l'hypothèse distributionnelle, de modéliser également son sens.

En allant un pas plus loin dans cette logique, un tel classifieur peut être vu comme un moyen de tester si un mot a le même sens que E. Il s'agit là d'un problème très proche de la notion de pseudo désambiguïsation sémantique, au sens de (Gale *et al.*, 1992) : un pseudo-mot est créé avec deux sens, E et $nonE$, $nonE$ prenant la forme d'un ou plusieurs mots dont le sens est supposé représentatif de sens autres que celui de E. L'objectif est alors de construire un classifieur permet de distinguer en contexte les pseudo-sens E et $nonE$. De cette vision, découle la décision d'adopter, pour construire notre classifieur, les mêmes traits que ceux utilisés les plus couramment en désambiguïsation sémantique. Nous avons sur ce plan suivi (Lee & Ng, 2002), un travail de référence dans ce domaine, en choisissant un classifieur à base de Machine à Vecteurs de Support (SVM) avec un noyau linéaire et les trois catégories de traits suivantes pour caractériser chaque occurrence de E et des mots représentatifs de $nonE$ [6] :

– les mots environnants ;
– la catégorie morphosyntaxique des mots environnants ;
– les collocations locales.

Pour les *mots environnants*, nous avons retenu tous les mots pleins (noms communs et noms propres, verbes et adjectifs) ainsi que les adverbes présents dans la même phrase qu'une occurrence de E. Chaque mot environnant est représenté par son lemme sous la forme d'un trait binaire dont la valeur est égale à 1 en cas de présence dans la même phrase qu'une occurrence de E. Pour le deuxième type de traits, nous avons considéré la catégorie morphosyntaxique des trois mots précédant et des trois mots suivant une occurrence de E. Chaque couple {catégorie, position} donne lieu à un trait binaire. Le symbole spécial *empty* est utilisé pour remplacer la catégorie morphosyntaxique lorsque la position se situe au-delà de la fin de la phrase ou précède son début. Enfin, les collocations locales correspondent à des couples de mots présents dans le voisinage d'une occurrence de E. Une collocation est notée $C_{i,j}$, avec i et j faisant référence aux positions respectives du premier et du second mot de la collocation. Dans notre cas, i et j prennent leurs valeurs dans l'intervalle $[-3, +3]$, à l'image des catégories morphosyntaxiques. Plus précisément, les 11 collocations suivantes sont extraites pour chaque occurrence de E :

– $C_{-3,-1}$, $C_{-2,-2}$, $C_{-2,-1}$, $C_{-1,-1}$
– $C_{-2,1}$, $C_{-1,1}$, $C_{-1,2}$
– $C_{1,1}$, $C_{1,2}$, $C_{1,3}$, $C_{2,2}$

Comme dans le cas des catégories morphosyntaxiques, le symbole spécial *empty* est utilisé pour les mots précédant le début de la phrase ou suivant la fin de celle-ci et à l'instar des mots environnants, les mots des collocations sont donnés sous une forme lemmatisée. Chaque instance de l'un de ces 11 schémas de collocations est représentée par un tuple ⟨lemme1, position1, lemme2, position2⟩, donnant lieu également à un trait binaire pour le classifieur SVM.

Conformément au processus global décrit à la section précédente, un classifieur SVM spécifique a été entraîné pour chaque entrée de notre thésaurus initial, ce qui requiert la sélection d'un ensemble d'exemples positifs et négatifs. Pour les exemples positifs, nous avons simplement choisi de manière aléatoire un nombre fixe de phrases issues du corpus AQUAINT-2 contenant au moins une occurrence de l'entrée, la première occurrence dans une phrase étant retenue comme exemple positif. La sélection des exemples négatifs a quant à elle été guidée par notre thésaurus, dans la perspective de s'appuyer sur des critères de proximité sémantique par rapport à l'entrée. De ce point de vue, le choix d'un voisin de l'entrée éloigné en termes de rang aurait garanti un faible nombre de faux exemples négatifs, c'est-à-dire de mots similaires à l'entrée [7], dans la mesure où les performances décroissent rapidement à mesure que le rang des voisins

6. Dans ce qui suit nous ferons référence aux traits associés aux occurrences de E mais les mêmes traits sont associés aux occurrences des mots représentatifs de $nonE$.
7. Formellement, les exemples sont des occurrences de mots mais nous parlerons parfois de mots pour simplifier l'expression.

augmente, comme le montre le tableau 2. En pratique, prendre comme exemples négatifs des voisins ayant un rang plus faible est une meilleure option dans la mesure où ils sont plus utiles en termes de discrimination, étant plus proches de la zone de transition entre exemples positifs et négatifs. Pour limiter les risques de faux exemples négatifs, nous avons réparti notre sélection sur trois rangs, en l'occurrence les rangs 10, 15 et 20, sans optimisation particulière. Pour chacun de ces exemples négatifs, un nombre fixe de phrases ont été sélectionnées de la même façon que pour les exemples positifs. En moyenne, le nombre d'exemples négatifs est donc trois fois plus important que le nombre d'exemples positifs, ce qui reflète la présence très majoritaire de voisins non sémantiquement liés à leur entrée dans le thésaurus initial.

4.2 Identification des mauvais voisins et réordonnancement du thésaurus

Pour l'identification des « mauvais » voisins d'une entrée, le classifieur développé pour reconnaître les occurrences de cette entrée est appliqué à un nombre fixe d'occurrences représentatives de chacun de ses voisins. La sélection de ces occurrences s'effectue de la même façon que la sélection des exemples positifs et négatifs décrite à la section précédente. L'application de ce classifieur vise à déterminer si le contexte de l'occurrence considérée est compatible avec le contexte d'une occurrence de l'entrée. En pratique, la décision de ce classifieur est rarement positive, ce qui n'est pas complètement surprenant : même si deux mots sont sémantiquement équivalents, chacun d'entre eux est caractérisé par des usages spécifiques, en particulier dans un corpus donné, et certains traits utilisés par notre classifieur, comme les collocations, sont plus susceptibles que les contextes distributionnels classiques de capturer de telles spécificités. De ce fait, nous faisons l'hypothèse qu'une décision positive du classifieur est un indice fort de la présence d'un voisin sémantiquement similaire à l'entrée et nous considérons un voisin comme potentiellement « bon » si au moins un nombre minimal G de ses occurrences sélectionnées sont classées positivement. À l'inverse, un voisin est considéré comme « mauvais » si le nombre de décisions positives du classifieur est inférieur à G. Les voisins identifiés comme mauvais ne sont pas complètement écartés mais font l'objet d'une rétrogradation en queue de la liste des voisins de l'entrée considérée, leur ordre initial relatif restant inchangé. Il est enfin à noter que le classifieur de mots en contexte n'est pas appliqué aux voisins dont certaines occurrences ont été utilisées en tant qu'exemples négatifs lors de son entraînement. Ces voisins se verraient en effet très souvent rétrogradés alors qu'ils occupent de faibles rangs et qu'ils sont de ce fait liés de façon effective à l'entrée dans une proportion de cas non négligeable.

5 Expérimentations et évaluation

5.1 Mise en œuvre

La mise en œuvre de la méthode que nous avons présentée ci-dessus nécessite de préciser plusieurs points. L'un d'entre eux concerne la taille des échantillons de phrases à sélectionner à la fois pour les entrées du thésaurus et pour leurs voisins. Ces phrases, et plus précisément les occurrences de mots qu'elles contiennent, sont utilisées à la fois pour l'entraînement du classifieur de mots en contexte et pour l'identification des mauvais voisins. En pratique, nous avons sélectionné aléatoirement un échantillon de 250 phrases pour chaque mot de notre vocabulaire et nous avons exploité l'ensemble ainsi constitué pour les deux tâches. Cet échantillonnage a été réalisé sur la base de la forme lemmatisée des mots du vocabulaire. Par ailleurs, le chiffre de 250 est une limite haute dans la mesure où beaucoup de mots du vocabulaire ont une fréquence inférieure à cette limite, la fréquence minimale étant égale à 11 et la fréquence médiane à 249. Cette limite représente une forme de moyen terme entre les 385 exemples d'entraînement en moyenne de la tâche *Lexical Sample* de l'évaluation *Senseval 1* et les 118 exemples de la même tâche de *Senseval 2*.

Les modalités d'entraînement de nos classifieurs de mots en contexte sont également à préciser. Ceux-ci étant des SVM linéaires, seul le paramètre de régularisation C peut être optimisé. Néanmoins, le nombre de ces classifieurs étant égal au nombre d'entrées du thésaurus, le coût d'une telle optimisation n'est pas complètement négligeable. Nous avons donc évalué dans un premier temps ces classifieurs par le biais d'une procédure de validation croisée à 5 volets, avec une valeur de C par défaut égale à 1. Le tableau 4 donne l'exactitude moyenne de ces classifieurs, ainsi que leur écart-type, pour l'ensemble des entrées du thésaurus ainsi que pour une segmentation selon trois tranches fréquentielles (basses : fréquence < 100 ; moyennes : $100 <$ fréquence ≤ 1000 ; hautes : fréquence > 1000).

Ce tableau montre que le niveau général de résultat de ces classifieurs est élevé, avec des valeurs d'exactitude très similaires pour les différentes tranches fréquentielles [8]. Par ailleurs, les tentatives d'optimisation du paramètre C que nous

8. L'écart-type pour les fréquences basses est un peu plus élevé mais peut s'expliquer par le nombre d'exemples assez faible pour ces fréquences, ce

	toutes	hautes	moyennes	basses
exactitude	86,2	86,1	86,0	86,5
écart-type	6,1	4,2	5,7	7,6

TABLE 4 – Résultats des classifieurs de mots en contexte

avons réalisées pour quelques entrées du thésaurus n'ont pas montré d'amélioration possible. Nous avons donc décidé, à l'échelle de toutes les entrées du thésaurus, de conserver le paramètre C à la valeur de 1.

	3	5	7	10	15
R-préc.	11,2	11,1	11,1	11,0	10,8
P@5	19,4	19,5	19,5	19,4	19,3

TABLE 5 – R-précision et précision au rang 5 pour différentes valeurs de G, avec la référence [WM]

Le dernier paramètre à fixer dans la méthode considérée ici est le paramètre G déterminant le nombre d'occurrences d'un voisin classées négativement par le classifieur en contexte de l'entrée en dessous duquel ce voisin est considéré comme mauvais. Le tableau 5 illustre l'influence de ce paramètre en donnant les résultats obtenus pour différentes valeurs de G en termes de R-précision et précision au rang 5 par rapport à la référence WM. Globalement, il laisse apparaître que cette influence est faible. Les valeurs les plus élevées de G dégradent les résultats globaux mais cette tendance n'est pas très marquée. Cette relative insensibilité aux valeurs de G laisse à penser que l'assimilation d'une occurrence d'un voisin à une entrée par le classifieur en contexte de celle-ci est un indice très fort du fait que ce voisin n'est probablement pas un mauvais voisin. Compte tenu de ce constat, la valeur $G = 3$ a été retenue pour les résultats de la section suivante.

5.2 Évaluation après réordonnancement

Le tableau 6 donne les résultats de l'évaluation, selon les mêmes modalités qu'à la section 2.2, du thésaurus obtenu après réordonnancement de ses voisins selon la méthode présentée ci-dessus. Chaque mesure est accompagnée de sa différence en valeur par rapport à la mesure correspondante pour le thésaurus initial. Les résultats du nouveau thésaurus (lignes *syntaxiques*) sont comme précédemment mis en balance avec les résultats déjà obtenus pour les cooccurrents graphiques.

La première observation suscitée par ce tableau est le fait qu'au niveau global, toutes les mesures sont améliorées de façon significative [9], ce qui distingue la méthode considérée favorablement par rapport à (Ferret, 2012) et (Ferret, 2013b) qui enregistraient certaines baisses avec WordNet comme référence, même si ces baisses étaient souvent non significatives. Cet état de fait trouve son explication probable dans le très faible taux d'erreur de la méthode d'identification des mauvais voisins. En prenant WordNet comme référence, seuls 670 voisins répartis sur 570 entrées ont été faussement identifiés comme de mauvais voisins, ce qui ne représente que 4,7% des voisins dégradés. Ce résultat valide en outre, au moins partiellement, la possibilité de capturer par le biais d'un classifieur discriminant intervenant au niveau des occurrences de mots une partie au moins des propriétés distributionnelles de ceux-ci. Il est d'ailleurs à noter que les traits utilisés dans le cas présent sont adaptés à la désambiguïsation sémantique, ce qui ne les prédispose pas pour autant à rendre compte de tous les aspects sémantiques des mots. Des traits plus spécifiques à la tâche considérée pourraient éventuellement conduire à de meilleurs résultats. Sur un autre plan, il faut préciser que les améliorations sont obtenues pour toutes les fréquences, y compris les plus faibles. Le petit nombre d'exemples en contexte pour ces entrées aurait pu faire craindre des améliorations moindres, voire des dégradations, ce qui n'est pas le cas. Même si l'influence du nombre d'exemples fournis reste à analyser, cette observation suggère que l'approche n'est pas très sensible à la valeur de ce paramètre.

Le deuxième grand enseignement apporté par le tableau 6 concerne le type des relations sémantiques. Les résultats avec Moby comme référence se trouvent en effet améliorés de façon plus importante que ceux obtenus avec WordNet comme référence. Cette tendance peut paraître surprenante au premier abord dans la mesure où la méthode présentée, en mettant l'accent sur la caractérisation en contexte d'un mot par opposition à tous les autres, pourrait *a priori* sembler favoriser les voisins les plus directement substituables à lui, donc des synonymes. Mais il faut aussi considérer que la synonymie n'est pas la seule relation sémantique de nature paradigmatique et que Moby abrite aussi beaucoup de ces relations. Ainsi, si

qui est source d'instabilité.

9. La significativité statistique des différences avec le thésaurus initial est évaluée grâce à un test de Wilcoxon pour échantillons appariés avec un seuil de significativité de 0,01.

type cooc.	réf.	R-préc.	MAP	P@1	P@5	P@10
syntaxiques	W	11,8 (0,3)	13,7 (0,4)	16,1 (0,6)	7,1 (0,2)	4,6 (0,1)
	M	9,6 (0,2)	4,9 (0,1)	31,6 (1,0)	22,3 (0,6)	17,8 (0,5)
	WM	11,2 (0,5)	8,2 (0,3)	30,3 (0,9)	19,4 (0,5)	15,1 (0,5)
graphiques	W	9,1 (0,9)	10,7 (0,9)	12,8 (1,1)	5,6 (0,5)	3,7 (0,3)
	M	7,2 (0,5)	3,5 (0,3)	26,5 (2,4)	17,9 (1,5)	14,0 (1,0)
	WM	8,4 (0,7)	6,1 (0,5)	24,8 (2,3)	15,4 (1,3)	11,7 (0,9)

TABLE 6 – Résultat du réordonnancement des voisins sémantiques

l'on utilise WordNet comme moyen d'analyse des relations présentes dans Moby, on constate que la cohyponymie y est la relation la plus fréquente et que l'hyponymie et l'hyperonymie directes y occupent respectivement les $6^{ème}$ et $7^{ème}$ rangs.

Le tableau 6 permet enfin de constater que la méthode décrite dans cet article est plus efficace pour les cooccurrents graphiques que pour les cooccurrents syntaxiques, même si dans ce dernier cas, elle apporte tout de même un plus significatif. Le niveau initial plus élevé des résultats dans le cas des cooccurrents graphiques est l'explication la plus évidente de ce constat mais il faut noter que l'absence de prise en compte des relations syntaxiques au niveau du classifieur de mots en contexte est une source d'améliorations non exploitée ici. Il faut aussi souligner que le thésaurus « syntaxique » est moins riche que le thésaurus fondé sur des cooccurrents graphiques en entrées de faible fréquence, entrées qui bénéficient précisément le plus des améliorations mises en œuvre.

WordNet	catastrophe, calamity, tragedy, disaster
Moby	accident, apoplexy, blow, breakdown, breakup, calamity, casualty, catastrophe, climax, collapse, collision, convulsion, crash, debacle + 35 mots supplémentaires
initial	divisionism, **calamity**, upheaval, **catastrophe**, schism, landslip, **disaster**, devastation, conflagration, **tragedy**, deterioration, shakeout …
réordonnancés	**calamity**, upheaval, **catastrophe**, schism, **disaster**, devastation, conflagration, **tragedy**, deterioration, shakeout, maneuvering, displacement …

TABLE 7 – Impact de notre réordonnancement pour l'entrée *cataclysm*

Le tableau 7 apporte quant à lui une vue plus qualitative des résultats de la procédure de réordonnancement présentée en l'illustrant pour l'entrée *cataclysm*. Les lignes **WordNet** et Moby donnent les voisins de référence dans ces deux ressources, la ligne *initial*, les plus proches voisins présents dans le thésaurus initial « syntaxique » et la ligne *réordonnancés*, ceux dans le thésaurus après le réordonnancement. Au niveau des premiers voisins qui sont montrés ici, les « bons » voisins sont communs aux deux ressources et l'on constate globalement que les voisins présents dans le thésaurus ne sont pas aberrants, même lorsqu'ils ne font pas partie des ressources de référence. La procédure de réordonnancement permet néanmoins d'améliorer la situation en éliminant le premier voisin, non pertinent, tout en préservant les voisins pertinents. De ce fait, le rang de ces « bons » voisins augmente mécaniquement, ce qui était l'effet recherché. On notera néanmoins que le voisin *schism*, lié sémantiquement au voisin dégradé *divisionism*, conserve sa place relative.

5.3 Combinaison des approches

La méthode appliquée ici et celles proposées par Ferret (2013b) s'appuient sur des critères très différents les uns des autres. Dans le prolongement de (Curran, 2002), il apparaît *a priori* intéressant de combiner les résultats de ces différentes 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 résultat en adoptant une méthode classique de vote. Le tableau 8 donne les résultats que nous avons obtenus avec quatre de ces méthodes. Trois d'entre elles, *Borda*, *Condorcet* (Nuray & Can, 2006) et *Reciprocal Rank Fusion* (RRF, avec le paramètre $k = 60$ de (Cormack *et al.*, 2009)), s'appuient uniquement sur les rangs tandis que *CombSum*, utilisée ici avec une normalisation des valeurs de type *Zero-one* (Wu *et al.*, 2006), exploite les valeurs de similarité. Quatre thésaurus sont ainsi fusionnés : le thésaurus initial, en l'occurrence celui construit à partir des cooccurrents graphiques puisqu'il est commun à tous les travaux considérés ; le thésaurus réordonné grâce au critère de symétrie de (Ferret, 2013b) ; celui réordonné grâce aux mots composés, toujours

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
composé	+0,1	-0,1	+2,0	+0,9	+0,6
déclassement	+0,7	+0,5	+2,3	+1,3	+0,9
RRF	+0,9	+0,7	+4,2	+2,3	+1,6
Borda	+0,8	+0,7	+4,1	+2,1	+1,5
Condorcet	+0,9	+0,8	+3,4	+2,4	+1,7
CombSum	**+1,2**	**+1,0**	**+5,1**	**+2,6**	**+1,8**

TABLE 8 – Évaluation de la fusion des différentes méthodes d'amélioration, avec la référence [WM]

issu de (Ferret, 2013b) ; enfin, le thésaurus produit grâce à la détection des mauvais voisins présentée dans cet article.

Outre les résultats pour ces quatre méthodes de fusion, le tableau 8 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 comme nous l'avons déjà indiqué précédemment. 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, *Borda* est l'option la moins bonne pour toutes les mesures. *Condorcet* se comporte quant à elle de façon intéressante mais souffre d'une faiblesse au rang 1, ce qui conduit à lui préférer RRF pour un usage général.

6 Discussion et travaux liés

Comme nous l'avons vu en introduction, une première façon d'améliorer un thésaurus distributionnel est d'intervenir sur le contenu des contextes distributionnels des mots ou sur la pondération de celui-ci. Les méthodes concernées sont donc dépendantes de la représentation de ces contextes. En définissant une méthode d'amélioration indépendante de la méthode initiale de constitution du thésaurus, l'approche que nous adoptons s'affranchit d'hypothèses *a priori* sur la représentation de l'information distributionnelle adoptée par cette méthode et peut donc en principe être appliquée à tout thésaurus distributionnel, quel que soit son mode de construction.

Le problème de l'amélioration d'un thésaurus distributionnel a aussi été abordé en exploitant ses spécificités, au-delà des données distributionnelles ayant permis sa constitution initiale. Ainsi, (Ferret, 2012) sélectionne de façon non supervisée un ensemble d'exemples positifs et négatifs de mots sémantiquement similaires en s'appuyant sur l'hypothèse de la symétrie de la relation de similarité sémantique entre mots. Cette sélection permet d'entraîner un classifieur utilisé ensuite pour réordonnancer les voisins du thésaurus. (Ferret, 2013b) s'inscrit dans le même cadre mais fait appel à un principe de sélection des exemples différent utilisant le contexte de mots composés similaires pour mettre au jour la similarité de leurs constituants. Dans les deux méthodes, les exemples sont de plus sélectionnés parmi les entrées de plus forte fréquence afin de minimiser les cas d'erreur. Plus globalement, l'idée sous-jacente est de produire un ensemble d'exemples comportant le moins d'erreurs possible afin de transposer, au travers de l'entraînement d'un classifieur statistique, les performances observées pour les entrées de forte fréquente vers les entrées de fréquence plus faible.

Cette approche se heurte néanmoins à une difficulté intrinsèque : les exemples étant issus du thésaurus, l'application d'un tel classifieur aux entrées de forte fréquence a tendance à faire décroître les résultats pour ces entrées. L'effet se comprend aisément : un classifieur statistique peut difficilement dépasser les performances de son ensemble d'apprentissage. Le fait de s'appuyer fortement sur le thésaurus initial a également pour effet négatif de biaiser les améliorations obtenues en faveur des voisins relevant de la proximité sémantique par opposition à ceux relevant davantage de la similarité sémantique du fait de la prévalence nette des premiers par rapport aux seconds dans le thésaurus initial. La méthode d'amélioration que nous considérons dans cet article ne présente pas ces deux difficultés : le critère de réordonnancement des voisins

qu'elle exploite est extérieur au thésaurus distributionnel puisqu'il repose sur les occurrences des mots et leur contexte environnant et non sur une agrégation des informations liées à ces occurrences. (Claveau *et al.*, 2014) représente une autre façon de résoudre les problèmes posés par (Ferret, 2012) et (Ferret, 2013b) : l'idée est ici de représenter un thésaurus comme un graphe de voisinage distributionnel et d'exploiter les relations de réciprocité dans ce graphe, soit par l'entremise de fonctions d'agrégation, soit pour calculer un score de confiance d'une liste de voisins permettant ensuite de réordonner ces voisins. Les améliorations obtenues, données en pourcentage, sont à peu près comparables à celles observées dans notre cas pour les cooccurrents graphiques et tendent à dépasser celles associées aux cooccurrents syntaxiques. Plus globalement, il serait intéressant d'intégrer ces résultats dans les approches de fusion que nous avons déjà étudiées.

7 Conclusion et perspectives

Dans cet article, nous avons présenté une approche de réordonnancement des voisins d'un thésaurus distributionnel fondée sur une modélisation discriminante du contexte distributionnel des mots. Plus précisément, cette modélisation repose sur la construction de classifieurs permettant de différencier en contexte un mot des autres mots dans une optique de tâche de pseudo-désambiguïsation sémantique. Dans le cas des voisins sémantiques d'une entrée de thésaurus, le classifieur construit pour l'entrée permet de détecter les voisins dont les contextes d'occurrence ne sont pas jugés compatibles avec les contextes de l'entrée et qui sont donc, en vertu de l'hypothèse distributionnelle, considérés comme sémantiquement distants. Le réordonnancement des voisins s'effectue en final par le déclassement des voisins détectés comme non similaires à l'entrée. La méthode présentée a été testée pour l'anglais sur un large thésaurus de noms constitué à partir de cooccurrents syntaxiques et a montré son efficacité par une amélioration significative des résultats dans le cadre d'une évaluation intrinsèque. Nous avons montré en outre que la combinaison de cette méthode avec les méthodes présentées dans (Ferret, 2012) et (Ferret, 2013b) selon une approche de type vote permet d'exploiter de façon intéressante les complémentarités de ces différentes méthodes.

Nous envisageons d'étendre ce travail en y intégrant la notion de sens de mot, à l'instar de (Reisinger & Mooney, 2010) ou de (Huang *et al.*, 2012). En l'absence de cette différenciation, le ou les sens majoritaires d'une entrée du thésaurus dans le corpus considéré ont tendance à être représentés de façon également très majoritaire parmi les voisins de cette entrée et à en limiter la diversité sur le plan sémantique. Dans le contexte de notre travail, cette extension devrait être assez directe puisqu'elle consisterait pour l'essentiel à transformer nos classifieurs de mots en contexte en véritables classifieurs de désambiguïsation sémantique.

Références

ALEXANDRESCU A. & KIRCHHOFF K. (2007). Data-driven graph construction for semi-supervised graph-based learning in NLP. In *NAACL HLT 2007*, p. 204–211, Rochester, New York.

BARONI M., DINU G. & KRUSZEWSKI G. (2014). Don't count, predict ! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL 2014*, Baltimore, Maryland.

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

CLAVEAU V., KIJAK E. & FERRET O. (2014). Improving distributional thesauri by exploring the graph of neighbors. In *COLING 2014*, p. 709–720, Dublin, Ireland.

CORMACK G. V., CLARKE C. L. A. & BUETTCHER S. (2009). Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *SIGIR'09*, p. 758–759.

CURRAN J. (2002). Ensemble methods for automatic thesaurus extraction. In *EMNLP 2002*, p. 222–229.

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

ERK K. & PADO S. (2010). Exemplar-based models for word meaning in context. In *ACL 2010*, p. 92–97, Uppsala, Sweden.

FERRET O. (2010). Testing semantic similarity measures for extracting synonyms from a corpus. In *LREC'10*, Valletta, Malta.

FERRET O. (2012). Combining bootstrapping and feature selection for improving a distributional thesaurus. In *ECAI 2012*, p. 336–341, Montpellier, France.

FERRET O. (2013a). Identifying bad semantic neighbors for improving distributional thesauri. In *ACL 2013*, p. 561–571, Sofia, Bulgaria.

FERRET O. (2013b). Sélection non supervisée de relations sémantiques pour améliorer un thésaurus distributionnel. In *TALN 2013*, p. 48–61, Les Sables d'Olonne, France.

GALE W. A., CHURCH K. W. & YAROWSKY D. (1992). Work on statistical methods for word sense disambiguation. In *AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, p. 54–60.

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

HENESTROZA ANGUIANO E. & CANDITO M. (2012). Probabilistic lexical generalization for french dependency parsing. In *ACL 2012 Joint Workshop on Statistical Parsing and Semantic Processing of Morphologically Rich Languages*, p. 1–11, Jeju, Republic of Korea.

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

HUANG E. H., SOCHER R., MANNING C. D. & NG A. Y. (2012). Improving word representations via global context and multiple word prototypes. In *ACL'12*, p. 873–882.

KANERVA P., KRISTOFERSON J. & HOLST A. (2000). Random indexing of text samples for latent semantic analysis. In *CogSci 2000*, p. 103–6 : Lawrence Erlbaum.

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

KIELA D. & CLARK S. (2014). A systematic study of semantic vector space model parameters. In *2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*, p. 21–30, Gothenburg, Sweden.

LEE Y. K. & NG H. T. (2002). An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In *EMNLP 2002*, p. 41–48.

LIN D. (1994). PRINCIPAR : An efficient, broad-coverage, principle-based parser. In *COLING'94*, p. 42–48, Kyoto, Japan.

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

MIKOLOV T., YIH W.-T. & ZWEIG G. (2013). Linguistic regularities in continuous space word representations. In *NAACL HLT 2013*, p. 746–751, Atlanta, Georgia.

MILLER G. A. (1990). WordNet : An On-Line Lexical Database. *International Journal of Lexicography*, **3**(4).

MIN B., SHI S., GRISHMAN R. & LIN C.-Y. (2012). Ensemble semantics for large-scale unsupervised relation extraction. In *EMNLP-CoNLL 2012*, p. 1027–1037, Jeju Island, Korea.

NURAY R. & CAN F. (2006). Automatic ranking of information retrieval systems using data fusion. *Information Processing and Management*, **42**(3), 595–614.

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

PENNINGTON J., SOCHER R. & MANNING C. (2014). Glove : Global vectors for word representation. In *EMNLP 2014*, p. 1532–1543, Doha, Qatar.

REISINGER J. & MOONEY R. J. (2010). Multi-prototype vector-space models of word meaning. In *HLT-NAACL 2010*, p. 109–117, Los Angeles, California.

RIEDL M. & BIEMANN C. (2013). Scaling to largeˆ3 data : An efficient and effective method to compute distributional thesauri. In *EMNLP 2013*, p. 884–890, Seattle, Washington, USA.

VAN DE CRUYS T. (2010). *Mining for Meaning. The Extraction of Lexico-semantic Knowledge from Text*. PhD thesis, University of Groningen, The Netherlands.

WARD G. (1996). Moby thesaurus. Moby Project.

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

WU S., CRESTANI F. & BI Y. (2006). Evaluating score normalization methods in data fusion. In *AIRS'06*, p. 642–648 : Springer-Verlag.

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, Beijing, China.

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