TALN 2004, Fès, 19-21 avril 2004

Découvrir des sens de mots à partir d'un réseau de cooccurrences lexicales

Olivier Ferret

CEA – LIST/LIC2M
92265 Fontenay-aux-Roses Cedex
ferreto@zoe.cea.fr

Résumé – Abstract

Les réseaux lexico-sémantiques de type WordNet ont fait l'objet de nombreuses critiques concernant la nature des sens qu'ils distinguent ainsi que la façon dont ils caractérisent ces distinctions de sens. Cet article présente une solution possible à ces limites, solution consistant à définir les sens des mots à partir de leur usage. Plus précisément, il propose de différencier les sens d'un mot à partir d'un réseau de cooccurrences lexicales construit sur la base d'un large corpus. Cette méthode a été testée à la fois pour le français et pour l'anglais et a fait l'objet dans ce dernier cas d'une première évaluation par comparaison avec WordNet.

Lexico-semantic networks such as WordNet have been criticized a lot on the nature of the senses they distinguish as well as on the way they define these senses. In this article, we present a possible solution to overcome these limits by defining the sense of words from the way they are used. More precisely, we propose to differentiate the senses of a word from a network of lexical cooccurrences built from a large corpus. This method was tested both for French and English and for English, was evaluated through a comparison with WordNet.

Keywords – Mots Clés

Sémantique lexicale, découverte du sens des mots, réseaux lexico-sémantiques
Lexical semantics, word sense discovery, lexico-semantic networks

1 Introduction

L'intérêt de l'utilisation de ressources sémantiques en recherche ou en extraction d'information a été montré depuis quelque temps déjà au travers de travaux allant de l'expansion de requêtes (de Loupy, El-Bèze, 2002) aux systèmes de question/réponse (Pasca, Harabagiu, 2001). Ces travaux ont également mis en avant le fait qu'une telle utilisation devait être entourée de précautions : une amélioration des performances n'est observée que si la levée d'ambiguïté sur le sens des mots est réalisée avec une très bonne fiabilité. Cette observation met l'accent sur l'un des rôles premiers de la notion de ressource sémantique : définir pour chacun de ses mots un inventaire de ses sens possibles ainsi qu'une caractérisation de chacun d'entre eux. Les principales ressources sémantiques exploitables sous forme électro-

nique et présentant une large couverture sont des réseaux lexico-sémantiques du type WordNet (Miller, 1995). De par leur mode de construction, essentiellement manuel, ces réseaux ne se démarquent pas fondamentalement des dictionnaires sous forme papier. Ils s'appuient avant tout sur une formalisation et une systématisation des pratiques lexicographiques existantes. Les critiques formulées quant à leur inadéquation vis-à-vis du traitement automatique des langues, comme par exemple dans (Harabagiu *et al.*, 1999), ne sont dès lors pas surprenantes. Ces critiques portent à la fois sur la nature des sens qu'ils distinguent et sur leur caractérisation. Ces sens sont jugés à la fois trop fins et incomplets. Par ailleurs, leur caractérisation, réalisée pour l'essentiel au travers des relations de synonymie, d'hyperonymie et d'hyponymie, manque d'éléments définissant leur contexte d'usage.

Deux grandes solutions ont été explorées pour remédier à cette situation. La première d'entre elles consiste à enrichir automatiquement les réseaux de type WordNet pour y introduire les informations permettant de répondre aux critiques formulées. Différents travaux, dont récemment (Agirre, Lopez de Lacalle, 2003), se sont ainsi donnés pour objectif de regrouper des sens de WordNet pour obtenir une granularité de sens à plusieurs niveaux, donc adaptable à la tâche considérée. D'autres travaux, en particulier dans le cadre du projet eXtended WordNet (Mihalcea, Moldovan, 2001), se sont orientés vers l'extraction de relations sémantiques plus diverses à partir des définitions (les « glosses ») associées aux synsets de WordNet, ce qui permet de caractériser davantage le contexte d'usage de chacun d'entre eux.

La seconde solution consiste à extraire les sens des mots automatiquement à partir de corpus, sans utilisation des dictionnaires existants. Chaque sens est alors décrit par une liste de mots ne se limitant pas à des synonymes ou des hyperonymes. Les travaux déjà menés dans ce cadre se répartissent en trois grandes tendances. La première, illustrée par (Pantel, Lin, 2002), ne place pas la découverte des différents sens des mots au centre de ses préoccupations. Son objectif premier est en effet de rassembler les mots en classes d'équivalence et donc plutôt de former des classes de synonymes. La découverte de sens est une conséquence indirecte : la méthode de classification utilisée, Clustering by Committee, autorisant l'appartenance d'un mot à plusieurs classes, chacune d'entre elles devient *de facto* un sens de ce mot. La deuxième tendance observée, que l'on retrouve dans (Schütze, 1998), (Pedersen, Bruce, 1997) et à sa suite (Purandare, 2003), caractérise pour sa part chaque occurrence d'un mot par un ensemble de traits liés à son environnement plus ou moins proche et procède à une classification non supervisée de toutes les occurrences du mot sur la base de ces traits. Les différentes classes formées constituent autant de sens du mot. La dernière approche enfin, représentée par (Véronis, 2003), (Dorow, Widdows, 2003) et (Rapp, 2003), prend comme point de départ les cooccurrents d'un mot enregistrés à partir d'un corpus et forme les différents sens de ce mot en regroupant ses cooccurrents suivant leur similarité ou au contraire leur dissimilarité. C'est dans cette dernière perspective que se situe le travail que nous décrivons dans cet article.

2 Principes

Le point de départ de la méthode que nous présentons est un réseau de cooccurrences lexicales, c'est-à-dire un graphe dont les nœuds sont les mots constituant le vocabulaire significatif d'un corpus et les arêtes représentent les cooccurrences observées entre ces mots dans le corpus. La découverte des sens des mots est réalisée mot par mot et le traitement d'un mot ne fait intervenir que le sous-graphe rassemblant les cooccurrents de ce mot. La première étape de la méthode consiste à construire une matrice de similarité de ces cooccurrents sur la base de leurs relations dans le sous-graphe. Une méthode de classification automatique non supervisée

est alors appliquée afin de regrouper ces cooccurrents et former les différents sens du mot considéré. L'hypothèse sous-jacente à cette méthode, hypothèse qu'elle partage avec les travaux relevant de la troisième tendance dégagée dans la section précédente, est bien entendu que la connectivité au sein du sous-graphe des cooccurrents formant le sens d'un mot est plus importante que leur connectivité avec les cooccurrents définissant les autres sens de ce mot. La méthode de classification que nous utilisons est une adaptation de la méthode Shared Nearest Neighbors (SNN), exposée dans (Ertöz et al., 2001). Cette méthode présente l'avantage de déterminer automatiquement le nombre de classes, c'est-à-dire le nombre de sens dans le cas présent, et de laisser de côté les éléments les moins représentatifs des classes formées. Ce dernier point est particulièrement utile pour cette application compte tenu du taux important de « bruit » parmi les cooccurrents d'un mot.

3 Les réseaux de cooccurrence lexicale

Dans le cadre de ce travail, nous avons testé notre méthode de découverte de sens à la fois sur le français et sur l'anglais. Nous avons donc construit un réseau de cooccurrences lexicales pour ces deux langues. Celui pour le français a été constitué à partir de 24 mois du journal *Le Monde* sélectionnés entre 1990 et 1994 ; celui pour l'anglais à partir de deux ans du journal *Los Angeles Times*, issus du corpus TREC. Dans chacun des cas, la taille du corpus est d'environ 40 millions de mots. Pour les deux réseaux, le corpus initial a d'abord été prétraité afin de caractériser les textes par leurs mots les plus discriminants sur le plan thématique, en l'occurrence les noms, les verbes et les adjectifs, donnés sous forme lemmatisée. Dans le cas du français, les noms étaient à la fois des noms simples et des noms composés. Les cooccurrences ont ensuite été extraites en utilisant une fenêtre glissante selon la méthode décrite dans (Church, Hanks, 1990). Les paramètres de cette extraction ont été fixés afin de favoriser la capture de relations sémantiques et thématiques : la fenêtre était assez large (20 mots), respectait la fin des textes et l'ordre des cooccurrences n'était pas conservé. Nous avons comme Church et Hanks adopté une évaluation de l'information mutuelle comme mesure de la cohésion de chaque cooccurrence, mesure normalisée dans notre cas par l'information mutuelle maximale relative au corpus. Après filtrage des cooccurrences les moins significatives (cohésion < 0,1 et moins de 10 occurrences), nous avons obtenu un réseau d'approximativement 23 000 mots et 5,2 millions de cooccurrences pour le français et un réseau de 30 000 mots et 4,8 millions de cooccurrences pour l'anglais.

4 Algorithme de découverte des sens

4.1 Construction de la matrice de similarité entre cooccurrents

Les algorithmes de classification sont en général suffisamment paramétrables pour influer sur le nombre et l'étendue des classes formées. Mais cette adaptabilité est implicitement limitée par la mesure de similarité définie pour comparer les éléments à classer, d'où son importance. Dans le cas présent, les éléments à classer sont les cooccurrents dans le réseau de cooccurrence lexicale du mot dont on cherche à découvrir les sens. Tout en conservant le même cadre général, nous avons souhaité tester deux mesures de similarité entre cooccurrents dans la perspective d'obtenir différents niveaux de granularité quant aux sens distingués. La première mesure reprend simplement la valeur de cohésion existant dans le réseau de cooccurrence entre les cooccurrents considérés. S'il n'existe pas de relation entre eux dans le réseau, leur similarité est considérée comme nulle. Cette mesure possède l'avantage de la simplicité et de l'efficacité algorithmique mais elle est limitée par le fait que la relation de cooccurrence ne

permet pas de capturer certaines proximités entre mots. On constate ainsi expérimentalement que l'on retrouve parmi les cooccurrents d'un mot assez peu de ses synonymes reconnus[1]. On peut donc s'attendre à ce que certains sens distingués en s'appuyant sur cette mesure ne soient en fait qu'un seul et même sens.

Pour prévenir ce risque, nous avons expérimenté une mesure de similarité entre cooccurrents reposant non seulement sur une relation de cooccurrence de premier niveau mais également de deuxième niveau, cette dernière étant réputée plus stable (Schütze, 1998). La mise en œuvre de cette mesure se fait de la façon suivante : chaque cooccurrent se voit associer un vecteur de taille égale au nombre de cooccurrents du mot traité et contenant la valeur de cohésion entre ce cooccurrent et chacun des autres cooccurrents de ce mot. Comme précédemment, cette valeur est nulle s'il n'y a pas de relation dans le réseau entre deux cooccurrents. La matrice de similarité entre cooccurrents est simplement construite en appliquant la mesure cosinus entre les vecteurs de chaque couple de cooccurrents. Avec cette seconde mesure de similarité, deux cooccurrents n'ont plus nécessairement besoin d'entretenir une relation de cooccurrence directe pour être jugés proches : ils peuvent se contenter de partager un ensemble de mots avec lesquels ils entretiennent une telle relation.

4.2 Algorithme SNN (Shared Nearest Neighbors)

L'algorithme SNN (Ertöz et al., 2001) s'inscrit dans la mouvance des algorithmes ramenant le problème de la classification à celui de la détection de composantes de forte densité dans un graphe de similarité. Dans un tel graphe, chaque nœud représente un élément à classer et une arête relie deux nœuds lorsque la similarité entre les éléments qu'ils représentent est non nulle. Lorsque la matrice de similarité est symétrique, comme c'est le cas ici, le graphe obtenu est non orienté. On pourra noter que dans le cas de la découverte des sens d'un mot, le problème est à la base un problème de détection de composantes de forte densité, les sens, au sein du graphe des cooccurrents de ce mot. Il est conservé tel quel avec la première mesure de similarité mais transposé en un problème plus général de classification avec la seconde.

Dans son principe général, l'algorithme SNN comporte deux grandes étapes : la première vise à mettre en évidence les éléments les plus représentatifs de leur voisinage en masquant les relations les moins importantes du graphe de similarité. Ces éléments constituent les embryons des futures classes, formées dans un second temps en agrégeant les autres éléments à ceux sélectionnés lors de la première phase. L'algorithme SNN, considéré dans le contexte de la découverte de sens, se décompose plus précisément comme suit :

1. « éclaircissement » du graphe de similarité : pour chaque cooccurrent, seules les arêtes en direction des k (k = 15 en l'occurrence) plus proches cooccurrents sont conservées.
2. construction du graphe des plus proches voisins partagés : cette étape consiste à remplacer dans le graphe « éclairci » la valeur portée par chaque arête par le nombre de voisins directs que les deux cooccurrents reliés par l'arête ont en commun.
3. calcul de la distribution en liens forts des cooccurrents : l'objectif de cette étape est, comme lors de l'étape 1, de procéder à une sorte d'éclaircissement. Il s'agit de repérer les

[1] Constatation faite en réalisant l'intersection pour chaque mot du réseau construit à partir du *Los Angeles Times* entre ses cooccurrents et ses synonymes dans WordNet.

cooccurrents autour desquels s'organisent un ensemble d'autres cooccurrents, *i.e.* des germes de sens, mais aussi de repérer ceux qui sont visiblement sans connexion véritable avec les autres. Pour ce faire, un seuil minimum est fixé concernant le nombre de voisins partagés par deux cooccurrents, seuil au-dessus duquel on considère les deux cooccurrents comme fortement liés. On caractérise ensuite chaque cooccurrent par le nombre de liens forts qu'il possède.

4. détermination des germes de sens et élimination du bruit : les germes de sens et les cooccurrents laissés de côté sont déterminés par simple comparaison de leur nombre de liens forts par rapport à un seuil.
5. construction des sens : cette étape consiste principalement à associer aux germes se sens trouvés à l'étape précédente les cooccurrents non déjà sélectionnés comme germe de sens ou bruit pour former des classes représentant les sens du mot considéré. Pour associer un cooccurrent à un germe de sens, la force du lien qui les unit doit être supérieure à un seuil. Si un rattachement à plusieurs germes est possible, est choisi le germe avec lequel la force du lien est la plus grande. Par ailleurs, cette étape est aussi l'occasion de rassembler plusieurs germes de sens considérés comme trop proches pour former des sens distincts : le rattachement des cooccurrents fait donc également intervenir les germes de sens.
6. élargissement des sens : à l'issue des étapes précédentes, un nombre plus ou moins important de cooccurrents n'ayant pas été considérés comme du bruit se retrouvent néanmoins sans affectation à un sens. Ce nombre dépend bien entendu de la sévérité du seuil de rattachement à un germe de sens mais l'objectif étant de former des classes homogènes, celle-ci doit être nécessairement assez forte. Néanmoins, il est également intéressant que les sens puissent être décrits de la façon la plus complète et la plus précise possible. Les sens à ce stade étant caractérisés de façon plus sûre qu'à l'issue de l'étape 4, il est possible de leur rattacher des cooccurrents dont la force de lien avec leurs constituants est plus faible.

4.3 Adaptation et modalités d'application de l'algorithme SNN

Les principes de l'algorithme SNN exposés dans la section précédente doivent être précisés sur certains points quant à leur mise en œuvre. Le principal de ces points est le mode de fixation de ses différents seuils. Nous avons opté pour un mode unique s'adaptant à la distribution des valeurs observées : chaque seuil est exprimé comme un certain quantile de ces valeurs. Dans le cas du seuil de détermination des germes de sens (égal à 0,9) et de celui de définition du bruit (égal à 0,2), il s'agit d'un quantile s'appliquant au nombre de liens forts des cooccurrents. Pour le seuil définissant la notion de lien fort (égal à 0,65), celui de rattachement des coocccurrents aux germes (égal à 0,5) et celui de rattachement des cooccurrents aux sens (égal à 0,7), le quantile est appliqué directement à la force des liens entre cooccurrents dans le graphe des plus proches voisins partagés.

Au-delà des modalités de mise en œuvre des principes, nous avons également introduit des adaptations. La plus importante d'entre elles est l'ajout d'une étape entre les deux dernières. Nous avons en effet observé qu'en dépit de la possibilité, au niveau de la phase de construction des sens, de fusionner des classes par l'intermédiaire du rattachement d'un germe de sens à un autre, certains sens restent divisés en plusieurs classes. Ce phénomène est observable même en faisant initialement appel à des cooccurrences d'ordre 2 et ne peut être efficacement

traité[2] par le seul ajustement du seuil contrôlant le rattachement des cooccurrents aux germes de sens. Dans un nombre significatif de cas, le sens « divisé » se répartit entre une ou plusieurs classes ne regroupant que 3 à 4 mots et une classe de plus large ampleur. En pratique, les germes de sens de ces classes « minoritaires » n'ont pas pu être rattachés à la classe « majoritaire » alors que la plupart des cooccurrents qui leur étaient liés s'y sont rattachés. Plutôt que de définir un mécanisme spécifique pour regrouper ces classes « minoritaires » avec la classe la plus importante, nous avons choisi de laisser l'algorithme dans sa forme actuelle le faire en détruisant ces classes (taille < 6) et en remettant leurs éléments dans l'ensemble des cooccurrents non rattachés. La dernière étape de l'algorithme permet alors dans la plupart des cas de rattacher ces cooccurrents à la classe « majoritaire ». De plus, ce mécanisme permet d'obtenir une plus grande stabilité des sens formés lorsque les paramètres de l'algorithme sont modifiés.

Une seconde adaptation, d'impact plus faible, a été opérée afin de s'assurer que les cooccurrents rattachés lors de la dernière étape n'introduisent pas de bruit. Nous avons ainsi imposé que la condition de rattachement ne porte pas seulement sur la force de la relation entre le cooccurrent à rattacher et l'un des membres de la classe mais sur la force moyenne des relations entre ce cooccurrent et les éléments de cette classe.

5 Expérimentation

Nous avons appliqué notre méthode de découverte de sens aux deux réseaux de cooccurrences lexicales (LM : français ; LAT : anglais) que nous avons construits avec les valeurs de paramètres précisées dans les sections précédentes. Pour chaque réseau, nous avons testé l'utilisation initiale de cooccurrences d'ordre 1 (LM-1 et LAT-1) et d'ordre 2 (LM-2 et LAT-2). Pour l'anglais, la seconde modalité n'a été testée que sur le sous-ensemble des mots utilisés pour l'évaluation de la section 6 (LAT-2.no). Le tableau 1 synthétise les informations concernant les sens découverts dans les différentes configurations. On remarquera qu'un pourcentage significatif de mots n'ont pas sens, même avec les cooccurrences d'ordre 2. Ce sont les mots dont les cooccurrents sont faiblement liés et dont le sens est probablement mal représenté au sein de leur réseau de cooccurrences. Par ailleurs, on notera que l'utilisation des cooccurrences d'ordre 2 conduit effectivement à réduire le nombre de sens par mot.

	LM-1	**LM-2**	**LAT-1**	**LAT-1.no**	**LAT-2.no**
nombre de mots	17.261	17.261	13.414	6.177	6.177
nombre de mots avec au moins un sens	7.373 (44,4%)	7.376 (42,7%)	5.338 (39,8%)	2.584 (41.8%)	2.406 (39%)
nombre moyen de sens par mot	2,8	2,2	1,6	1,9	1,5
nombre moyen de mots décrivant un sens	16,1	16,3	18,7	20,2	18,9

Tableau 1 : Statistiques concernant les résultats de la découverte de sens

À l'instar de Véronis (2003), nous illustrerons les résultats de notre algorithme en donnant quelques uns des mots caractérisant les sens trouvés pour le mot *barrage* :

[2] C'est-à-dire sans regrouper des classes correspondant à des sens différents.

LM-1	1.1	manifestant, forces_de_l'ordre, préfecture, agriculteur, protester, incendier, calme, pierre
	1.2	conducteur, routier, véhicule, poids_lourd, camion, permis, trafic, bloquer, voiture, autoroute
	1.3	fleuve, lac, rivière, bassin, mètre_cube, crue, amont, pollution, affluent, saumon, poisson
	1.4	blessé, casque_bleu, soldat, milicien, tir, milice, convoi, évacuer, croate, milicien, combattant
LM-2	2.1	eau, mètre, lac, pluie, rivière, bassin, fleuve, site, poisson, affluent, montagne, crue, vallée
	2.2	conducteur, trafic, routier, route, camion, chauffeur, voiture, chauffeur_routier, poids_lourd
	2.3	casque_bleu, soldat, tir, convoi, milicien, blindé, milice, aéroport, blessé, incident, croate

On retrouve dans les deux cas 3 des 4 sens distingués dans (Véronis, 2003) : *barrage hydraulique* (sens 1.3 et 2.1), *barrage routier* (sens 1.2 et 2.2), *barrage frontière* (sens 1.4 et 2.3). Le sens *match de barrage* n'est pas représenté car faiblement présent au niveau des cooccurrences et de plus, au travers de certains mots ambigus, comme *division*, qui renvoient aussi à d'autres domaines que le sport. Il faut préciser que *barrage* ne comporte ici que 1104 occurrences, à comparer avec environ 7000 occurrences pour (Véronis, 2003). Cet exemple illustre également la différence de granularité des sens induite par l'utilisation des cooccurrences d'ordre 1 ou 2. Le sens 1.1, qui est assez proche du sens 1.2, les deux faisant référence à des manifestations de colère liée à une profession, disparaît ainsi lorsqu'on fait appel aux cooccurrences d'ordre 2. Nous donnons à la suite les sens pour d'autres mots en français et en anglais avec des cooccurrences d'ordre 1 :

organe (1300)	patient, transplantation, greffe, malade, thérapeutique, médical, médecine, greffer, rein
	procréation, embryon, éthique, humain, relatif, bioéthique, corps_humain, gène, cellule
	constitutionnel, consultatif, constitution, instituer, exécutif, législatif, siéger, disposition
	article, hebdomadaire, publication, rédaction, quotidien, journal, éditorial, rédacteur
mouse (563)	compatible, sofware, computer, machine, user, desktop, pc, graphics, keyboard, device
	laboratory, researcher, cell, gene, generic, human, hormone, research, scientist, rat
party (16999)	candidate, democrat, republican, gubernatorial, presidential, partisan, reapportionment
	ballroom, cocktail, champagne, guest, bash, gala, wedding, birthday, invitation, festivity
	caterer, uninvited, party-goers, black-tie, hostess, buffet, glitches, napkins, catering

6 Évaluation

La découverte de sens se heurte, comme les autres tâches de construction de ressources linguistiques, à la difficulté de l'évaluation du résultat obtenu. La voie la plus directe pour ce faire est la comparaison avec une ressource de référence que l'on considère comme proche. Dans le cas présent, les réseaux lexico-sémantiques de type WordNet s'imposent comme la ressource de référence la plus proche. Utiliser ce type de réseaux pour évaluer les sens trouvés est certes critiquable puisqu'un des objectifs d'une telle découverte est de dépasser les limites de ces réseaux. Néanmoins, compte tenu du caractère contrôlé de ces derniers, une telle évaluation apporte au moins un élément de jugement important quant à la fiabilité des sens mis en évidence. Nous avons choisi de reprendre le protocole d'évaluation défini dans (Pantel, Lin, 2002), protocole qui s'appuie sur WordNet et dont l'accord avec un jugement manuel est raisonnablement bon (88% pour Pantel et Lin). Notre évaluation ne portera donc que sur l'anglais et a été réalisée avec Wordnet 1.7.1. Ce protocole consiste à essayer de mettre en correspondance chaque sens trouvé pour un mot avec l'un de ses synsets dans WordNet et ce, au moyen d'une mesure de similarité. Il s'agit donc d'une mesure de précision. Pantel et Lin précisent qu'une mesure de rappel n'est dans le cas présent que faiblement significative : un sens découvert peut être valide et non présent dans WordNet et à l'inverse certaines distinctions de sens dans WordNet ne sont pas nécessairement souhaitables. Ils définissent néan-

moins une mesure de rappel mais destinée seulement à classer un ensemble de systèmes. Elle n'est donc pas applicable à notre seule méthode.

La mesure de similarité entre un sens et un synset utilisée pour le calcul de la précision s'appuie sur la mesure de similarité entre synsets définie par Lin :

$$sim(s1, s2) = \frac{2 \times \log P(s)}{\log P(s1) + \log P(s2)} \quad (1)$$

où *s* est le synset le plus spécifique subsumant les synset *s1* et *s2* dans la hiérarchie de WordNet et où *P(s)* représente la probabilité du synset *s* calculée à partir d'un corpus de référence, en l'occurrence le SemCor. Pour le calcul de cette mesure, nous nous avons utilisé le module Perl WordNet::Similarity v0.06 (Patwardhan, Pedersen, 2003).
La similarité entre un sens et un synset est plus précisément définie comme la moyenne des similarités entre les mots composant le sens, ou une partie de ceux-ci, et le synset. La similarité entre un mot et un synset est elle-même donnée par la plus forte des similarités entre le synset et les synsets auxquels le mot considéré appartient, celles-ci reposant sur (1). Un sens est affecté au synset qui lui est le plus similaire, à condition toutefois que la similarité entre les deux soit supérieure à un seuil (égal ici à 0,25 comme dans (Pantel, Lin, 2002)). Finalement, la précision pour un mot est donnée par le rapport entre le nombre de ses sens s'appariant avec un de ses synsets et le nombre total de ses sens.

	LAT-1.no	**LAT-2.no**
nombre de liens forts	19,4	20,8
choix optimum	56,2	63,7

Tableau 2 : Précision moyenne des sens découverts pour l'anglais par rapport à WordNet

Le tableau 2 donne le résultat de l'évaluation de notre algorithme de découverte de sens pour les mots du réseau de cooccurrences anglais qui ne sont que des noms et qui ont au moins un sens. Deux mesures sont données. Comme Pantel et Lin, nous ne prenons en compte que 4 mots de chaque sens pour l'évaluation. Mais contrairement à eux, nous n'avons pas de mesure spécifique de la proximité des mots d'un sens par rapport au mot qu'il décrit. Nous donnons donc la précision moyenne obtenue en choisissant les 4 mots d'un sens ayant le plus grand nombre de liens forts (les critères de fréquence ou de cohésion dans le réseau de cooccurrences donnent les mêmes résultats) et celle obtenue en choisissant les 4 mots d'un sens permettant d'avoir un score maximal. Nous constatons à l'évidence un écart important entre ces deux mesures : la pertinence des sens distingués est comparable à celle obtenue par Pantel et Lin lorsque le choix des 4 mots représentatifs d'un sens est optimal (Pantel et Lin obtiennent une précision de 60,8 pour un nombre de mots par sens égal à 14) mais les mots choisis pour représenter un sens dans notre cas ne sont pas fortement liés dans WordNet (selon la mesure de Lin) au mot caractérisé par ce sens. Cela ne signifie d'ailleurs pas que ces mots ne soient pas intéressants pour décrire un sens mais plus sûrement que leur lien avec lui repose sur des relations sémantiques autres que l'hyperonymie. Le meilleur résultat obtenu par Pantel et Lin sur ce point s'explique par le fait que la base de leur méthode est le regroupement de mots similaires et non la classification des cooccurrents d'un mot, lesquels ne comportent pas beaucoup de synonymes de leur mot source. Enfin, il est à noter que leur corpus de départ est beaucoup plus large (de l'ordre de 144 millions de mots) et qu'ils font appel à des moyens d'analyse plus élaborés, en l'occurrence un analyseur syntaxique.

Sans surprise, les résultats obtenus avec les cooccurrences d'ordre 1 (LAT-1.no), qui conduisent à un nombre de sens plus important, sont inférieurs à ceux obtenus avec les cooccurrences d'ordre 2 (LAT-2.no). En l'absence de rappel, il est néanmoins difficile d'en tirer une conclusion claire : il est probable que des sens se trouvent divisés de façon artificielle dans le cas de LAT-1.no mais ce phénomène peut simultanément masquer la couverture d'un plus grand ensemble de sens effectifs permise par la meilleure homogénéité des classes formées.

7 Discussion

De par sa nature, notre méthode se compare le plus directement à (Véronis, 2003) et (Dorow, Widdows, 2003). Malgré une proximité d'approche générale avec (Rapp, 2003), la distance avec ce dernier est plus grande car il ne repose pas sur la détection de composantes de forte densité dans un graphe de cooccurrence. (Véronis, 2003) et (Dorow, Widdows, 2003) ne présentant pas d'évaluation formelle, seule une comparaison qualitative est possible. Deux différences principales sont à noter avec notre travail. La première est l'utilisation directe du graphe de cooccurrence. Nous avons opté pour notre part pour une approche plus générale en travaillant au niveau d'un graphe de similarité : lorsque la similarité entre deux mots est donnée par leur relation de cooccurrence, notre situation est la même que celle des travaux cités mais nous pouvons prendre en compte dans le même cadre des relations de similarité plus générales, telles que les cooccurrences de second ordre. La seconde différence est l'utilisation d'une procédure itérative de distinction des sens. Cette procédure consiste à sélectionner à chaque étape le sens se détachant le plus clairement puis à actualiser le graphe de cooccurrence en en éliminant les constituants du sens formé, ce qui permet de faire apparaître plus distinctement les sens résiduels. Nous avons préféré quant à nous mettre l'accent sur la possibilité de réunir des sens très proches, voire identiques, artificiellement séparés par la seule utilisation de formes de surface (cf. section 4.3). Plus globalement, les deux différences pointées ont pour conséquence principale de conduire à des distinctions de sens plus fines que celles que nous mettons en évidence. Cependant, les méthodes de découverte de sens à partir de corpus ayant plutôt tendance à distinguer un trop grand nombre de sens proches, il nous a semblé plus important de favoriser la mise en évidence de sens stables et nettement délimités que de rechercher une très grande finesse dans les distinctions de sens réalisées.

8 Conclusion et perspectives

Nous avons présenté dans cet article une nouvelle méthode pour différencier et caractériser le sens des mots à partir d'un réseau de cooccurrences lexicales. Cette méthode applique un algorithme de classification non supervisé, l'algorithme SNN, aux cooccurrents des mots dont on veut différencier les sens en se fondant sur les relations que ces cooccurrents entretiennent dans le réseau. Nous en avons réalisé une première évaluation suivant le protocole défini dans (Pantel, Lin, 2002), évaluation montrant que la pertinence des sens formés est comparable à celle des sens formés par Pantel et Lin. Cette évaluation doit cependant être approfondie. Il semble en particulier nécessaire de s'appuyer sur une mesure de similarité entre synset et sens formé permettant de prendre en compte un ensemble plus vaste de relations sémantiques telles que celles implicitement présentes dans les « glosses » associées aux synsets. Par ailleurs, une évaluation au travers d'une utilisation dans une tâche telle que l'expansion de requêtes nous semble également nécessaire afin de juger de l'apport véritable de ce type de ressource par rapport à une ressource de type WordNet.

Références

AGIRRE E., LOPEZ DE LACALLE O. (2003), Clustering WordNet Word Senses, Actes de *RANLP 2003*.

CHURCH K.W., HANKS P. (1990), Word Association Norms, Mutual Information, And Lexicography, *Computational Linguistics*, Vol. 16(1), pp. 177-210.

DOROW B., WIDDOWS D. (2003), Discovering Corpus-Specific Word Senses, Actes de *EACL 2003*, pp. 79-82.

ERTÖZ L., STEINBACH M., KUMAR V. (2001), Finding Topics in Collections of Documents: A Shared Nearest Neighbor Approach, Actes de *Text Mine'01, Workshop of the 1st SIAM International Conference on Data Mining*.

HARABAGIU S, MILLER G.A., MOLDOVAN D (1999), WordNet 2 - A Morphologically and Semantically Enhanced Resource, Actes de *SIGLEX'99*, pp. 1-8.

DE LOUPY C., EL-BÈZE M. (2002), Managing Synonymy and Polysemy in a Document Retrieval, Actes de *LREC 2002 Workshop on Creating and Using Semantics for Information Retrieval*.

MILLER G.A. (1995), WordNet: A lexical Database, *Communications of the ACM.*

MIHALCEA R., MOLDOVAN D. (2001), eXtended WordNet: Progress Report, Actes de *NAACL 2001 Worshop on WordNet and Other Lexical Resources*, pp. 95-100.

PASCA M AND HARABAGIU S. (2001), The informative role of WordNet in Open-Domain Question Answering, Actes de *NAACL 2001 Worshop on WordNet and Other Lexical Resources*, pp. 138-143.

PANTEL P., LIN D. (2002), Discovering Word Senses from Text, Actes de *ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2002*, pp. 613-619.

PATWARDHAN S., PEDERSEN T. (2003), WordNet::Similarity, http://www.d.umn.edu/~tpederse/similarity.html.

PEDERSEN T., BRUCE R. (1997), Distinguishing Word Senses in Untagged Text, Actes de *EMNLP'97*, pp. 197-207.

PURANDARE A. (2003), Discriminating Among Word Senses Using Mcquitty's Similarity Analysis, Actes de *HLT-NAACL 03 - Student Research Workshop*.

RAPP R. (2003), Word Sense Discovery Based on Sense Descriptor Dissimilarity, Actes de *Machine Translation Summit IX*.

SCHÜTZE H. (1998), Automatic Word Sense Discrimination, *Computational Linguistics*, Vol. 24(1), pp. 97-123.

VERONIS J. (2003), Cartographie lexicale pour la recherche d'information, Actes de *TALN 2003*, pp. 265-274.