Étude des critères de désambiguïsation sémantique automatique : résultats sur les cooccurrences

Laurent Audibert

Etude des critères de désambiguïsation sémantique automatique : résultats sur les cooccurrences

Laurent AUDIBERT

Jeune équipe DELIC – Université de Provence
29 Av. Robert SCHUMAN - 13621 Aix-en-Provence Cedex 1
laurent.audibert@up.univ-aix.fr

Résumé – Abstract

Nous présentons dans cet article une étude sur les critères de désambiguïsation sémantique automatique basés sur les cooccurrences. L'algorithme de désambiguïsation utilisé est du type liste de décision, il sélectionne une cooccurrence unique supposée véhiculer l'information la plus fiable dans le contexte ciblé. Cette étude porte sur 60 vocables répartis, de manière égale, en trois classes grammaticales (nom, adjectif et verbe) avec une granularité fine au niveau des sens. Nous commentons les résultats obtenus par chacun des critères évalués de manière indépendante et nous nous intéressons aux particularités qui différencient les trois classes grammaticales étudiées. Cette étude s'appuie sur un corpus français étiqueté sémantiquement dans le cadre du projet SyntSem.

This paper describes a study on cooccurrence-based criteria for automatic word sense disambiguation. We use a decision-list algorithm which selects the best disambiguating cue in the target context. The algorithm is tested on 60 words equally distributed among three parts of speech (noun, adjective and verb) with a fine sense granularity. We present the results obtained by each criterion evaluated in an independent way and we discuss the characteristics which differentiate the three parts of speech studied. The study uses a French sense-tagged corpus developed in the SyntSem project.

Mots Clés – Keywords

Désambiguïsation sémantique automatique, corpus sémantiquement étiqueté, cooccurrences.

Word sense disambiguation, sense tagged corpora, cooccurrences.

1 Introduction

La désambiguïsation sémantique automatique est un enjeu important dans la plupart des applications de traitement automatique des langues : recherche d'information, traduction automatique, reconnaissance de la parole, etc. (Ide, Véronis, 1998). Cependant, les ressources nécessaires pour aborder correctement ce problème commencent à peine à être disponibles. Ceci est particulièrement vrai pour le français.

Nous avons déjà présenté les débuts d'un travail visant à rechercher et à étudier les critères de désambiguïsation sémantique automatique (Audibert, 2002). Cette étude préliminaire portait sur 7 noms avec une granularité grossière au niveau des sens (2 à 3 sens par mot). Nous étendons ici notre étude à 60 vocables répartis, de manière égale, en trois classes grammaticales (nom, adjectif et verbe) avec une granularité de sens bien plus fine (18 lexies par vocable en moyenne). Nous présentons une série de résultats sur le pouvoir désambiguïsateur de critères basés sur les cooccurrences[1] sans chercher à combiner ces critères. Nous appelons critère, un ensemble de « phénomènes » susceptibles de survenir dans le contexte d'un vocable (ex : lemme des cooccurrences).

2 Méthodologie

2.1 Corpus de travail

La première phase de notre travail est l'étiquetage de notre corpus avec le logiciel Cordial Analyseur (développé par la société Synapse Développement), qui offre une lemmatisation et un étiquetage morpho-syntaxique d'une exactitude satisfaisante (Valli, Véronis, 1999). Cette phase nous affranchit de toute ambiguïté catégorielle comme le préconise (Kilgarriff, 1997).

L'une des difficultés majeures de l'étiquetage sémantique automatique réside dans l'inadéquation des dictionnaires traditionnels (Véronis, 2001) ou dédiés (Palmer, 1998) pour cette tâche. Une autre difficulté (Gale, Church, Yarowsky, 1993) provient du manque de corpus sémantiquement étiquetés sur lesquels des méthodes d'apprentissage supervisé peuvent être entraînées. Ce manque se transforme même en absence totale pour une langue comme le français alors qu'il commence à apparaître de tels corpus pour l'anglais, notamment dans le cadre de l'action d'évaluation Senseval (Kilgarriff, Rosenzweig, 2000). Pour ces multiples raisons, notre équipe a entrepris la construction d'un dictionnaire distributionnel en se basant sur un ensemble de critères différentiels stricts (Reymond, 2001). Ce dictionnaire comporte pour l'instant la description détaillée de 20 noms, 20 verbes et 20 adjectifs totalisant plus de 53000 occurrences dans le corpus du projet SyntSem[2] (Corpus d'environ 5.5 millions de

---

[1] Nous emploierons, dans cet article, le mot « cooccurrence » dans son acception la plus large, c'est-à-dire des mots apparaissant dans le contexte, sans contrainte de fréquence, de figement ou de lien syntaxique.

[2] Le projet SyntSem, financé par l'ELRA/ELDA, vise à produire un corpus étiqueté au niveau morpho-syntaxique avec en plus un marquage syntaxique peu profond et un marquage sémantique de mots sélectionnés.

mots, composé de textes de genres variés). C'est sur ce corpus que nous réalisons l'entraînement et l'évaluation de nos algorithmes de désambiguïsation sémantique.

Les données d'apprentissage, sur lesquelles nous entraînons et évaluons nos algorithmes, sont fonction du critère étudié. Nous avons développé un outil (Audibert, 2001) qui permet de modéliser un critère et de l'appliquer au corpus pour générer les données d'apprentissage.

2.2 Vocables étudiés

NOMS					ADJECTIFS					VERBES				
Vocable	**freq**	**lex**	**H**	**2^H**	**Vocable**	**freq**	**lex**	**H**	**2^H**	**Vocable**	**freq**	**lex**	**H**	**2^H**
barrage	92	5	1,18	2,26	correct	116	5	1,81	3,50	couvrir	518	21	3,25	9,51
restauration	104	5	1,85	3,60	sain	129	10	2,45	5,46	importer	576	8	2,57	5,93
suspension	110	5	1,50	2,82	courant	168	4	0,63	1,55	parvenir	653	8	2,31	4,97
détention	112	2	0,85	1,80	régulier	181	11	2,54	5,82	exercer	698	8	1,52	2,88
lancement	138	5	0,99	1,99	frais	182	18	3,10	8,57	conclure	727	16	2,36	5,13
concentration	246	6	1,98	3,93	secondaire	195	5	1,69	3,23	arrêter	913	15	2,97	7,85
station	266	8	2,58	5,98	strict	220	9	2,23	4,69	ouvrir	919	41	3,80	13,92
vol	278	10	2,20	4,61	exceptionnel	226	3	1,45	2,73	poursuivre	978	16	2,71	6,53
organe	366	6	2,24	4,71	utile	359	9	2,39	5,23	tirer	1001	47	3,88	14,72
compagnie	412	12	1,62	3,08	vaste	368	6	2,08	4,22	conduire	1082	15	2,28	4,85
constitution	422	6	1,64	3,13	sensible	425	11	2,63	6,19	entrer	1210	38	3,65	12,55
degré	507	18	2,47	5,53	traditionnel	447	2	0,49	1,40	connaître	1635	16	2,24	4,71
observation	572	3	0,68	1,60	populaire	457	5	2,02	4,05	rendre	1985	27	2,88	7,35
passage	601	19	2,70	6,49	biologique	475	4	0,55	1,46	comprendre	2136	13	2,76	6,79
solution	880	4	0,44	1,36	clair	556	20	3,10	8,57	présenter	2140	18	2,56	5,90
économie	930	10	2,16	4,46	historique	620	3	0,67	1,59	porter	2328	59	4,01	16,07
pied	960	62	3,55	11,70	sûr	645	14	2,61	6,12	répondre	2529	9	0,99	1,99
chef	1133	11	1,47	2,77	plein	844	35	3,99	15,93	passer	2547	83	4,49	22,49
formation	1528	9	1,66	3,17	haut	1016	29	3,46	10,98	venir	3788	33	3,21	9,29
communication	1703	13	2,44	5,42	simple	1051	14	2,14	4,41	mettre	5095	140	3,65	12,55
Total	**11360**				**Total**	**8680**				**Total**	**33458**			

Tableau 1 : Les 60 vocables avec la fréquence (freq), le nombre de lexies (lex), l'entropie de la fréquence des lexies (H) et le nombre de lexies équiprobables nécessaires pour une même entropie (perplexité : 2^H)

Le Tableau 1 détaille l'ensemble des 60 vocables de notre étude. On notera la grande disparité de la fréquence de ces vocables. Le moins fréquent étant *barrage*, avec une fréquence de 92, et le plus fréquent étant le verbe *mettre*, avec une fréquence de 5095. On remarquera également que le nombre de lexies peut atteindre 140 pour le verbe *mettre*. Nous travaillons avec une granularité au niveau des sens relativement importante : environ 11 lexies par vocable en moyenne pour les noms et les adjectifs et plus de 30 pour les verbes.

Cependant, le nombre de lexies n'est pas un bon indice de la difficulté de la tâche. En effet, il est plus facile de lever l'ambiguïté d'un vocable ayant 10 lexies, mais dont la quasi-totalité des occurrences est regroupée sous une seule lexie, que de lever l'ambiguïté d'un vocable comportant 2 lexies équiprobables. L'entropie de la répartition des occurrences du vocable sur ses différentes lexies est un meilleur indicateur de la difficulté de la levée de l'ambiguïté pour ce vocable, d'où la présence de la colonne (H), pour l'entropie, et de la colonne (2^H) qui mesure la perplexité, ce qui peut s'avérer plus parlant. Ainsi, pour une entropie inchangée, si les lexies étaient équiprobables, les noms auraient une moyenne de 4 lexies par vocables, les adjectifs de plus de 5 et les verbes de pratiquement 9. Il s'agit là d'un indice qui laisse présager une plus grande difficulté pour lever l'ambiguïté des adjectifs, et encore plus des verbes, par rapport aux noms.

2.3 Définition des critères

Il existe de nombreuses sources d'information pour lever l'ambiguïté du sens des mots. Comme l'ont montré (McRoy, 1992), (Wilks, Stevenson, 1998) ou encore (Ng, Lee, 1996) toutes ces sources peuvent être utilisées simultanément pour aboutir à une meilleure désambiguïsation. Nous avons déjà présenté un inventaire non exhaustif des critères qui peuvent être étudiés (Audibert, 2002).

De nombreuses études, comme par exemple (Ng, Lee, 1996), (Mooney, 1996) ou encore (Yarowsky, 1993), montrent que les cooccurrences constituent un bon critère pour identifier le sens d'un mot. Dans cette étude, nous nous proposons d'étudier des critères élémentaires, basés sur les cooccurrences, sans chercher à les combiner. Notre objectif est de fournir des informations de référence pour l'élaboration de critères plus complexes et de répondre à des questions comme l'intérêt de la lemmatisation, l'utilité des fenêtres de mots sans distinction de position (*unordered set of surrounding words* en anglais), l'importance des mots grammaticaux ou encore les différences de comportement entre les catégories grammaticales.

2.4 Algorithme de désambiguïsation

L'algorithme de classification des lexies utilisé est du type *liste de décision* (Rivest, 1987) pour sa simplicité de mise en œuvre et son efficacité. Cette approche ne combine pas l'information de tous les attributs de la description dont on cherche à déterminer la classe, mais se focalise sur un attribut unique, supposé véhiculer l'information la plus fiable. L'algorithme ici utilisé diffère de celui de l'étude préliminaire (Audibert, 2002) basé sur une mesure de dispersion. Il s'agit d'un algorithme très proche de celui de (Golding, 1995) qui constitue une généralisation à plus de deux classes de l'algorithme de (Yarowsky, 1995).

Soit un exemple E dont nous désirons déterminer la lexie la plus probable l_E, parmi un ensemble de lexies possibles L, en se basant sur la description D de E composée d'un certain nombre d'indices $D=\{i_1 \ldots i_n\}$ générés par l'application du critère étudié sur l'exemple E. Soit A l'ensemble des indices des exemples d'apprentissage générés par l'application du critère étudié sur le corpus d'apprentissage.

La lexie choisie est déterminée en se basant sur l'indice considéré comme étant le plus fiable dans la liste de décision : $l_E = \underset{lexie \in L}{\operatorname{argmax}}\left(p(lexie/IndFia)\right)$

L'indice le plus fiable de la liste est : $IndFia = \underset{indice \in D \cap A}{\operatorname{argmax}}\left(fiabilité(indice)\right)$.

La mesure utilisée pour ordonner les indices est : $fiabilité(indice) = \underset{lexie \in L}{\max}\left(p(lexie/indice)\right)$.

Lorsque les indices de la description n'ont jamais été rencontrés dans le corpus d'apprentissage, $D \cap A$ est vide et il n'y a pas d'indice le plus fiable *IndFia*. Dans ce cas, la lexie choisie est la lexie la plus fréquente.

L'estimation des probabilités *p(lexie/indice)* se fait sur les exemples d'apprentissage. Nous utilisons une *m*-estimation (Cussens, 1993) en raison de certains dénombrements faibles et parfois nuls : $p(lexie/indice) = \frac{n_{lexie,indice} + m.p_{lexie,indice}}{n_{indice} + m}$.

- $n_{lexie,indice}$ est le nombre d'exemples d'apprentissage dont la description contient l'indice *indice* et dont la lexie du vocable étudié est *lexie* ;

- n_{indice} est le nombre d'exemples d'apprentissage dont la description contient l'indice *indice* ;
- $p_{lexie,indice}$ est une estimation a priori de la probabilité recherchée, comme nous ne connaissons pas cette estimation, nous supposons une répartition uniforme de probabilité et nous posons $p_{lexie,indice}=\frac{1}{card(L)}$;
- *m* est une constante à déterminer.

Le lissage réalisé dans (Golding, 1995) revient à fixer *m=card(L).* D'après notre expérience, poser $m=\sqrt{fréquence_{indice}}$ donne de bien meilleurs résultats.

Pour évaluer un critère sur le corpus, nous utilisons une méthode d'évaluation croisée *k* fois (*k-fold cross validation* en anglais), conformément à l'usage commun, k=10 dans notre expérience. Cette méthode est coûteuse en temps de calcul mais permet d'évaluer le critère sur la totalité du corpus.

3 Résultats de notre étude

Nous appelons **précision de l'étiquetage majoritaire** (Gale, Church, Yarowsky, 1992) la précision obtenue en étiquetant toutes les occurrences avec la lexie la plus fréquente.

Nous appelons **précision** le rapport entre le nombre d'étiquetages corrects et le nombre d'étiquetages effectués :

Précision = (Nombre d'étiquetages corrects) / (Nombre d'étiquetages effectués).

Nous appelons **gain** l'amélioration de la précision obtenue par rapport à la précision d'un étiquetage majoritaire :

Gain = (Précision(étiq.) - Précision(étiq. majoritaire)) / (1 - Précision(étiq. majoritaire)).

3.1 Evaluation de différents critères

La Figure 1 montre la précision de la désambiguïsation obtenue par huit critères. Les noms de ces critères précisent leur nature et sont de la forme [*info1*]-[*info2*]-[*info3*]. *info3* indique si le critère considère tous les mots ou seulement les mots pleins. *info2* indique si les mots considérés sont différenciés par leur position ou pas. *info1* indique si l'on regarde la forme brute des mots ou leur lemme. Le Tableau 2 permet de synthétiser les informations de la Figure 1.

On peut remarquer que tous ces critères atteignent leur meilleure précision pour de petites fenêtres allant de ±1 mot à ±4 mots. Dans tous les cas, le retrait des mots grammaticaux se traduit par une baisse significative des performances. Traiter les mots sans tenir compte de leur position par rapport au mot à désambiguïser entraîne également une baisse des performances. De plus, les performances des critères qui tiennent compte de la position des mots présentent une dynamique moins importante. Ces critères sont ainsi moins sensibles que les autres au choix de la taille du contexte. Cette robustesse constitue une raison de plus de les privilégier. Enfin, la lemmatisation n'a permis une augmentation significative des performances que pour les adjectifs.

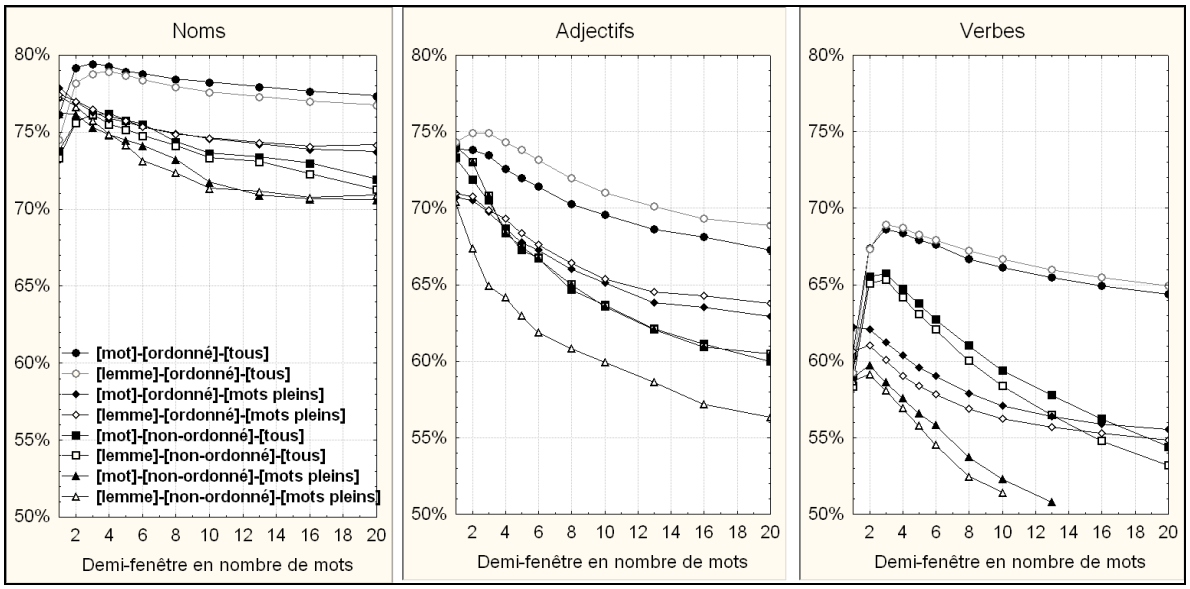

Figure 1 : Précision des 8 critères pour les 3 catégories de vocables

Critère	Catégorie Grammaticale											
	Noms				Adjectifs				Verbes			
	R	T	P %	G %	R	T	P %	G %	R	T	P %	G %
[lemme]- [ordonné] -[tous]	2	4	78,9	50,7	1	3	74,9	53,2	1	3	68,9	50,4
[mot]- [ordonné] -[tous]	1	3	79,4	51,8	3	1	73,8	51,2	2	3	68,6	49,9
[lemme]- [non-ordonné] -[tous]	8	3	76,1	44,1	2	1	74,1	51,7	4	3	65,3	44,6
[mot]- [non-ordonné] -[tous]	7	4	76,2	44,2	4	1	73,3	50,1	3	3	65,7	45,3
[lemme]- [ordonné] -[mots pleins]	3	1	77,9	48,2	6	1	70,7	45,3	5	1	62,2	39,7
[mot]- [ordonné] -[mots pleins]	4	1	77,4	47,2	5	1	71,0	45,8	6	2	61,0	37,8
[lemme]- [non-ordonné] -[mots pleins]	5	1	77,3	47,0	8	1	70,4	44,8	8	2	59,1	34,7
[mot]- [non-ordonné] -[mots pleins]	6	1	76,2	44,3	7	1	70,5	45,0	7	2	59,7	35,7
Etiquetage majoritaire	*9*	*x*	*57,3*	*0*	*9*	*x*	*46,4*	*0*	*9*	*x*	*37,4*	*0*

Tableau 2 : Meilleure précision (colonne P) et donc meilleur gain (colonne G) des 8 critères pour les 3 catégories de vocables, la colonne R indique le rang du critère (du meilleur 1 au moins bon 8) et la colonne T indique la taille du demi-contexte

3.2 Particularités des catégories grammaticales

En observant la Figure 1 on peut déjà relever trois différences de comportement entre les noms, les adjectifs et les verbes. La première différence se situe au niveau des précisions atteintes. La précision de l'étiquetage réalisé est la meilleure pour les noms, elle est moins bonne pour les adjectifs et encore moins bonne pour les verbes. Comme nous l'avions prédit dans la section 2.2, cela peut s'expliquer par le nombre moyen de lexies par vocable. Si l'on étiquette chaque occurrence d'un vocable avec sa lexie la plus fréquente (étiquetage majoritaire) on obtient une précision de 57% pour les noms, 46% pour les adjectifs et 37% pour les verbes (cf. Tableau 2). Ainsi le gain réalisé par le meilleur critère pour chaque catégorie de vocable est de 52% pour les noms, 53% pour les adjectifs et 50% pour les verbes (cf. Tableau 2). Ces chiffres montrent bien que l'algorithme et les critères de désambiguïsation utilisés fonctionnent aussi bien pour nos trois catégories de vocables.

La seconde différence est la pente de la décroissance de la précision qui est bien plus importante pour les adjectifs et les verbes que pour les noms (cf. Figure 1). Nous avons tenté

de désambiguïser chacun de nos 60 vocables en regardant une fenêtre de 4 mots pleins située à une distance de $\pm x$ mots de la cible. Les courbes de la Figure 2 montrent le gain obtenu en fonction de cette distance de x mots. On observe immédiatement que lorsque l'on s'éloigne de la cible, le gain tend vers zéro de manière beaucoup plus rapide pour les adjectifs et les verbes que pour les noms. L'information qui permet de lever l'ambiguïté d'un vocable est donc plus concentrée autour de ce vocable quand il s'agit d'un verbe ou d'un adjectif que quand il s'agit d'un nom.

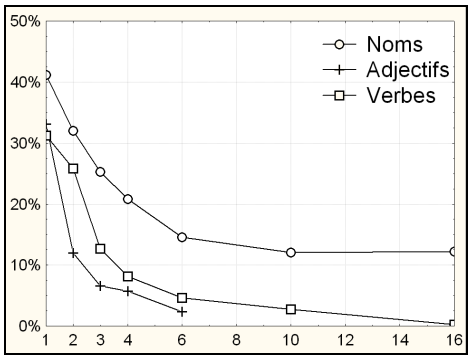

Figure 2 : Gain des 3 catégories

La troisième différence tient dans la médiocrité de la désambiguïsation des verbes pour un contexte de ±1 mot, et ce, même s'il s'agit de mots pleins. Nous apporterons une explication à ce phénomène dans la suite de cette section.

En utilisant le critère [mot]-[ordonné]-[tous], qui considère tous les mots en tenant compte de leur position et sans lemmatisation, avec un contexte de ±3 mots, nous nous sommes intéressé à la catégorie grammaticale du mot ayant servi à la levée de l'ambiguïté (Tableau 3). Cette expérience permet d'observer que les adjectifs sont les mots qui permettent le mieux de désambiguïser les noms (précision de 95.3%) et qu'ils sont utilisés dans 12.4% des cas (sur un total de 11360 cas, cf. Tableau 1). Les noms donnent de bons résultats pour désambiguïser nos trois catégories de vocables, et spécialement les adjectifs et les verbes pour lesquels ils sont utilisés dans environ 25% des cas. Les verbes à l'infinitif fonctionnent bien pour les trois catégories, ce qui n'est pas forcément le cas des autres formes verbales. Les adverbes fonctionnent bien pour désambiguïser les adjectifs. Au niveau des mots grammaticaux, on observe que les prépositions sont intéressantes pour les noms, les déterminants pour les adjectifs et les pronoms personnels pour les verbes.

Nous avons également cherché à observer comment les principales catégories grammaticales se répartissaient autour de nos trois catégories de vocables à désambiguïser. La Figure 3 montre les pourcentages des principales catégories grammaticales utilisées pour désambiguïser le vocable en fonction de la position où se trouvait le mot utilisé pour la levée de l'ambiguïté.

NOMS			**ADJECTIFS**			**VERBES**		
Catégorie	**P %**	**Utilis. %**	**Catégorie**	**P %**	**Utilis. %**	**Catégorie**	**P %**	**Utilis. %**
Adjectifs	95,3	12,4	Noms	93,4	24,2	Noms	87,6	25,3
Verbes à l'inf.	87,3	0,6	Adverbes	81,1	8,9	Conj. de sub.	82,6	2,6
Noms	82,0	32,6	Verbes à l'inf.	80,7	0,7	Verbes à l'inf.	75,6	3,7
Verbes au par.	81,1	0,3	Adjectifs	68,8	19,3	Pronoms pers.	68,9	9,8
Prépositions	79,9	19,1	Déterminants	68,5	15,6	Adjectifs	67,1	3,0
Conjonctions	75,0	3,1	Pronoms pers.	68,0	2,6	Prépositions	65,4	15,2
Déterminants	72,2	20,8	Conj. de coord.	66,1	3,4	Adverbes	63,8	4,8
Verbes conj.	63,6	1,0	Verbes restants	48,1	2,4	Verbes restants	59,1	16,5
Autres	67,6	10,0	Autres	60,6	23,1	Autres	52,6	19,2
TOTAL	**79,4**	**100,0**	**TOTAL**	**73,3**	**100,0**	**TOTAL**	**68,6**	**100,0**

Tableau 3 : Principales catégories grammaticales des mots utilisées pour lever l'ambiguïté

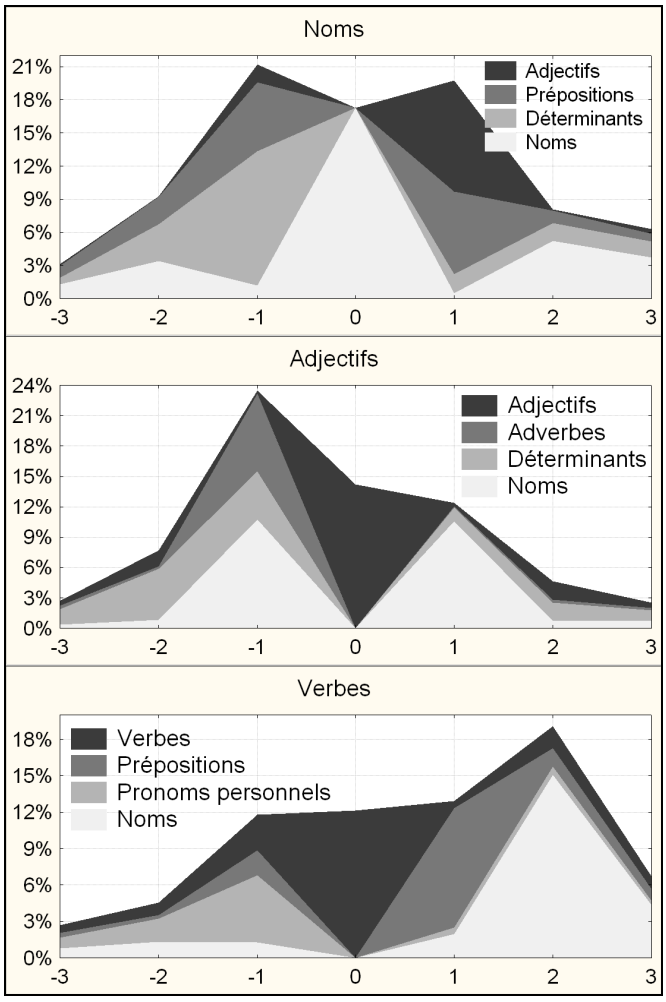

Figure 3 : Répartition, autour du mot à désambiguïser, des principales catégories grammaticales utilisées pour lever l'ambiguïté.

La position 0 est celle du vocable à désambiguïser. On peut remarquer que la forme du vocable à désambiguïser est utilisée dans 17% des cas pour les noms, 14 % pour les adjectifs et 12% pour les verbes. On pourrait penser que la précision de l'étiquetage, quand elle est basée sur la seule forme du mot à désambiguïser, doit être proche de la précision d'un étiquetage majoritaire. En fait, il n'en est rien. Le gain obtenu, lorsque c'est la forme du mot à désambiguïser qui a permis de lever l'ambiguïté, est de 45% pour les noms, 42% pour les adjectifs et 37% pour les verbes.

Les noms qui permettent de lever l'ambiguïté des adjectifs sont directement collés à cet adjectif et se trouvent indifféremment à droite ou à gauche.

La dissymétrie de la figure des verbes est très instructive. Tout d'abord elle permet d'expliquer la forme des courbes de la Figure 1. En effet, nous avions remarqué la médiocrité de la désambiguïsation des verbes pour un contexte de ±1 mot. On comprend maintenant pourquoi, contrairement aux noms et aux adjectifs où la majeure partie de l'information permettant la levée de l'ambiguïté était puisée en position –1 et +1, la majeure partie de l'information pour les verbes se trouve en position +2 et une part non négligeable se trouve en +3. Ensuite, on peut se rendre compte que la désambiguïsation des verbes se fait plus en fonction de leur objet que de leur sujet puisque la forme *sujet-verbe-complément* est la plus fréquente. Enfin, la forme de ce graphique inciterait à ne pas utiliser un contexte symétrique mais plutôt un contexte dissymétrique de la forme –2 +4 par exemple.

Sur les particularités des catégories grammaticales (Yarowsky, 1993) obtenait des résultats analogues sur l'anglais en se limitant à deux lexies par mot et en utilisant des pseudo-mots possédant deux « sens ». Ces pseudo-mots peuvent être obtenus en fusionnant deux mots quelconques, ou homographes dans une autre langue, ou encore ne se distinguant que par une seule lettre, en un seul en gardant l'information du mot d'origine. Ces pseudo-mots permettent d'obtenir directement des corpus de grande taille en s'affranchissant de la phase d'étiquetage manuel.

4 Perspectives et conclusion

Le travail présenté dans cet article sera étendu à d'autres critères pour mesurer, par exemple, l'utilité des étiquettes morpho-syntaxique ou des n-grammes. Il faudra également étudier les interactions entre ces critères de manière à pouvoir les utiliser conjointement pour aboutir à une désambiguïsation automatique plus efficace et plus robuste.

On peut remarquer que, parmi les travaux similaires déjà réalisés, peu l'ont été sur des corpus manuellement étiquetés en raison de leur rareté. Pour pallier ce problème, les chercheurs utilisent souvent des pseudo-mots qui ne comportent que deux « sens » et dont les contextes sont parfois très distincts, ce qui facilite leur désambiguïsation et biaise les résultats. Notre étude porte sur de « vrais » mots et s'appuie sur un corpus de taille suffisante manuellement étiqueté. D'autre part, l'une des difficultés de l'étiquetage sémantique automatique réside dans l'inadéquation des dictionnaires traditionnels. Pour cette raison, notre corpus a été étiqueté en utilisant les définitions d'un dictionnaire distributionnel établi sur un ensemble de critères différentiels stricts.

Cette étude a porté sur 60 vocables répartis en 20 noms, 20 adjectifs et 20 verbes. Le nombre de lexies est de pratiquement 18 en moyenne pour ces 60 mots. Les résultats obtenus sur des critères simples et sans combinaison de plusieurs critères sont encourageants. La précision moyenne obtenue atteint 79% pour les noms, 75% pour les adjectifs et 69% pour les verbes, ce qui constitue, par rapport à un étiquetage majoritaire basé sur la lexie la plus fréquente, un gain respectivement de 52%, 53% et 50%. Les meilleurs algorithmes de l'action d'évaluation Senseval (Kilgarriff, Rosenzweig, 2000) atteignent des performances de plus de 80% pour les noms, 70% pour les verbes et environ 75% pour les adjectifs. Nous sommes très proche de ces résultats, mais la comparaison est difficile car nous ne travaillons ni sur les mêmes corpus, ni sur la même langue, ni avec le même dictionnaire.

Références

Audibert L. (2001), LoX : outil polyvalent pour l'exploration de corpus annotés, Actes de *RECITAL (TALN) 2001*, pp.411-419.

Audibert L. (2002), Etude des critères de désambiguïsation sémantique automatique : présentation et premiers résultats sur les cooccurrences, Actes de *RECITAL (TALN) 2002*, pp.415-424.

Cussens J. (1993), Bayes and Pseudo-Bayes Estimates of Conditional Probabilities and Their Reliability, Actes de *European Conference on Machine Learning (Machine Learning: ECML-93)*, pp.136-152.

Gale W. A., Church K. W., Yarowsky D. (1992), Estimating upper and lower bounds on the performance of word-sense disambiguation programs, *30th Annual Meeting of the Association for Computational Linguistics*, pp.249-256.

Gale W. A., Church K. W., Yarowsky D. (1993), A method for disambiguating word senses in a large corpus, Actes de *Computers and the Humanities*, pp.415-439.

Golding A. R. (1995), A bayesian hybrid method for context-sensitive spelling correction, Actes de *Third Workshop on Very Large Corpora*, pp.39-53.

Ide N., Véronis J. (1998), Word sense disambiguation : the state of the art, *Special Issue on Word Sense Disambiguation*, Presses de l'Université de Montréal, pp.1-40.

Kilgarriff A. (1997), Evaluating word sense disambiguation programs : progress report, *SALT Workshop on Evaluation in Speech and Language Technology*, pp.114-120.

Kilgarriff A., Rosenzweig J. (2000), English SENSEVAL: Report and Results, Actes de *2nd International Conference on Language Resources and Evaluation*, pp.1239-1244.

McRoy S. (1992), Using multiple knowledge sources for word sense discrimination, Actes de *Computational Linguistics*, pp.1-30.

Mooney R. J. (1996), Comparative experiments on disambiguating word senses : an illustration of the role of bias in machine learning, *Conference on Empirical Methods in Natural Language Processing*, pp.82-91.

Ng H. T., Lee H. B. (1996), Integrating multiple knowledge sources to disambiguate word sense : an exemplar-based approach, Actes de *34th Annual Meeting of the Society for Computational Linguistics*, pp.40-47.

Palmer M. (1998), Are WordNet sense distinctions appropriate for computational lexicons ?, Actes de *SIGLEX-98, SENSEVAL*.

Reymond D. (2001), Dictionnaires distributionnels et étiquetage lexical de corpus, Actes de *RECITAL (TALN) 2001*, pp.479-488.

Rivest R. L. (1987), Learning Decision Lists, Actes de *Machine Learning*, pp.229-246.

Valli A., Véronis J. (1999), Etiquetage grammatical de corpus oraux : problèmes et perpectives, *Revue Française de Linguistique Appliquée*, Association pour le traitement informatique des langues (ASSTRIL), pp.113-133.

Véronis J. (2001), Sense tagging : does it makes sense ?, Actes de *Corpus Linguistics'2001*.

Wilks Y., Stevenson M. (1998), Word sense disambiguation using optimised combinations of knowledge sources, Actes de *COLING-ACL'98*, pp.1398-1402.

Yarowsky D. (1993), One sense per collocation, Actes de *ARPA Human Language Technology Workshop*, pp.266-271.

Yarowsky D. (1995), Decision lists for lexical ambiguity resolution : application to accent restoration in spanish and french, Actes de *33rd Annual Meeting of the Association for Computational Linguistics*, pp.88-95.