Désambiguïsation lexicale automatique : sélection automatique d'indices

Laurent AUDIBERT
Laboratoire d'Informatique de l'université Paris-Nord (LIPN)
99, avenue Jean-Baptiste Clément – 93430 Villetaneuse, France
laurent.audibert@lipn.univ-paris13.fr

Résumé. Nous exposons dans cet article une expérience de sélection automatique des indices du contexte pour la désambiguïsation lexicale automatique. Notre point de vue est qu'il est plus judicieux de privilégier la pertinence des indices du contexte plutôt que la sophistication des algorithmes de désambiguïsation utilisés. La sélection automatique des indices par le biais d'un algorithme génétique améliore significativement les résultats obtenus dans nos expériences précédentes tout en confortant des observations que nous avions faites sur la nature et la répartition des indices les plus pertinents.

Abstract. This article describes an experiment on automatic features selection for word sense disambiguation. Our point of view is that word sense disambiguation success is more dependent on the features used to represent the context in which an ambiguous word occurs than on the sophistication of the learning techniques used. Automatic features selection using a genetic algorithm improves significantly our last experiment bests results and is consistent with the observations we have made on the nature and space distribution of the most reliable features.

Mots-clés : désambiguïsation lexicale automatique, corpus sémantiquement étiqueté, cooccurrences, sélection d'indices, algorithmes génétiques.

Keywords: word sense disambiguation, sense tagged corpora, cooccurrences, features selection, genetic algorithms.

1 Introduction

La plupart des mots ont plusieurs significations. La *désambiguïsation lexicale* consiste à choisir la bonne signification d'un mot polysémique dans un contexte donné. Cette opération est utile ou indispensable pour la plupart des applications de traitement automatique des langues : recherche d'information, traduction automatique, reconnaissance de la parole, etc. (Ide & Véronis, 1998). La campagne d'évaluation trisannuel SensEval (Edmonds, 2002) atteste de l'importance de cette tâche.

La désambiguïsation lexicale s'effectue toujours en utilisant l'information présente dans le contexte du mot à désambiguïser. Cette information peut être enrichie par un certain nombre d'annotations (étiquette morphosyntaxique, lemmatisation, etc.). Il n'est cependant pas pos-

sible d'utiliser toute l'information disponible car elle est bien trop importante et bruitée. Il faut donc se focaliser sur un certain nombre d'indices. Le choix de ces indices, déterminé par ce que nous appelons des critères de désambiguïsation lexicale, est primordial et constitue un enjeu important dans le domaine de la désambiguïsation lexicale automatique (Bruce *et al.*, 1996; Ng & Zelle, 1997; Pedersen, 2001b).

Notre approche s'inscrit dans celles qui utilisent des techniques de classification supervisée sur un corpus lexicalement désambiguïsé. Dans ce type d'approche, de nombreux travaux cherchent à améliorer la précision de la désambiguïsation en améliorant les techniques de classification. Le choix des indices utilisés est généralement déterminé plus ou moins arbitrairement par la connaissance, l'expérience et l'intuition du chercheur. Peu de travaux avaient étudié systématiquement l'impact du choix des indices utilisés sur la précision de la désambiguïsation. Pour cette raison, nous avons présenté une étude des critères de désambiguïsation sémantique automatique (Audibert, 2003a) basés sur les unigrammes (*i.e.* cooccurrences de mots isolés). Nous avons complété cette étude en explorant des indices basés sur des bigrammes et des trigrammes (Audibert, 2004). Dans ces travaux, les critères étudiés étaient *homogènes* dans le sens où ils étaient constitués d'indices de même nature : par exemple, soit des lemmes, soit des étiquettes morphosyntaxiques, mais pas une combinaison des deux.

Dans le présent article, nous présentons, dans un premier temps, une petite étude comparative de différents algorithmes de classification. Nous nous intéressons ensuite à la sélection automatique des meilleurs indices du contexte pour former des critères de désambiguïsation hétérogènes sur lesquels un algorithme de classification peut s'appuyer efficacement pour effectuer de la désambiguïsation lexicale. Ce travail s'appuie toujours sur les 60 mots cibles (20 noms, 20 adjectifs et 20 verbes) des travaux précédents (Audibert, 2003a; Audibert, 2004).

2 Corpus, indices et critères

2.1 Corpus

Notre corpus de travail est composé de textes de genres variés et comporte $6\,468\,522$ mots. Il a été constitué dans le cadre du projet *SyntSem* qui vise à produire un corpus français d'amorçage étiqueté au niveau morphosyntaxique, lemmatisé et comportant un étiquetage syntaxique peu profond ainsi qu'un étiquetage lexical de 60 mots-cibles sélectionnés pour leur caractère fortement polysémique (Véronis, 1998). Ces 60 mots-cibles, qui totalisent 53796 occurrences dans le corpus, sont également répartis en 20 noms, 20 adjectifs et 20 verbes et sont détaillés dans le tableau 1.

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. Pour remédier à ce problème, l'équipe DELIC [1] a entrepris la construction d'un dictionnaire distributionnel en se basant sur un ensemble de critères différentiels stricts (Reymond, 2001). C'est ce dictionnaire qui a été utilisé pour étiqueter les occurrences des 60 mots-cibles du projet *SyntSem*. Dans ce dictionnaire, le nombre de lexies par vocable est important car il inclut les locutions figées ou composées comme *mettre sur pied*, *mettre à pied*, *pied de nez*, etc.

Un consensus semble émerger selon lequel l'étiquetage morphosyntaxique, et plus particuliè-

[1] Équipe DELIC, Université de Provence, 29 Avenue Robert SCHUMAN, 13621 Aix-en-Provence Cedex 1.

Noms				Adjectifs				Verbes			
Vocable	**freq**	**lex**	**H**	**Vocable**	**freq**	**lex**	**H**	**Vocable**	**freq**	**lex**	**H**
barrage	92	5	1,18	correct	116	5	1,81	couvrir	518	21	3,25
restauration	104	5	1,85	sain	129	10	2,45	importer	576	8	2,57
suspension	110	5	1,50	courant	168	4	0,63	parvenir	653	8	2,31
détention	112	2	0,85	régulier	181	11	2,54	exercer	698	8	1,52
lancement	138	5	0,99	frais	182	18	3,10	conclure	727	16	2,36
concentration	246	6	1,98	secondaire	195	5	1,69	arrêter	913	15	2,97
station	266	8	2,58	strict	220	9	2,23	ouvrir	919	41	3,80
vol	278	10	2,20	exceptionnel	226	3	1,45	poursuivre	978	16	2,71
organe	366	6	2,24	utile	359	9	2,39	tirer	1001	47	3,88
compagnie	412	12	1,62	vaste	368	6	2,08	conduire	1082	15	2,28
constitution	422	6	1,64	sensible	425	11	2,63	entrer	1210	38	3,65
degré	507	18	2,47	traditionnel	447	2	0,49	connaître	1635	16	2,24
observation	572	3	0,68	populaire	457	5	2,02	rendre	1985	27	2,88
passage	601	19	2,70	biologique	475	4	0,55	comprendre	2136	13	2,76
solution	880	4	0,44	clair	556	20	3,10	présenter	2140	18	2,56
économie	930	10	2,16	historique	620	3	0,67	porter	2328	59	4,01
pied	960	62	3,55	sûr	645	14	2,61	répondre	2529	9	0,99
chef	1133	11	1,47	plein	844	35	3,99	passer	2547	83	4,49
formation	1528	9	1,66	haut	1016	29	3,46	venir	3788	33	3,21
communication	1703	13	2,44	simple	1051	14	2,14	mettre	5095	140	3,65
Moyenne	568	14,2	1,9	**Moyenne**	434,4	14,1	2,3	**Moyenne**	1687,6	47,4	3,1

TAB. 1 – Fréquence moyenne des occurrences des vocables (**freq**), nombre moyen de lexies (**lex**) et entropie de la répartition des occurrences sur les lexies (**H**).

rement la levée de l'ambiguïté sur la catégorie grammaticale des vocables, n'est pas du ressort de la désambiguïsation lexicale (Kilgarriff, 1997; Ng & Zelle, 1997). Nous avons confié l'étiquetage morphosyntaxique de notre corpus au logiciel *Cordial Analyseur* (développé par la société Synapse Développement), qui offre une lemmatisation et un étiquetage morphosyntaxique d'une exactitude satisfaisante (Valli & Véronis, 1999).

jeton	**lemme**	**ems**	**smallems**	**lexie**
pouvait	pouvoir	VINDI3S	VCON	
mettre	mettre	VINF	VINF	1.12.7
fin	fin	NCFS	NCOM	
à	à	PREP	PREP	
la	le	DETDFS	DET	
pratique	pratique	NCFS	NCOM	
des	de	DETDPIG	DET	
détentions	détention	NCFP	NCOM	1

TAB. 2 – Extrait du corpus *SyntSem*

Le Tableau 2 présente un extrait du corpus *SyntSem*. Il permet de visualiser l'ensemble des étiquettes que possède un mot. C'est l'information de ces étiquettes que nous utilisons dans nos critères de désambiguïsation lexicale.

2.2 Indices et critères

Nous désignons par le terme d'*indice* une source potentielle d'information pouvant participer à la levée de l'ambiguïté d'un mot cible dont nous cherchons la bonne lexie. Un indice peut être le lemme du mot qui précède par exemple. Un *critère* est simplement la donnée d'un ensemble d'indices.

Nous avons étudié une grande variété de critères dans (Audibert, 2004). Les noms de ces critères précisent leur nature et sont de la forme `[P1|P2|P3|P4]`. Le paramètre `P1` indique si le critère considère des unigrammes (`P1=1gr`), des bigrammes (`P1=2gr`) ou des trigrammes (`P1=3gr`) ; un n-gramme étant la juxtaposition de n mots. Le paramètre `P2` indique si l'on regarde la forme brute des mots (`P2=jeton`), leur lemme (`P2=lemme`), leur étiquette morphosyntaxique (`P2=ems`) ou leur étiquette morphosyntaxique simplifiée (`P2=smallems`). Le paramètre `P3` indique si les mots considérés sont différenciés par leur position (`P3=ordonne`), différenciés suivant qu'ils appartiennent au contexte droit ou gauche (`P3=differencie`), ou non différenciés (`P3=non-ordonne`). Enfin, le paramètre `P4` indique si le critère considère tous les mots (`P4=mot`) ou seulement les mots pleins (`P4=mot-plein`). Nous qualifions ces critères de *critères homogènes* dans la mesure où l'ensemble des indices de désambiguïsation sont de la même nature puisque entièrement déterminés par l'instanciation des quatre paramètres.

3 Comparaison de différents algorithmes de désambiguïsation

Dans cette expérience, nous comparons différents algorithmes de classification supervisée en utilisant un critère assez standard constitué du lemme des mots en tenant compte de leur position (*i.e.* `[1gr|lemme|ordonne|mot]`) dans une fenêtre de ± 3 mots. Les algorithmes de classification évalués sont les suivants :

MAJ est un classifieur qui retourne toujours la lexie la plus fréquente ; nous l'utilisons comme borne inférieure à la précision de la désambiguïsation ;

PCM est un algorithme basé sur une liste de décisions, proche de celui utilisé par (Yarowsky, 1994) et détaillé dans (Audibert, 2003a) ;

NB est notre implémentation du classifieur Naïf de Bayes ;

KPPV est une implémentation élémentaire d'un classifieur du type k plus proches voisins ;

PEBLS est classifieur du type k plus proches voisins possédant une métrique bien plus sophistiqué que celle de `KPPV` ;

NBW est l'implémentation du projet Weka du classifieur Naïf de Bayes ;

C45W est l'implémentation du projet Weka du classifieur C45.

Le tableau 3 montre les résultats de cette expérience comparative. Dans toutes les expériences de désambiguïsation de cet article, toutes les occurrences reçoivent une classification. Le rappel étant égal à la précision dans ce cas, nous ne mentionnons que la précision obtenue.

Les temps d'exécution des deux algorithmes du projet Weka que nous avons utilisés (`NBW` et `C45W`) sont rédhibitoires pour nos expériences. Les raisons de ces temps d'exécution sont, ou peuvent être, la non optimisation de l'implémentation, l'utilisation du langage java et le format,

	MAJ	PCM	NB	KPPV	PEBLS	NBW	C45W
Précision	$42,9\%$	$72,3\%$	$74,5\%$	$65,5\%$	$70,9\%$	$58,2\%$	$74,6\%$
Intervalle de confiance		$\pm 0,38\%$	$\pm 0,37\%$	$\pm 0,40\%$	$\pm 0,38\%$	$\pm 0,42\%$	$\pm 0,37\%$
Temps	3s	3s	5s	26mn	2h33mn	1h47mn	35h43mn

TAB. 3 – Comparaison de la précision, avec intervalle de confiance de l'estimation à 95%, et des temps d'exécution de différents algorithmes de classification.

peu adapté au problème, de la représentation des données d'apprentissage. Le temps d'exécution du classifieur `PEBLS` est également bien trop important et est une conséquence de la complexité de la métrique utilisée.

Les classifieurs `NB` et `PCM`, nécessitent tous deux des estimations de probabilités. En raison des observations souvent rares et parfois nulles qui interviennent dans ces estimations, nous utilisons la m-estimation (Cussens, 1993) plutôt que l'estimation classique des probabilités. Cette différence explique certainement l'écart de performance des classifieurs `NB` et `NBW`.

Nous pouvons tirer deux enseignements de cette expérience. La premier est qu'il est souvent difficile et parfois préjudiciable d'utiliser un algorithme de classification comme une boîte noire (cf. la comparaison entre `NB` et `NBW`). Le second est que la complexité et la sophistication des algorithmes de classification n'apportent pas forcément un gain important pour notre tâche (cf. la comparaison entre `NB`, `PEBLS` et `C45`). Actuellement, des gains bien plus importants sont à attendre des indices fournis aux classifieurs plutôt que des classifieurs eux-mêmes.

4 Sélection automatique des indices

4.1 Méthodologie

En prenant tous les indices générés par tous les critères homogènes `[P1|P2|P3|P4]` correspondants aux différentes instanciations possibles des quatre paramètres `P1` à `P4`, et en considérant une fenêtre de ± 12 mots, nous obtenons 3 (`1gr`, `2gr` ou `3gr`) $\times 4$ (`jeton`, `lemme`, `ems` ou `smallems`) $\times 3$ (`ordonne`, `differentie` ou `non-ordonne`) $\times 2$ (`mot` ou `mot-plein`) $\times 24$ (contexte de ± 12 mots[2]) soit 1728 indices différents.

En réduisant la taille du contexte considéré, nous avons généré un deuxième jeu d'indices réduit à 888 indices. Dans ce jeu d'indices, la taille du contexte pour les critères basés sur les étiquettes `lemme` et `jeton` et composés d'unigrammes (respectivement de bigrammes et trigrammes) est de ± 6 mots (respectivement ± 8 et ± 10), et pour les critères basés sur les étiquettes `ems` et `smallems` et composés d'unigrammes (respectivement de bigrammes et trigrammes) est de ± 4 mots (respectivement ± 5 et ± 6).

La question est de savoir quels indices retenir, parmi les 1728 du premier jeu d'indices ou parmi les 888 du second, pour former un critère hétérogène efficace pour la levée de l'ambiguïté. Pour représenter un critère nous utilisons une chaîne de bits, appelée un génome, composée de 1728

[2] Le calcul est ici simplifié, en réalité, le nombre d'indices considérés sans sortir du contexte de ± 12 mots est de $12 + 1 + 12 = 25$ pour les unigrammes, $12 + 1 + 11 = 24$ pour les bigrammes et $12 + 1 + 10 = 23$ pour les trigrammes ce qui fait bien 24 en moyenne.

bits pour le premier jeu et de 888 bits pour le second. La valeur de chaque bit permet de préciser si l'indice associé est retenu ou pas. Un génome caractérise donc un critère (une sélection d'indices) hétérogène (tous les indices ne sont pas forcément de la même nature). En raison de la complexité combinatoire de notre problème d'optimisation de sélection d'indices, il n'existe pas de méthode exacte pour le résoudre en un temps raisonnable. Il faut donc se contenter de solutions approchées que nous obtenons en utilisant deux techniques classiques d'optimisation : les algorithmes gloutons et les algorithmes génétiques[3]. Le principe de l'algorithme glouton est de rechercher le meilleur indice pris individuellement, puis de chercher quel indice lui associer pour améliorer au maximum la précision, et ainsi de suite jusqu'à ne plus obtenir d'amélioration. Les algorithmes génétiques, quant à eux, tentent de mettre en œuvre le principe de la sélection naturelle (croisements et mutations) sur des populations de solutions potentielles (*i.e* des génomes) et se rapprochent de la solution au cours de générations successives.

Pour mettre en œuvre ces techniques, le corpus de départ est scindé en deux sous-corpus. Le premier sous-corpus contient 60% des exemples d'apprentissage. Il est utilisé dans un premier temps pour effectuer la sélection des indices en utilisant l'algorithme glouton ou l'algorithme génétique. Cette sélection se fait en générant une famille de génomes en suivant les règles propres à l'algorithme glouton ou génétique. L'évaluation de la performance de chacun des génomes (*i.e.* sous-ensemble d'indices) est réalisée par l'estimation de la précision obtenue par le classifieur `NB` en utilisant une méthode d'évaluation croisée k fois (avec $k = 10$) toujours sur ce même sous-corpus. Cette méthode est coûteuse en temps de calcul, mais permet l'évaluation des critères (*i.e.* des génomes) sur la totalité du sous-corpus. Une nouvelle génération de génomes est ensuite calculée en fonction de la génération précédente et des règles de l'algorithme glouton ou génétique. L'expérience est répétée tant que des génomes plus performants émergent des générations successives.

Les indices sélectionnés par le génome obtenant la meilleure performance constituent un critère hétérogène utilisé pour l'apprentissage du classifieur `NB` sur la totalité du sous-corpus contenant 60% des exemples. Le deuxième sous-corpus, qui contient 40% des exemples d'apprentissage, est enfin utilisé pour estimer la précision de désambiguïsation obtenue par le classifieur `NB` précédemment entraîné.

Cette expérience a été conduite d'un côté sur chacun des vocables indépendamment (*i.e.* un génome est sélectionné pour chacun des vocables) et d'un autre côté par catégorie grammaticale (*i.e.* un unique génome est sélectionné pour les 20 vocables d'une catégorie). L'expérience par catégorie grammaticale n'a pas été menée pour le jeu contenant 1728 indices en raisons des temps de calcul déjà de l'ordre de la dizaine de jours pour le jeu contenant 888 indices. Le tableau 4 rend compte des résultats de notre expérience.

4.2 Résultats des différentes expériences de sélection

La lecture du tableau 4 permet d'observer immédiatement que chacune des expériences de sélection automatique des indices à permis de surpasser la précision obtenue par le meilleur critère homogène identifié dans (Audibert, 2004).

Nous avons systématiquement obtenu de meilleurs résultats en sélectionnant les indices avec l'algorithme génétique plutôt qu'avec l'algorithme glouton qui est incapable de se sortir d'un

[3] Ce type d'approche n'est pas original, par exemple (Daelemans *et al.*, 2003) montrent comment obtenir une amélioration significative des performances en réalisant une optimisation simultanée des paramètres de l'algorithme d'apprentissage et de la sélection des indices en utilisant justement des algorithmes génétiques.

	Noms			Adjectifs			Verbes			Moyenne		
	P (%)	Am.	ICA	P (%)	Am.	ICA	P (%)	Am.	ICA	P (%)	Am.	ICA
Baseline (MAJ)	$57,2$			$46,3$			$37,2$			$42,9$		
Critère homogène	$81,4$	$0,0$		$75,1$	$0,0$		$72,3$	$0,0$		$74,7$	$0,0$	
Glou/Voc (1728)	$82,5$	$1,0$	$\pm 1,6$	$75,7$	$0,5$	$\pm 2,0$	$74,3$	$2,0$	$\pm 1,1$	$76,3$	$1,6$	$\pm 0,8$
Géné/Voc (1728)	$83,5$	$2,1$	$\pm 1,6$	$75,9$	$0,7$	$\pm 2,0$	$75,1$	$2,8$	$\pm 1,0$	$77,0$	$2,3$	$\pm 0,8$
Glou/Voc (888)	$82,9$	$1,5$	$\pm 1,6$	$75,7$	$0,6$	$\pm 2,0$	$75,0$	$2,7$	$\pm 1,0$	$76,8$	$2,1$	$\pm 0,8$
Géné/Voc (888)	$85,3$	$3,9$	$\pm 1,5$	$77,3$	$2,2$	$\pm 2,0$	$77,3$	$5,0$	$\pm 1,0$	$79,0$	$4,3$	$\pm 0,8$
Glou/Cat (888)	$83,7$	$2,3$	$\pm 1,6$	$75,9$	$0,7$	$\pm 2,0$	$76,8$	$4,5$	$\pm 1,0$	$78,1$	$3,4$	$\pm 0,8$
Géné/Cat (888)	$85,9$	$4,4$	$\pm 1,5$	$78,2$	$3,1$	$\pm 2,0$	$77,7$	$5,4$	$\pm 1,0$	$79,5$	$4,8$	$\pm 0,8$

TAB. 4 – Précision d'un critère hétérogène constitué par sélection automatique des indices. La ligne *Baseline (MAJ)* donne la précision obtenue par l'algorithme retournant systématiquement la lexie majoritaire. La ligne *Critère homogène* donne la précision obtenue par le meilleur critère homogène, c'est-à-dire le critère `[2gr|lemme|differencie|mot]`) avec une taille de fenêtre de ± 4 mots pour les noms et les verbes et ± 3 mots pour les adjectifs. Dans les lignes suivantes, *Glou* signifie que la technique de sélection d'indices utilisée est de type algorithme glouton tandis que *Géné* signifie que la technique utilisée est de type algorithme génétique. *Voc* signifie que la sélection d'indices est indépendante pour chacun des vocables et *Cat* qu'elle est commune aux 20 vocables de la catégorie grammaticale. *(1728)* et *(888)* précisent la taille du jeux d'indices de l'expérience. La colonne *P (%)* donne la précision obtenue en pourcentage, *Am.* l'amélioration réalisée par rapport à la précision du meilleur critère homogène (ligne *Critère homogène*) et *ICA* l'intervalle de confiance à 95% de l'amélioration réalisée (l'intervalle de confiance de la précision étant bien inférieur).

minimum local. Dans nos expériences, l'algorithme glouton sélectionne environ 20 indices. D'un autre côté, un algorithme génétique est capable, par définition, de se sortir d'un minimum local. Cependant, il ne garantit pas que tous les indices sélectionnés sont utiles et il sélectionne, dans nos expériences, environ 180 indices.

Un autre phénomène qui ressort de la lecture de ces résultats est que le jeu d'indices qui n'en contient que 888 permet d'aboutir à de meilleurs résultats que le jeu d'indices en contenant 1728. Les deux raisons de ce comportement sont la taille du corpus d'apprentissage, probablement trop faible pour une telle quantité d'indices, et le piège du surapprentissage sensible dans notre approche.

De manière surprenante, nous obtenons de meilleurs résultats en opérant la sélection sur l'ensemble d'une catégorie grammaticale plutôt que sur chacun des vocables pris individuellement. Opérer la sélection sur l'ensemble d'une catégorie grammaticale permet de limiter le phénomène de surapprentissage et d'augmenter le nombre d'exemples sur lesquels se fait la sélection. Le gain obtenu par la limitation du phénomène de surapprentissage et l'augmentation du nombre d'exemples est ici supérieur à celui obtenu par l'ajustement de la sélection des indices individuellement pour chaque vocable.

L'accord entre plusieurs annotateurs (*ITA* pour *InTer-annotator Agreement* en anglais) a été estimé à 96.4% (Audibert, 2003b) sur notre corpus. La précision moyenne de $79,5\%$ obtenue en effectuant une sélection automatique des indices permet de gagner $4,8pt$ (avec un intervalle de confiance de $\pm 0,8$) sur la précision obtenue par le meilleur critère homogène, ce qui correspond à 22% de l'écart avec la borne maximale estimée. Il s'agit donc d'une amélioration très substantielle.

4.3 Forme et répartition des indices sélectionnés

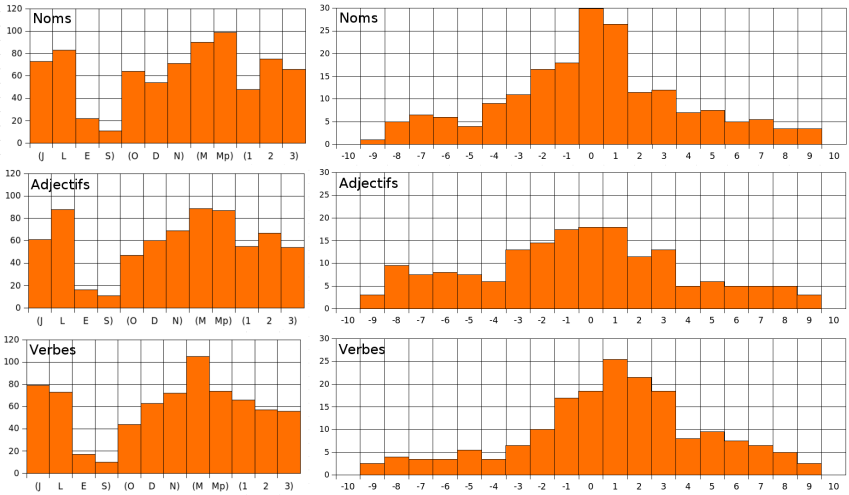

FIG. 1 – Forme et répartition spatiale des indices sélectionnés par l'algorithme génétique appliqué par catégorie grammaticale sur le jeu de 888 indices. Les graphiques de gauche montrent : les proportions d'indices constitués des étiquettes `jeton` (J), `lemme` (L), `ems` (E) ou `smalems` (S) ; les proportions d'indices différenciés par leur position (O), différenciés suivant qu'ils appartiennent au contexte droit ou gauche (D), ou non différenciés (N) ; les proportions d'indices constitués de mots sans distinction (M) ou seulement de mots pleins (Mp) ; et enfin les proportions d'indices constitués d'unigrammes (1), de bigrammes (2) ou de trigrammes (3). Les graphiques de droite montrent où se situent les indices (en prenant la position médiane pour les bigrammes et trigrammes) par rapport au mot à désambiguïser.

Nous avons cherché à en savoir plus sur les indices sélectionnés par l'algorithme génétique appliqué par catégorie grammaticale sur le jeu de 888 indices, c'est-à-dire par la sélection qui obtient les meilleurs résultat et qui correspond à la dernière ligne du tableau 4. La figure 1 résume ces observations pour chacune des catégories grammaticales.

Les graphiques de gauche permettent de remarquer que les étiquettes `jeton` et `lemme` sont bien plus utilisées que les étiquettes `ems` et `smallems` ce qui paraît logique et cohérent avec la littérature. Ils permettent également d'observer que les indices sélectionnés sont constitués en proportions comparables d'unigrammes, de bigrammes et de trigrammes. Cette observation conforte celle que nous avions faite dans (Audibert, 2004), à savoir que les bigrammes[4] et les trigrammes véhiculent une information importante qui ne se retrouve pas dans les unigrammes. Ces graphiques permettent enfin d'observer que la sélection opérée par l'algorithme génétique ne privilégie pas les indices constitués uniquement de mots pleins. Comme nous l'avions remarqué dans (Audibert, 2004), le filtrage consistant à supprimer les mots grammaticaux n'apparaît absolument pas pertinent.

[4] cf. également (Pedersen, 2001a) concernant l'utilisation des bigrammes.

Les graphiques de droite de la figure 1 montrent que la répartition spatiale des indices sélectionnés par l'algorithme génétique diffère suivant la catégorie grammaticale du mot à désambiguïser. Concernant les noms, la répartition des indices est grossièrement symétrique par rapport au mots à désambiguïser et les indices les plus proches sont privilégiés. La répartition des indices pour la désambiguïsation des adjectifs est bien plus aplatie que pour les deux autres catégories grammaticales. De plus, ce sont les adjectifs qui bénéficient le moins de l'amélioration de la précision apportée par la sélection automatique des indices : $3,1pt$ contre $4,4pt$ pour les noms et $5,4pt$ pour les verbes. Comme nous l'avions déjà observé dans (Audibert, 2004), la répartition des indices pour la désambiguïsation des verbes est fortement dissymétrique probablement parce que la désambiguïsation des verbes se fait plus en fonction de leur objet que de leur sujet, la forme sujet-verbe-complément étant la plus fréquente.

5 Conclusion et perspectives

Comme (Mohammad & Pedersen, 2004; Ng & Lee, 2002; Pedersen, 2001a), entre autres, nous pensons que les performance d'un algorithmes de désambiguïsation dépendent principalement de la qualité des indices du contexte considéré plutôt que de la sophistication des algorithmes de désambiguïsation utilisés. Dans cet article, nous avons exposé une expérience consistant à automatiser une sélection d'indices de natures différentes. Ainsi, en réalisant une sélection automatique basée sur un algorithme génétique, nous sommes parvenus à une précision de désambiguïsation, sur les 60 vocables de notre étude, de 79.5%, soit $4,8pt$ de plus que la précision obtenue par le meilleur critère homogène identifié lors de notre étude systématique précédente (Audibert, 2004).

Cette amélioration est importante, mais d'autres espoirs d'améliorations sont à attendre de l'enrichissement des indices disponibles en utilisant, par exemple :
– des indices issus de relations syntaxiques binaires (nom-nom, nom-verbe, adjectif-nom, etc.) ;
– des thésaurus ou des ontologies pour effectuer des généralisations sur les mots du contexte du mot à désambiguïser ;
– des informations sur le thème du texte.

Références

AUDIBERT L. (2003a). Etude des critères de désambiguïsation sémantique automatique : résultats sur les cooccurrences. In *10ème conférence sur le Traitement Automatique des Langues Naturelles (TALN-2003)*, p. 35–44, Batz-sur-Mer.

AUDIBERT L. (2003b). *Outils d'exploration de corpus et désambiguïsation lexicale automatique*. PhD thesis, Université de Provence.

AUDIBERT L. (2004). Word sense disambiguation criteria : a systematic study. In *20th International Conference on Computational Linguistics (COLING-2004)*, p. 910–916, Geneva.

BRUCE R., WIEBE J. & PERDERSEN T. (1996). The measure of a model. In E. BRILL & K. W. CHURCH, Eds., *1st Conference on Empirical Methods in Natural Language Processing (EMNLP-1996)*, p. 101–112, Somerset, New Jersey : Association for Computational Linguistics.

CUSSENS J. (1993). Bayes and pseudo-bayes estimates of conditional probability and their reliability. In P. B. BRAZDIL, Ed., *6th European Conference on Machine Learning (ECML-1993)*, p. 136–152, Springer-Verlag, Berlin.

DAELEMANS W., HOSTE V., MEULDER F. D. & NAUDTS B. (2003). Combined optimization of feature selection and algorithm parameter interaction in machine learning of language. In *14th European Conference on Machine Learning (ECML-2003)*, Cavtat-Dubrovnik, Croatia.

EDMONDS P. (2002). Introduction to senseval. In *ELRA Newsletter*.

IDE N. & VÉRONIS J. (1998). Word sense disambiguation : The state of the art. In *Computational Linguistics : Special Issue on Word Sense Disambiguation*, volume 24, p. 1–40.

KILGARRIFF A. (1997). Evaluating word sense disambiguation programs : Progress report. In *Speech and Language Technology (SALT-1997) Workshop on Evaluation in Speech and Language Technology*, p. 114–120, Sheffield University, United Kingdom.

MOHAMMAD S. & PEDERSEN T. (2004). Combining lexical and syntactic features for supervised word sense disambiguation. In *Proceedings of CoNLL-2004*, p. 25–32, Boston, MA, USA.

NG H. T. & LEE Y. K. (2002). An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In *7th Conference on Empirical Methods in Natural Language Processing (EMNLP-2002)*, p. 41–48, Philadelphia, Pennsylvania, USA.

NG H. T. & ZELLE J. (1997). Corpus-based approaches to semantic interpretation in natural language processing. In *Artificial Intelligence Magazine - Special Issue on Natural Language Processing*, volume 18, p. 45–64.

PALMER M. (1998). Are WordNet sense distinctions appropriate for computational lexicons. In *Association for Computational Linguistics Special Interest Group on the Lexicon (ACL-SIGLEX-1998) : SENSEVAL*, Herstmonceux, Sussex, UK.

PEDERSEN T. (2001a). A decision tree of bigrams is an accurate predictor of word sense. In *Second Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, p. 79 ?–86, Pittsburgh.

PEDERSEN T. (2001b). Machine learning with lexical features : The duluth approach to senseval-2. In *2nd International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-2)*, p. 139–142.

REYMOND D. (2001). Dictionnaires distributionnels et étiquetage lexical de corpus. In *5ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL-2002)*, volume 1, p. 479–488, Tours.

VALLI A. & VÉRONIS J. (1999). Etiquetage grammatical de corpus oraux : Problèmes et perpectives. In *Revue Française de Linguistique Appliquée*, volume IV, p. 113–133. Champs-sur-Marne : Association pour le traitement informatique des langues (ASSTRIL).

VÉRONIS J. (1998). A study of polysemy judgements and inter-annotator agreement. In *Programme and Advanced Papers of the Senseval Workshop*, Herstmonceux Castle, England.

VÉRONIS J. (2001). Sense tagging : Does it makes sense. In *Corpus Linguistics*, Lancaster, U.K.

YAROWSKY D. (1994). A comparision of corpus-based techniques for restoring accents in spanish and french text. In *2nd Annual Workshop on Very Large Text Corpora*, p. 19–32, Las Cruces.