Un étiquetage morphologique pour une résolution des ambiguïtés morphologiques en anglais

Gaëlle BIROCHEAU (1)

(1) Centre de recherches Lucien TESNIERE – Université de Franche-Comté
rue Mégevand 25000 BESANCON FRANCE
gaelle.birocheau@univ-fcomte.fr
Date de la thèse (septembre 2003)

Mots-clefs

Anglais, TAL, étiquetage morphologique, ambiguïtés dues à la morphologie, grammaires locales, contexte

Keywords

English, NLP, morphological tagging, morphological ambiguities, local grammars, context.

Résumé

Cet article expose la recherche effectuée dans le cadre de mon doctorat visant à élaborer un étiquetage morphologique de l'anglais et à désambiguïser automatiquement les ambiguïtés dues à la morphologie dans le cadre du projet LABELGRAM [9]. Nous montrons qu'il est très pertinent et efficace de travailler conjointement sur l'étiquetage et la désambiguïsation. Nous décrivons de manière précise notre contribution au système qui a consisté à mettre en place la partie anglaise. Pour ce faire, nous avons établi un dictionnaire en intention, nous avons évalué quantitativement le phénomène d'ambiguïté morphologique et établi la validité de la méthode de désambiguïsation par règles contextuelles pour l'anglais.

Abstract

The issue of this paper is to present a morphological tagging of English dedicated to resolve morphological ambiguities within the LABELGRAM project. We show it is useful to work on both fields of NLP at the same time since they make their mutual contribution to each other. To establish the base of the English part of the LABELGRAM software, we completely built up an intention dictionary of the English language, we quantitatively brought the ambiguity to the fore, and we precisely described ambiguous contexts of simple words and proved the context based disambiguation valid for English. Finally we give a tool, useful for many NLP fields, even semantic processing.

1 Introduction

Cet article expose la recherche effectuée dans le cadre du projet LABELGRAM [9]. Nous appelons ambiguïté due à la morphologie un mot présentant une graphie identique et pouvant potentiellement appartenir à plusieurs catégories grammaticales. Il suffit d'ouvrir un dictionnaire pour se rendre compte que les mots sont souvent ambigus. Or, ces ambiguïtés sont les pierres d'achoppement des techniques de TAL car elles en limitent les performances, quel que soit le domaine d'application.

L'étiquetage morphologique semble efficace en ce sens qu'il autorise une description du lexique en intention par rapport aux descriptions extensives des dictionnaires usuels. De plus, il apporte une contribution à la phase de désambiguïsation ultérieure en effectuant certaines lemmatisations. Enfin, la description des ambiguïtés morphologiques fournit des critères formels pour la définition des catégories grammaticales. La désambiguïsation morphologique basée sur le contexte local donne des résultats probants tout en évitant la lourdeur d'une analyse syntaxique complète ou d'un système devant gérer des connaissances encyclopédiques. Basée uniquement sur les catégories grammaticales, il s'avère que la désambiguïsation morphologique réduit l'ambiguïté sémantique et offre même un cadre pour sa résolution ultérieure. Il est donc très pertinent et efficace de travailler conjointement sur l'étiquetage et la désambiguïsation.

2 Etiquetage

Notre étude s'est basée sur les données de l'Oxford English Dictionary électronique [6]. Nous avons procédé à une analyse détaillée systématique de l'OED électronique, à partir de laquelle nous avons établi un mode d'étiquetage automatique, original, du lexique anglais, compatible avec les formats de l'analyseur du projet LABELGRAM.

2.1 Ambiguïtés

L'étude de l'OED montre que les ambiguïtés sont souvent plus nombreuses que ce que l'on croit (Figure 1), et relativement mal répertoriées : les dictionnaires ne s'accordent pas sur le nombre et la nature des ambiguïtés morphologiques ce qui rend leur traitement particulièrement délicat.

catégories	***tous***	***non ambigus***	***ambigus***	***% ambig / total catégorie***	***% ambig / total lexique***	***% ambig / total ambig***
adjectifs	49232	45656	3576	7,26	3,56	14,94
adverbes	7901	7205	696	8,81	0,69	2,91
conjonctions	41	25	16	39,02	0,02	0,07
interjections	685	476	209	30,51	0,21	0,87
noms	16269	6484	9785	60,15	9,75	40,89
prépositions	215	151	64	29,77	0,06	0,27
pronoms	150	112	38	25,33	0,04	0,16
verbes	25914	16366	9548	36,84	9,51	39,90
TOTAL	100407	76475	23932	23,83	23,83	100,00

Figure 1 : Importance des ambiguïtés par catégorie grammaticale

2.2 Méthodologie

Nous avons d'abord redéfini les ambiguïtés en opérant des croisements de fichiers de chacune des catégories grammaticales de l'OED. Puis, nous les avons analysées automatiquement, et sans à priori, en fonction de leur terminaison afin d'élaborer des règles de reconnaissance. Notre travail se limite au traitement des mots simples. Il a permis de valider la méthode. Nous étudierons ensuite son adaptabilité au traitement des composés. Cette méthode déjà expérimentée pour le français[1] a donné des résultats très prometteurs. Sa pertinence réside dans sa faculté à mettre à jour des règles morphologiques masquées qui peuvent être particulièrement utiles pour les apprenants étrangers, et à pouvoir prendre en compte (dans une certaine mesure) la créativité lexicale.

2.3 Résultats

Les résultats de cette analyse du lexique sont éloquents à plusieurs niveaux : d'un point de vue numérique, ils décrivent précisément l'ampleur des ambiguïtés morphologiques en langue anglaise et permettent de mettre le doigt sur les phénomènes linguistiques sous-jacents. Le mode d'étiquetage offre la possibilité de mettre en évidence les caractéristiques linguistiques et de résoudre certaines difficultés. En effet, il permet de regrouper les variantes morphologiques par prise en compte de la dérivation ou des flexions ce qui accroît les performances des systèmes TAL (Krovetz, 2000). De plus, les orientations choisies confèrent une souplesse facilitant une optimisation du traitement informatique des données. Les différentes étapes de cette démarche, précisément décrites et consignées, offrent la perspective d'une automatisation du procédé pour le traitement d'autres lexiques.

3 Désambiguïsation

La désambiguïsation est une phase préliminaire indispensable à tout traitement du langage (Rosenthal et al., 1989). De sa réussite dépendent les phases d'analyses ultérieures.

3.1 Méthodologie

Le seconde étape a consisté à élaborer un analyseur morphologique fiable, facilement adaptable et modifiable et évitant les erreurs habituelles des taggers liés à des dictionnaires.

3.1.1 Erreurs des autres taggers

L'examen de quelques résultats d'analyse d'ambiguïté par différents systèmes, qui comptent pourtant parmi les systèmes performants, montre que les ambiguïtés morphologiques posent encore beaucoup des problèmes aux analyseurs.

Résultats de l'analyseur Xerox MLTT [8] (MultiLingual Theories and Technologies)

Entrée : That very old man **works** on computer. [2]

---

[1] Thèse de Zarah El Harouchy (1998), BESANCON.

[2] Les mots en gras symbolisent les erreurs d'analyse.

L'analyseur analyse *that* comme conjonction de subordination et *works* comme un nom pluriel.

Résultats du BNC étiqueté par CLAWS

Lorsque l'on effectue un recherche en ligne sur le BNC, on constate que le système peine à désambiguïser certaines formes. Par exemple, 10,56% des occurrences de *work* ne sont pas étiquetées, 3,09% des occurrences de *surprise*, 2,61% de *program*…Le nombre d'occurrences se comptant souvent en centaines de milliers et parfois en millions, ces chiffres représentent un nombre important de données non traitées. Quant à la correction des analyses, voilà ce que nous obtenons, par exemple, pour *down* après vérification manuelle des résultats obtenus :

Catégories	PREP	ADJ	ADV	N,pro	N	V,inf	V
Nombre d'analyses correctes / Nombre d'analyses obtenues	49/50	32/50	50/50	34/50	7/50	5/7	3/42
Taux d'erreurs en pourcentage	2%	36%	0%	32%	86%	28,57%	92,8%

Figure 2 : L'analyse de *down* par l'analyseur CLAWS du BNC

3.1.2 Notre méthode

Nous avons d'abord défini deux grands types de classes grammaticales : les classes majeures et les classes mineures. Ces concepts s'entendent quantitativement. Les classes mineures sont ce que nous appelons aussi parfois les mots outils ou grammaticaux. Les classes majeures sont au nombre de 4 : les adjectifs, les adverbes, les noms et les verbes. Pour chacune des classes, nous avons défini des sous-ensembles ambigus et non ambigus.

	Total	formes non ambig.	formes ambig.	% formes ambig.
Simples				
ADJ	49232	45656	3576	7,26
ADV	7901	7205	696	8,81
N	16269	6484	9785	60,15
V	25914	16366	9548	36,84
TOTAL	**99316**	**75711**	**23605**	**23,77**
Triples				
N/V/ADJ	973	891	82	8,43
N/V/ADV	156	74	82	52,56
N/ADJ/ADV	109	27	82	75,23
V/ADJ/ADV	117	35	82	70,09
TOTAL	**1355**	**1027**	**328**	**24,21**

	Total	formes non ambig.	formes ambig.	% formes ambig.
Quadruple				
N/V/ADJ/ADV	82	82	0	0,00
TOTAL	**82**	**82**	**0**	**0,00**
Doubles				
N/V	8293	7246	1047	12,63
N/ADJ	2097	1097	1000	47,69
N/ADV	245	62	183	74,69
V/ADJ	2102	1094	1008	47,95
V/ADV	272	81	191	70,22
ADJ/ADV	479	335	144	30,06
TOTAL	**13488**	**9915**	**3573**	**26,49**

Figure 3: Les ambiguïtés dans l'OED

Puis nous avons utilisé essentiellement le British National Corpus [7] pour décrire formellement les contextes afin d'établir des règles de désambiguïsation basées sur le contexte local. Notre choix est basé sur deux faits essentiels :

- Les théories syntaxiques visant à une description complète des structures de la langue échouent à traiter l'ensemble du problème de l'ambiguïté alors qu'une simple analyse du contexte proche parvient à réduire considérablement ce problème (Leffa, 1998). On s'abstrait ainsi de la nécessité de gérer des connaissances encyclopédiques tant sur le plan théorique (linguistique) que pratique (informatique).
- Des études montrent qu'en matière de désambiguïsation, les règles nécessaires à la résolution d'ambiguïtés sont aussi efficaces qu'elles soient générées manuellement ou automatiquement (Weiss, 1973).

Nous avons donc établi les règles de désambiguïsation basée sur le contexte local en nous conformant aux critères scientifiques de systématicité, de rigueur et de reproductibilité. Nous sommes parvenue à établir des règles valables pour une catégorie entière ce qui permet de mettre en avant le principe même du fonctionnement de la catégorie et de définir ainsi, de manière formelle, la catégorie elle-même, ce qui n'est pas toujours le cas dans la linguistique traditionnelle. Parfois, il a cependant été nécessaire d'écrire des règles pour un ensemble de termes très restreints, voire pour un mot seul, notamment pour ce qui concerne les classes mineures dont le fonctionnement est loin d'être normalisé et donc normalisable. Notre procédé de désambiguïsation est uniquement basé sur les catégories grammaticales et ne recourt jamais à une information à caractère sémantique. Mais on a montré (Wilks, 1998) qu'il était possible de désambiguïser près de 92% du contenu des mots en utilisant uniquement l'information fournie pour un étiqueteur grammatical.

3.2 Résultats

Notre système n'a pas encore été implémenté mais le sera très prochainement. Cependant, il est dors et déjà possible de dire que les règles en contexte immédiat (ne prenant en considération qu'un mot à droite et à gauche de la forme ambiguë) suffisent à réduire l'ambiguïté morphologique de 35 à 50%. Ce taux augmente considérablement avec les contextes longs.
Un exemple de désambiguïsation du LABELGRAM anglais :

Entrée : That very old man works on computer.

Analyse Morphologique :

Début	Début
That	PRO,rel/PRO,dem/DET/CONJ,sub
very	ADJ/ADV
old	ADJ/N
man	N/V
works	N,plu/V,3,sing
on	PREP/ADV
computers	N,plu
.	PONCT

Désambiguïsation :

Début	.+Maj	Début
That	that	DET
very	very	ADV
old	old	ADJ
man	man	N
works	work	V
on	on	PREP
computers	computer	N
.	.	PONCT

Les étapes de la désambiguïsation :

Les règles appliquées	**Les résultats**
1/ PREP/ADV+Nplu+PONCT	PREP+Nplu+PONCT
2/ Début+PROrel/PROdem/DET/CONJsub	Début+PROdem/DET/CONJsub
3/ Recherche des verbes conjugués	
2 verbes possibles :	*man* et *works*
Recherche du nom	
2 noms possibles avant :	*old* et *man*

Vérification de l'accord nom-verbe	
Soit *old*=N est le sujet de *man*=V	*accord
Soit *man*=N est le sujet de *works*=V	accord
4/ Les composés N+N ne sont pas pris en compte	*old*=ADJ
5/ ADJ/ADV+ADJ+N	ADV+ADJ+N
6/ [ADV+ADJ+N] = SN sujet	
7/ Il n'y a qu'un verbe conjugué	that=PROdem/DET
8/ PROdem = SN and *Début+SN+SN sujet+Vconj	that=DET

⇒ Début+DET+ADV+ADJ+N+V+PREP+N+PONCT

4 Conclusion

L'ambiguïté reste encore aujourd'hui un des grands problèmes du TAL. Les progrès accomplis dans de nombreux domaines gagneraient à voir ce problème définitivement résolu, exemple : *"We describe the interaction between morphology and lexical ambiguity, and how resolving that ambiguity will lead to further improvements in performance."* (Krovetz, 2000)

Nous avons établi la description linguistique nécessaire à la mise en place de la partie anglaise du logiciel LABELGRAM.

- Nous avons construit un dictionnaire complet de la langue anglaise en intention.
- Nous avons analysé précisément le problème de l'ambiguïté.
- Nous avons décrit dans le détail les contextes susceptibles de contenir des formes ambiguës simples.
- Nous avons prouvé que la méthode de désambiguïsation basée sur le contexte pouvait s'appliquer à l'anglais.

Nous avons constitué un outil efficace et utile pour de nombreux domaines du TAL qui présente l'avantage d'être léger, adaptable et facilement modifiable.

Références

[1] Krovetz R.; Croft, W. Bruce (2000), Lexical ambiguity and information retrieval : Viewing morphology as an inference process, *Artificial Intelligence,* Vol. 118, pp 277-294.

[2] Rosenthal V., Goldblum M.C. (1989), On certain grammatical prerequisites for agrammatic behaviour in comprehension, *Journal of Neurolinguistics***,** Vol. 4, pp 179-211.

[3] Leffa V. J. (1998), Textual constraints in L2 lexical disambiguation, *System,* Vol. 26, pp 183-194.

[4] Weiss S. F. (1973), Learning to disambiguate, *Information Storage and Retrieval,* Vol. 9, pp 33-41.

[5] Wilks, Stevenson (1998), NLP in support of decision making: phrases and part-of-speech tagging in Information, *Processing & Management*, Vol. 37, pp 769-787.

[6] Robert & Collins Senior, fourth edition, 1995
Oxford English Dictionary, version électronique 1.1, 1994 : http://www.uhb.fr/scd/oed2.htm

[7] http://nlp01.cs.ul.ie - http://www.hcu.ox.ac.uk/BNC/what/gramtag.html

[8] http://www.xrce.xerox.com/competencies/content-analysis/toolhome.en.html

[9] http://tesniere.univ-fcomte.fr/projets/labelgram/labelgram.html