Etiquetage et désambigüisation par classification basés sur l'approche microsystémique

SOUMANA Ibrahim, CARDEY Sylviane & GREENFIELD Peter
Centre de Recherche en Linguistique Lucien Tesnière
Université de Franche Comté
Besançon, France

Résumé
L'étiquetage et la désambigüisation des catégories grammaticales interviennent en premier lieu dans la plupart des applications basées sur l'analyse linguistique. Cette phase est généralement liée au besoin de l'application. Dans cet article, nous présentons la phase de prétraitement en vue de l'analyse linguistique pour un Système de Question-Réponse. Elle consiste à l'identification, à l'étiquetage et à la désambigüisation des unités lexicales. La méthode utilisée établie des classes d'ambigüité qui sont traitées par l'approche microsystémique. Une évaluation sur le corpus Europarl atteint un résultat de 99,43%.

1. Introduction

Cet article présente la phase de prétraitement pour un Système de Question-Réponse pour les bases de données. Contrairement à l'approche par mots clés, l'analyse linguistique nécessite de préparer le texte de la question. Dans cet article, le prétraitement consiste à étiqueter les unités lexicales, à bien identifier les anaphores, à préserver la cohésion du groupe nominal. Une unité lexicale est soit un seul mot, une locution, un nom composé ou une

expression. La phase du prétraitement est nécessaire pour lever certaines ambigüités de la langue. Par exemple dans la phrase « les entreprises devant rembourser la banque », le mot « devant » est étiqueté comme une préposition par la plupart des outils utilisés dans la littérature alors qu'il est un verbe, ce qui change le sens de la phrase. L'identification de l'anaphore n'est pas triviale car pour résoudre le problème des anaphores, il est nécessaire de d'abord savoir les identifier. Plusieurs travaux ont été effectués dans le domaine du prétraitement du texte. Mais les travaux sont intimement liés aux besoins de l'application. Ceci limite la réutilisation des outils déjà existant dans l'état où ils sont.

L'approche adoptée pour cette tâche consiste à établir des classes d'ambigüités et de les traiter en utilisant l'approche microsystémique [1].

La section suivante présente une vue générale du lexique de la langue. C'est une vision schématique de la granularité des difficultés dont il faut tenir compte lors la conception des outils. La section 3 présente la méthodologie. L'évaluation fait l'objet de la section 4 et la dernière section se termine par la conclusion.

2. Vue générale du lexique de la langue

Du point de vue de la catégorisation automatique des unités lexicales, le lexique peut être organisé en quatre ensembles : le lexique général, les nombres, les noms propres, le lexique spécialisé.

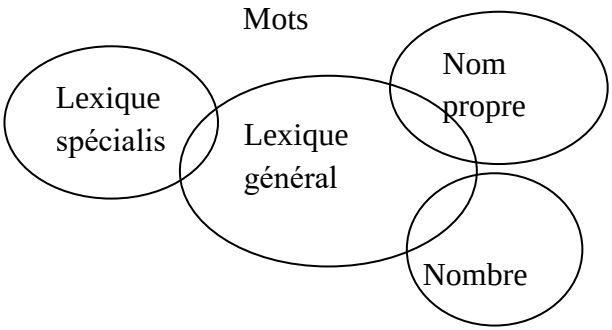

Figure 1. Vue générale du lexique

Le lexique général est l'ensemble des mots qu'un locuteur quelconque de la langue est susceptible de comprendre. Il comprend les mots simples, les mots composés, les locutions, les expressions (figées, moins figés, idiomatiques etc.). Le lexique général est fini en un instant donné et peut faire l'objet d'un dictionnaire. Les nombres servent à compter. La reconnaissance des nombres est aisée car ils sont bien formalisés. Certains nombres comme douzaine, centaine ou les chiffres (un à neuf) peuvent se retrouver dans un dictionnaire et sont donc considérés aussi comme élément du lexique général. Le nombre de noms propres est indéterminé. De ce fait, ils ne peuvent pas être recensés dans un dictionnaire de manière exhaustive. La délimitation précise des noms propres est souvent très délicate. Les noms propres partagent des éléments avec le lexique général (exemple : Orange, nom de société, couleur ou fruit). Le lexique spécialisé est le lexique des domaines spécialisés comme la biologie, la chimie ou les mathématiques par exemple. La

méthodologie doit prendre en compte les particularités du domaine.

3. Méthodologie

La méthode d'étiquetage consiste à établir des classes d'ambigüité (ambigüité de catégorie grammaticales) à partir du dictionnaire. Une classe d'ambiguïté C est par exemple :

$$C_{\text{determinant, pronom}} = \{\text{le, l', les, ce}\} \quad (1)$$

$$C_{\text{déterminant, pronom, nom}} = \{\text{tout, la}\} \quad (2)$$

$$C_{\text{pronom, verbe}} = \{\text{cela, lui, tu}\} \quad (3)$$

L'exemple (1) donne des éléments qui peuvent être *déterminant* ou *pronom*. L'exemple (2) est l'ensemble des éléments auxquels l'étiquette *déterminant*, *pronom* ou *nom* peuvent être attribué. Le mot « la »[1] est un nom quand il désigne une note de musique (sixième note de la gamme en *do* majeur). L'exemple (3) correspond à l'ensemble des mots qui peuvent avoir l'attribut *pronom* ou *verbe*. Le mot « cela » peut être un pronom démonstratif ou le verbe *celer* à la troisième personne du singulier au passé simple de l'indicatif. Le mot « lui » peut être un pronom ou le participe passé du verbe *luire*. De même le mot « tu » est

[1] http://www.cnrtl.fr/definition/la

pronom ou participe passé du verbe *taire*. Chaque classe de règles est traitée en utilisant l'approche microsystémique. Cette approche permet également de désambigüiser sans utiliser un dictionnaire en extension [2]. Le processus d'étiquetage commence par les catégories non ambiguës pour avoir un contexte local pour les mots ambigües. Le dictionnaire utilisé contient la catégorie grammaticale, le genre et nombre, le temps, la personne et le mode pour les verbes conjugués ainsi que le lemme. Il contient 683 824 entrées fléchies pour les mots simples et 108 436 entrées fléchies pour les mots composés. Les catégories du dictionnaire sont regroupées en 11 catégories auxquelles sont ajoutées trois autres catégories : la ponctuation, les symboles, les nombre et les mots inconnus.

Des nouvelles classes d'ambigüités peuvent apparaitre selon le domaine. Ces nouvelles classes peuvent être traitées par des modules liés au domaine en adoptant la même démarche.

4. Evaluation

L'évaluation du système est faite sur le corpus Europarl [3] en français. 16 959 phrases ont été extraites du corpus et étiquetées. Comme le corpus utilisé n'est pas annoté, l'évaluation est faite manuellement. 383 phrases totalisant 10 008 unités lexicales ont été évaluées. 9951 unités lexicales ont été étiquetées correctement. La précision du système est de **99,43%** sur ce corpus. Après la phase d'initialisation, la vitesse d'étiquetage est de 757 mots par seconde en moyenne sur une machine Windows XP à deux processeurs, 2 Go de RAM.

En général, les erreurs les plus fréquentes se retrouve au niveau des ambigüités qu'on peut qualifier de *liées*.

L'ambigüité *liée* survient lorsque deux mots se suivent et l'étiquette de l'un dépend de l'étiquette de l'autre. Elle peut être illustrée par l'exemple suivant :

L'entreprise la concurrence (4)

L'entreprise le concurrence (5)

Le petit (6)

Pris isolement, le mot « la » et « concurrence » sont ambigus. Si les deux sont ensemble comme en (4), si le mot « la » est un article alors le mot « concurrence » est un nom. Si le mot « la » est un pronom alors le mot « concurrence » est un verbe. L'ambigüité *liée* diffère de l'ambigüité qu'on peut qualifié de *libre* comme en (6). Le mot « le » et « petit » sont ambigus pris isolément. Mais quelque soit l'étiquette possible {adjectif ou nom} pour le mot « petit », le mot « le » a toujours la même étiquette. Donc l'ambigüité entre « le » et « petit » peut être qualifié de *libre*. Dans l'exemple 4, le mot « la » est étiqueté comme un article et le mot « concurrence » est étiqueté comme un nom. Ceci est du au fait qu'au départ la priorité est accordée à la cohésion du groupe nominal. Les mots « la » et « concurrence » s'accorde en genre et nombre. La cohésion du groupe nominal est très importante pour un Système de Question-Réponse pour des bases de données lors de la phase d'interprétation. Des règles de correction appliquées ultérieurement peuvent lever l'ambigüité. Par contre l'exemple (5) est étiqueté correctement car le mot « le » ne s'accorde pas avec le mot « concurrence » dans le groupe nominal. L'expression « le concurrence » n'est donc pas un groupe nominal.

5. Conclusion

L'étiquetage et la désambiguïsation des unités lexicales est une étape importante pour l'analyse linguistique. Dans cet article, nous avons utilisé une approche par classe d'ambigüités à partir d'un dictionnaire. Les classes d'ambigüités permettent d'avoir une couverture complète du lexique général et s'adapte également dans les autres catégories de lexique de la langue. L'approche microsystémique permet de traiter avec fiabilité les différentes classes.

Références

[1] CARDEY S., GREENFIELD P., (2005), *A Core Model of Systemic Linguistic Analysis,* In Proceedings of the International Conference RANLP-2005 Recent Advances in Natural Language Processing, Borovets, Bulgaria, 21-23 September 2005, pp. 134-138.

[2] CARDEY S., GREENFIELD P., (2003), *Disambiguating and Tagging Using Systemic Grammar*, in Actes du 8th International Symposium on Social Communication, Santiago de Cuba, January 20-24, 2003, pp. 559-564.

[3] KOEHN P. (2005). Europarl: A parallel corpus for statistical machine translation. *Conference Proceedings: the Tenth Machine Translation Summit*. Phuket, pp. 79-86.