

AMBIGUÏTES ET TRAITEMENT AUTOMATIQUE DES LANGUES. QUE PEUT FAIRE L'ORDINATEUR ?

Sylviane CARDEY[1], Peter GREENFIELD[2]

[1]Centre de recherche Lucien TESNIERE - Faculté des Lettres
30, rue Mégevand - 25030 BESANCON CEDEX - France
[2]Laboratoire d'Informatique de Besançon,
16, route de Gray - 25030 BESANCON CEDEX - France

Abstract: Ambiguities due to language in general and languages (langues) in particular which humans notice in speech or text are few in number compared with those that the computer must deal with if natural language processing is to succeed. Speech or text processing very often presents the problem of ambiguity in at least one of the domains of phonetics, phonology, morphology, syntax, semantics and pragmatics. In reality these domains have no hard frontiers. This can sometimes complicate the problem, but can sometimes help in solving it. We present techniques and methods which can contribute to solving, in written texts, certain of the problems due to ambiguity that arise at the different levels of analysis.

Keywords: ambiguity, Labelgram, natural language processing

Les ambiguïtés dues au langage en général et aux langues en particulier que l'humain (par opposition à la machine) remarque dans les énoncés oraux ou écrits sont peu nombreuses comparées à celles qui vont faire échouer les traitements par l'ordinateur.
Lors de l'analyse d'énoncés oraux ou écrits en langues naturelles, il est en effet rare de ne pas rencontrer un problème d'ambiguïté relevant au moins d'un des domaines phonétique,

ISBN: 0 08 043 438X

phonologique, morphologique, syntaxique, sémantique ou pragmatique. Ces domaines, en réalité, n'ont pas de vraies frontières, on parle d'ailleurs d'analyses morpho-phonologique, morpho-syntaxique, syntactico-sémantique et même d'analyse prama-sémantique.

Nous dirons d'un énoncé ou d'une unité qu'il(elle) est ambigu(ë) quand la machine peut l'interpréter d'au moins deux façons différentes. Nous ne parlerons que du traitement des ambiguïtés rencontrées à l'écrit.

La première étape en analyse automatique d'un énoncé, bien souvent, consiste à découper cet énoncé en unités. Cette étape n'est pas aussi simple que l'on pourrait le penser. Il est d'usage aujourd'hui de dire, pour simplifier la tâche de la machine, qu'une unité est une chaîne comprise entre blanc(s) et/ou séparateur(s). Aussi un composé n'est pas forcément la même chose pour une machine que pour un humain ; il suffit pour s'en persuader de considérer les unités suivantes :

porte, portefeuille, porte-monnaie, porte-à-faux, porte à porte, œil d'une porte, il **porte la main à.**

Pour notre exemple, seule la première unité sera simple pour l'humain alors que pour la machine les deux ou quatre premières unités seront des unités simples. En ce qui concerne les composés, grâce à nos connaissances du monde et de la langue, nous pouvons les repérer facilement ; pour la machine il en est tout autrement. Même si la machine a toutes les unités composées en mémoire, il reste encore des problèmes difficiles à résoudre. Comment une machine saura-t-elle que dans "en bas de la côte", nous avons trois unités alors que dans "en bas de soie", il y en a quatre ?

Supposons que cette étape de segmentation de l'énoncé soit possible, ce qui est, en très bonne voie. Que faire des unités polycatégorielles du genre

"ferme" qui peut être un adjectif, un nom, un verbe conjugué, un adverbe ?

Une des solutions est tout d'abord de créer des dictionnaires de mots simples, de composés ou de locutions. Les dictionnaires actuels fonctionnent à partir de la liste des mots de la langue traitée et d'informations, entre autres, sur leur(s) catégorie(s) grammaticale(s) sans lever les ambiguïtés dues à la polycatégorie. Il existe d'autres solutions qui s'éloignent quelque peu des habitudes linguistiques et qui sont à base de méthodes tout à fait empiriques. Par exemple, un de nos dictionnaires analyse la terminaison des mots simples du français et donne ainsi la ou les catégorie(s) grammaticale(s) à laquelle (auxquelles) ils peuvent appartenir. Il est également possible, à partir de ce dictionnaire "en intention" de reconstruire un dictionnaire "en extension" qui listera les mots simples de la langue française (Cardey, *et al.*, 1997). Ainsi voit-on apparaître des méthodes pour le traitement automatique des langues qui n'ont plus rien à voir avec les analyses linguistiques traditionnelles. Cependant, nous pensons que pour pouvoir s'éloigner des méthodes linguistiques traditionnelles, il faut d'abord être au courant de ces dernières. Bien souvent les méthodes qui donnent des résultats en linguistique computationnelle prennent appui sur la linguistique traditionnelle. Si l'on ne connaît pas les

ISBN: 0 08 043 438X

méthodes d'analyse en morpho-syntaxe ou syntaxe, il semble difficile de réussir à créer un système traitant des ambiguïtés dues à la morphologie et/ou à la syntaxe.

Après la création de ces dictionnaires, il reste à trouver une solution au problème des unités qui changent de catégorie grammaticale selon le contexte où elles apparaissent ; il existe différentes approches.
Une approche assez fiable, à notre avis, consiste à créer un système organisé de règles qui examinent le contexte immédiat de chacune des unités posant problème. Un travail en ce sens (El Harouchy, 1997) a été réalisé à partir du dictionnaire de terminaisons dont nous avons parlé, en effectuant un regroupement des mots polycatégoriels sous la forme de 28 ensembles en fonction des catégories grammaticales auxquelles ces mots peuvent appartenir ; aussi, on peut obtenir à partir de notre système informatisé "Labelgram" :

la méchante rigole car la petite est malade

Mot forme	Catégorie	Dict. réf.	Proc n°	Règle réf.
la	Art.	2.12/	5	45/
méchante	Nom	41.2/	28	339/
rigole	Verbe conj.	360.4/	8	79/
car	Conj.	47.5/	10	144/
la	Art.	2.12/	5	45/
petit	Nom	281.4/	28	339/
est	Verbe conj.	pre_dict	8	74/
malade	Adj.	13.1/	28	346a/

D'autres approches consistent à donner à chaque unité qui peut poser problème un ensemble d'informations aux niveaux lexicale, morphologique, syntaxique, sémantique. Cette méthode, non entièrement nouvelle, est très utilisée actuellement et porte des appellations différentes. Nous l'appellerons LeSSI (Lexique, Syntaxe, Sémantique Information).
Par exemple, une méthode a été élaborée pour permettre de lever les ambiguïtés au niveau du verbe dans un contexte de traduction automatique français-coréen, coréen-français (Hong, 1997).
Aussi, nous trouvons, dans notre dictionnaire des verbes français, les descriptions de "griller" ("griller1" pour "oléagineux" et "griller2" pour "autre que oléagineux") :

(griller,1,T,'C',nomP(X,[oléa]),FVX) et (griller,2,T,'C',nomA(X,[+com,+ins]),FVX)

où chaque symbole représente une contrainte pour la traduction.
Dans la phase de génération, le système réalise la construction-cible en partant de la description du verbe-cible et en respectant les informations enregistrées à l'issue de l'analyse de la construction-source et, s'il y en a, les règles de transfert. La procédure se termine par des transformations nécessaires comme par exemple, la pronominalisation, les ajustements morphologiques et syntaxiques.

ISBN: 0 08 043 438X

Ainsi que nous le voyons les phénomènes de polycatégorie et de polysémie sont autant de difficultés que la machine va devoir affronter.
Les méthodes probabilistes peuvent parfois donner d'assez bons résultats mais le linguiste se sent souvent sur un terrain mouvant lors de leur utilisation.
Pour terminer, donnons quelques exemples d'ambiguïtés que la machine actuellement ne parvient qu'extrêmement rarement à résoudre.
Concernant l'analyse syntaxique qui réunit les unités trouvées, lors de la phase de segmentation de l'énoncé, en unités fonctionnelles, elle doit répondre à des questions telles que : qu'est-ce qu'une phrase ? ("Elle ? un accident ! non !" ; A-t-on une, deux ou trois phrase(s) ?). Comment lever l'ambiguïté dans "Hélène a filé une toge à Paris" ?
D'autres phénomènes syntaxico-sémantiques tels que les problèmes d'accord ne sont pas simples à résoudre. Les correcteurs orthographiques avancent très lentement dans ce domaine. Une étude (Cardey, 1996) nous a permis d'établir une structure "globale" à partir de laquelle certains accords du participe passé en français employé avec avoir peuvent être automatiquement réalisés.

Structure globale :
P1(,)(adv)P2(adv)(l')(P3)(adv)(P4)(adv)avoir(adj)(P3)(P4)(adv)p.p(adv)(prép)(P5)inf(P5)

Chaque symbole représente une contrainte optionnelle si elle est entre parenthèses et obligatoire sinon.
Cependant, comment l'ordinateur pourra-t-il faire l'accord dans "la chanson que j'ai entendu? fredonner", "la fille que j'ai entendu? fredonner" ?
Saura-t-on donner une démarche, un jour, à l'ordinateur pour qu'il puisse résoudre les problèmes dus aux ambiguïtés relevant du contexte situationnel (en français : "être au violon", en anglais : "she is not laughing because he is silly") ?

Pour conclure, nous dirons que, pour le problème de l'ambiguïté, l'étape indispensable consiste en la création de dictionnaires électroniques de mots (au sens large) et de règles apportant des informations au niveau de la morphologie, de la syntaxe et de la sémantique.

REFERENCES

Cardey, S. (1996) La correction automatique : le cas du participe passé français. *In "Proceedings of the 16th European Conference on Grammar and Lexicon of Romance Languages", Munich, 19-21 septembre 1996* (à paraître).

Cardey, S., El Harouchy, Z., Greenfield, P.G. (1997) La forme des mots nous renseigne-t-elle sur leur nature ?. *In Actes des 5èmes journées scientifiques, Réseau Lexicologie, Terminologie, Traduction, LA MEMOIRE DES MOTS*, Tunis, 22-23 septembre 1997 (à paraître dans la revue de l'Association Tunisienne de Linguistique et dans la revue Meta (Journal des traducteurs)).

El Harouchy, Z (1997) Dictionnaire et grammaire pour le traitement automatique des ambiguïtés morphologiques des mots simples en français. Thèse, Université de Franche-Comté, 1997.

Hong, M. (1997) Dictionnaire automatique coréen-français. *In Actes du Colloque International FRACTAL 1997. BULAG 1996-1997- numéro Hors-Série.* 215-225.

ISBN: 0 08 043 438X