METHODES D'APPRENTISSAGE POUR L'ANALYSE AUTOMATIQUE MORPHOSYNTAXIQUE ET LEXICALE-SEMANTIQUE DE LA LANGUE ESPAGNOLE

A.Andreewsky[+], M.Desi, C.Fluhr

[+]LIMSI - CNRS, B.P.30, 91406 Orsay, France

Summary: TRAINING METHODS FOR THE AUTOMATIC MORPHOSYNTACTIC AND LEXICO-SEMANTIC ANALYSIS OF SPANISH

We describe herein the setting up of an automatic morphosyntactic and lexico-semantic analysis system for Spanish. This system uses learning methods analogous to those used for French (cf. Andreewski et al.: 1973, Dunod, 1974 and 1977 IFIP proceedings).

The learning is going on step by step (1.000 words each) and a syntactical parsing (specific to Spanish) attributes grammatical labels to specific words and suffixes, chosen for their low rate of grammatical ambiguities.

An ambiguity accumulation dictionary and morphosyntactic rules allowing for the resolution of these ambiguities are obtained automatically. They are progressively stabilized with the growing of the corpus.

The method is discussed:
- first, how to obtain lexico-semantic relations by filtering methods;
- second, how to handle the linguistic processing in Spanish with the "SPIRIT" automatic system (indexing and retrieval in natural language).

INTRODUCTION

Proposée en 1971-72 dans le but de résoudre un certain niveau d'ambiguïtés du langage, la méthode d'apprentissage que nous appliquons ici à la langue espagnole est très analogue à celle utilisée pour la langue française et pour laquelle de très bons résultats ont été obtenus (1).

Rappelons que le concept d'apprentissage auquel nous faisons appel repose sur le fait que dans toute chaîne lexicale des ambiguïtés apparaissent constamment, et que l'on doit supposer (pour l'homme) qu'à chaque fois des procédures de désambiguation sont dynamiquement mises en oeuvre pour les traiter.

L'ambiguïté que l'on traite est celle qui provient de la réutilisation d'un même mot ou groupe de mots avec des valeurs syntaxiques (catégorisation grammaticale) ou sémantiques différentes. Pour un traitement satisfaisant de ces ambiguïtés dans les applications à l'indexation automatique et l'interrogation de bases de données en langage naturel, on a été amené à formuler deux hypothèses essentielles:

- le contexte limité aux termes voisins doit permettre de lever un nombre très important d'ambiguïtés syntaxiques;
- l'étiquetage grammatical du texte doit permettre d'obtenir par "filtrage" des relations dites "lexicales sémantiques" et de traiter les ambiguïtés sémantiques. (Hypothèses qui sur le français ont donné de bons résultats.)

D'autre part le texte d'apprentissage espagnol est accompagné de la traduction correspondante française, afin de mieux étudier et préciser les problèmes de la micro-idiomatique dans le processus de la traduction (pas nécessairement automatique).

LA METHODE D'APPRENTISSAGE

Le principe de la méthode d'apprentissage, largement décrit dans (1), est le suivant: on analyse manuellement un texte T dit d'"apprentissage" accompagné de sa traduction, d'une analyse de terminaison et d'une analyse grammaticale, comme cela est indiqué dans l'exemple ci-dessous où l'on trouve: dans la première colonne le texte T lui-meme, dans la deuxième colonne la traduction, dans la troisième colonne la terminaison éventuelle du mot espagnol, et dans la quatrième collone le catégorie grammaticale réalisée dans le texte.

Remarques:

1. On s'efforce de faire une traduction aussi proche que possible du texte, mais intelligible. Les mots indispensables à l'intelligibilité et qui ne sont pas dans le texte espagnol, sont mis entre parenthèses en français.
2. Les terminaisons sont choisies en fonction de leur caractère discriminant, c'est-à-dire qu'elles ne sont caractéristiques que d'une seule catégorie grammaticale en général,

deux au plus. Si elles ont deux catégories grammaticales, il est supposé que le contexte voisin permettra de lever l'ambiguïté, ce qui est vérifié dans l'autocohérence.

3. Des rangements par ordre alphabétique de chacune des quatre colonnes, permettent au cours de l'apprentissage de vérifier la qualité du codage à savoir: correction des erreurs orthographiques, incohérences dans les codes grammaticaux dans la terminaison (deux trerminaisons différentes pour un même mot) dans la traduction.

como	comme	Ø	conjonction subordination
son	(ils) sont	Ø	verbe d'état indicatif
interesantes	intéressants	antes	attribut
para	pour	Ø	préposition
todos	tous	Ø	pronom général complément
,	,	Ø	,
los	les	Ø	article défini
documentos	documents	mentos	substantif
no	ne	Ø	négation no
estan	sont (pas)	Ø	verbe d'état indicatif
en	dans	Ø	préposition
la	la	Ø	article défini
biblioteca	bibliothèque	teca	substantif
sino	mais	Ø	élément de la négation
sobre	sur	Ø	préposition
la	la	Ø	article défini
mesa	table	Ø	substantif.

Un grand nombre de mots de cette phrase sont ambigus, comme on peut le constater en examinant les phrases qui suivent: baïlar al son de guitarra; como una naranja; el tiempo se para y mi sino se juega ahora; los hechos importantes son los de la experimentacíon; la cuenta está en el sobre; el la y el mi de mi piano suenan mal.

Si ensuite, à partir du texte d'apprentissage, on constitue un dictionaire de cumul, il aura la forme (ici ne figurent que les mots ambigus):

```
como  :  conjonction de subordination, verbe indicatif, ...
la    :  article défini, substantif, ...
los   :  article défini, pronom attribut, ...
para  :  préposition, verbe conjugué, ...
sino  :  préposition, substantif, ...
sobre :  préposition, substantif, ...
son   :  verbe état indicatif, substantif, ...
```

De meme, est constitué un dictionaire de cumul des terminaisons, par exemple:

antes : attribut, adjectif postérieur, substantif
mentos : substantif
teca : substantif

Dés que le texte devient assez long, les items lexicaux se rencontrent avec des étiquettes syntaxiques et des acceptions différentes, mais pour les terminaisons cela se produit assez vite. C'est pourquoi on a effectué un apprentissage mixte qui porte à la fois sur les mots pleins sans terminaison, les terminaisons et les mots relationnels.

A partir du texte initial T et du dictionnaire de cumul, un texte ambigu T_A est créé (les terminaisons sont précédées d'un tiret). Il a la forme:

como	(conjonction de subordination, verbe indicatif)
son	(verbe état indicatif, substantif)
-antes	(attribut, adjectif postérieur, substantif)
para	(préposition, verbe conjugué)
todos	(pronom général complément)
los	(article défini, pronom attribut)
-mentos	(substantif)
no	(négation no)
estan	(verbe d'état indicatif)
en	(préposition)
la	(article défini, substantif)
-teca	(substantif)
sino	(élément de la négation, substantif)
sobre	(préposition, substantif)
la	(article défini, substantif)
mesa	(substantif)

Et la comparaison de T et T_A permet d'obtenir des régles de résolution qui par exemple à l'ordre trois avec le texte chosi auront la forme:

(conj sub, verb ind) * (verb état ind, substantif) * (attribut, adj p,substantif)

où l'astérisque * se lit: suivi de, et où nous avons surligné les résolutions obtenues par comparaison de T_A avec T. On remarque que (attribut, adj p, subst) est une ambiguïté cumulée par une terminaison.

RESULTATS ET CONCLUSIONS

Le corpus d'apprentissage a été constitué à partir de textes variés littéraires ou scientifiques. Il est actuellement de cinq mille mots, se qui nous a amené à effectuer une catégorisation grammaticale assez complète (120 catégories actuellement) et nos a permis d'obtenir une diversité syntaxique suffisante pour les applications envisagées.

Ces dernières sont essentiellement orientées vers l'indexation automatique et l'interrogation en langage naturel dans le cadre du système SPIRIT qui impose tout d'abord une normalisation correcte des mots du texte afin d'en faire des comptages cohérents. Pour obtenir cette normalisation, on suppose que sont identifiés les singuliers et pluriels des substantifs, les flexions de la conjugaison, etc.., ce qui doit être fait en relation avec l'analyse syntaxique, grâce à un dictionnaire en formes complètes du même type que le dictionnaire de cumul décrit plus haut. La normalisation se fait alors suivant le schéma:

texte + lexique en formes complètes → texte ambigu → syntaxe → → normalisation

D'autre part, le système SPIRIT prend en compte les mots composés qui, grâce à la syntaxe décrite, peuvent être obtenus par filtrage. Ce problème a été étudié par comparaison avec les méthodes étudiées en français.

Rappelons que (2) le filtrage consiste à trier automatiquement l'ensemble des chaînes du corpus de structure grammaticale donnée: par exemple:

substantif * adjectif : estudios metalográficos
substantif * del * substantif: energía del átomo
substantif * adjectif * de la * substantif: control constante de la radiactividad
infinitif * las * substantif * adjectif: absorber las radiaciones peligrosas

Par ailleurs la structure donnée peut, selon le contexte, représenter ou non un concept. En conséquence, la structure du contexte doit être précisée: par exemple la structure: substantif « de un » substantif est un concept dans: fijación al ayuntamiento de un aviso et n'en est pas un dans: fijación al ayuntamiento de un pueblo. Par contre, précédée d'un point et un article et suivie d'un verbe conjugué, cette structure n'est plus ambigué.

Les filtres obtenus en français semblent s'appliquer à l'espagnol avec toutefois certaines modifications dues aux différences par rapport au français (pronoms agglutinés au verbe, absence fréquente du pronom personnel, de l'article indéfini au pluriel, de l'inversion de sujet par rapport au verbe, etc...).

BIBLIOGRAPHIE

(1) A. Andreewski, C. Fluhr

- Apprentissage - Analyse automatique du langage, application à la documentation. Dunod - documents de linguistique quantitative, no 21, 1973 (livre 250 pages).
- Analyse comparative du contenu, indexation automatique. Séminaire IRIA, février 1974.
- A learning method for natural language processing and application to information retrieval. IFIP congress, août 1974, pp. 924-926, Stockholm, éd. North-Holland.

(2) A. Andreewski, F. Debili, C. Fluhr

- Computational learning of semantic lexical relations for the generation and automatic analysis of content. (pp. 667-673), IFIP congress Toronto, août 1977.
- Apprentissage en syntaxe et sémantique. Revue du Palais de la Découverte, Vol. 9, No 83, pp. 17-40, décembre 1980.

(3) Y. N. Marchuk

- Dictionnaire contexto-logique de traduction des polysèmes de l'anglais en russe. Moscou, 1976.