Disambiguation strategies for Spanish used in PrADo

PrADo

Project PrADo was a project of two universities in Catalonia, Autonomous University of Barcelona and Pompeu Fabra University, with the aim of creating a grammar checker prototype for Catalan and another one for Spanish, from the texts tagged with a morphosyntactic tagger. The project was realised from 2001-2003 and the results of the project are summarized in "PrADo: Preparación Automatizada de Documentos" [1] from March 2004.

Module for Spanish language

Before any work on the tools was started, a corpus was created. The size of the corpus was 238.766 words with texts from web, newspapers, literature and e-mails (without specialisation) and also with texts specialised: on law and linguistics. The corpus was created with specific users in mind: trilingual (Spanish, Catalan and English -- all texts from Iberian Peninsula), contemporary (texts no older than 1.1.2000) and users with higher level of writing. The reason was to create tools that would effectively work with these types of users (and their texts) and also to create a corpus of errors for correction tools development.

Preprocessing

To preprocess the text, an external tool named TACO+ from Polytechnic University of Catalonia was used. TACO+ is in principle a tagger, but incorporates its own text preprocessing. This made the Spanish text preprocessing easier, but on the other hand it was necessary to alter the output to comply with Constraint Grammar.

First of all, there was a need to detect abbreviations. MACO+ gives every abbreviation a tag, but it is not possible to use them directly for creating grammar rules. Moreover, MACO+ cannot assign a morphological category to a concrete abbreviation.

MACO+ also lacks the ability to mark boundaries of sentences and cannot distinguish between simple verbs and verbs with clitics, so additional modules were created for marking sentences with SGML and preserving information about clitics in the format used by Constraint Grammar.

Morphology

MACO+, as stated before, was used for morphological tagging. Tags made by MACO+ were then automatically converted to Constraint Grammar format.

Grammar used for Spanish language was structured in similar matter as the one for Catalan to two blocks. The first one consists of rules that eliminate ambiguity that was caused by tagging words. To eliminate ambiguity correctly, specific user cases mentioned before are taken into account (the way they use anachronisms and Americanisms for instance).

The rest of the rules are organised according to different morphological categories. First rules refer to closed categories (determiners, prepositions, conjunctions, pronouns, adverbs) because based on them it is possible to better categorise other words and detect their characterictics (grammatical number, gender, etc.).

Apart from corpus mentioned before, additional sources of data are used: dictionaries (especially DRAE [2]) and grammars (GDLE [3]).

Disambiguation

Disambiguation strategies for nouns and verbs

Distinguishing between noun and verb is one of the most important disambiguation problems and affects almost 6 % of the corpus.

- Missing concordance: if for example gender of an article doesn't match the word, it is probably a verb.
- Presence of other verb in the phrase: if there is other verb, treat the word as noun.
- Other restrictions based on whether concrete interpretation is possible in verbal and non-verbal contexts.

With these techniques it was possible to reduce ambiguous cases by 81 %.

Disambiguation strategies for pronouns and determiners

Nearly 3 % of cases where a word can be either pronoun or determiner have to be resolved.

- If the following word is not a verb and is not ambiguous noun, it is not pronoun. (*las flores* vs. *las cantas, las niñas* vs. *las cuentas*)
- If the following word is no-ambiguous verb, it is a pronoun.

With these techniques it was possible to reduce ambiguous cases by 84.7 %.

Disambiguation strategies for prepositions

- Preposition and adverb: in majority of cases it is possible to resolve ambiguity between these two categories, because preposition is always followed by some nominal structure.
- Preposition and noun: possible combinations are considered here, for instance presence of two prepositions in a row or presence of a determiner before preposition.
- Preposition and verb: for this kind of cases (like „bajo“ and „entre“) were introduced rules more ad hoc that are difficult to generalise.

With these techniques it was possible to reduce ambiguous cases by 99.7 %.

Disambiguation strategies for conjunctions

For conjunctions, similar strategies like for prepositions were used and it was possible to reduce ambiguity by 82.14 %.

Disambiguation strategies between verbs

A lot of verbs in Spanish has forms that are the same words but everytime it refers to different person, tense or verb. 10.38 % of the corpus is affected by this ambiguity. Again, for resolving this kind of words, concordance for number and person is used. Also, grammar mood is considered and it is possible to omit some verb forms with respect to specific type of user who wrote the text. Effectiveness of this rules was 87.41 %.

References:

[1] http://mutis.upf.es/glicom/Papers/inftecn/pradov2.pdf (accessed January 31, 2013)

[2] http://lema.rae.es/drae/ (accessed January 31, 2013)

[3] http://es.wikipedia.org/wiki/Gramática_descriptiva_de_la_lengua_española (accessed January 31, 2013)

Tools & Resources

Petra Tag - Spanish POS Tagger. http://petrapostagger.sourceforge.net/ (accessed January 31, 2013)

FreeLing - library providing language analysis services (including POS tagging) for various languages (including Spanish). http://nlp.lsi.upc.edu/freeling/ (accessed January 31, 2013)

TreeTagger - a language independent part-of-speech tagger developed at University of Stuttgart (also for Spanish). http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/ (accessed January 31, 2013)

Corpus de Referencia del Español Actual (CREA) – corpus of contemporary Spanish by Real Academia Española (The Royal Spanish Academy). http://corpus.rae.es/creanet.html (accessed January 31, 2013)

Corpus Diacrónico del Español (CORDE) – historical corpus of Spanish by Real Academia Española (The Royal Spanish Academy). http://corpus.rae.es/cordenet.html (accessed January 31, 2013)

Corpus del Español - free online Spanish (historical) corpus with 100 million words. http://www.corpusdelespanol.org/x.asp (accessed January 31, 2013)

CRATER - Multilingual Aligned Annotated Corpus for English, French and Spanish http://www.comp.lancs.ac.uk/linguistics/crater/corpus.html (accessed January 31, 2013)

Various Spanish corpuses from the Laboratorio de Lingüística Informática: http://www.lllf.uam.es/ING/Recursos.html (accessed January 31, 2013)

A Universal Part-of-Speech Tagset including Spanish http://code.google.com/p/universal-pos-tags/ (accessed January 31, 2013)