Tagging a spontaneous speech corpus of Spanish

Antonio Moreno
Dept. of Linguistics
Autonomous University of Madrid
sandoval@maria.lllf.uam.es

José Ma. Guirao
Dept. of Software Engineering
University of Granada
jmguirao@ugr.es

The C-ORAL-ROM corpus

–
– A comparable corpora in the main Romace languages: French, Italian, Portuguese and Spanish, funded by EU Commission under the contract IST 2000-28226.
– Over 300.000 words of spontaneous speech, recorded in real contexts without any restriction or script.
– Great variety of language register: formal vs. informal, media, telephone conversations.
– Balanced sociolinguistic features like sex, age or education.
– High acoustic quality: digital recording.
– Visit C-ORAL-ROM project home site: http://lablita.dit.unifi.it/coralrom/

Tagging spoken corpus vs written corpus

	Written corpus	Spoken corpus
Syntax	Sentential and discourse coherence, marked by grammatical means (conjunctions) and orthografic punctuation (commas, periods, etc.). A fixed or cannonical word order. Absence of repetition or retracting, that is, no agrammatical constructions.	Free, relaxed word order. Repetition. Retracting, resulting in agrammatical constructions. Sub-sentential fragments. No punctuation marks.
Lexicon	Proper Names recognition. Many new terms.	Absence of the Proper Names recognition problem. Low presence of new terms. Importance of derivative preffixes and suffixes that do not change the systactic category (mostly appreciative morphemes).
	Un mutante sospechoso Células infectadas por el virus observadas mediante microscopio en la Universidad de Hong Kong. REUTERS El anuncio de un equipo de investigadores canadienses que ha conseguido descifrar el código genético del virus sospechoso de haber provocado el síndrome respiratorio agudo severo (SRAS) se ha convertido en un importante primer paso para desarrollar pruebas diagnósticas y tratamientos para esta mortífera enfermedad, y en el último escalón de una carrera científica sin descanso por dar con el culpable de esta pandemia global. El genoma parece ser el de un coronavirus "completamente nuevo", una nueva cepa, una mutación de alguno de los tres microorganismos conocidos hasta ahora. Éste virus, nunca detectado en humanos, podría ser el cuarto.	@Place: Madrid @Situation: chat between friends in the living-room, hidden, researcher not present @Topic: dogs, comics, glasses and messages @Source: C-ORAL-ROM @Class: informal, familiar/private, dialogue *LET: pues / la vas a llamar // *DAN: <no recuerdo lo de los xxx> // *LET: [<] <porque / Nesca> / ha tenido camada / y ha tenido diez perros // *DAN: si // *LET: / pues / le encantan los boxer atigrados // entonces le quiero regalar uno // ya he visto los perritos nacidos y todo / encima que claro / casi me llevo un mordisco de Nesca / y + *DAN: por celosa // *LET: eh ? claro // por [/] no / por protección / <de madre> // *DAN: [<] <por eso / por> celosa / por proteger a sus <cachorrillos>

Tokenization and tagging

Tokenization: Sentence or paragraph boundaries, and punctuation marks make no sense in spontaneous speech. Instead, dialog turns and prosodic tags are used for identifying utterances boundaries.

Tagging: Our tagger relies on a morphological analyser, GRAMPAL, that assigns all possible tags to a particular word.

GRAMPAL is based on a rich morpheme lexicon of around 40.000 lexical units. The advantage of the "lexicon" approach is to provide the search space for every possible ambiguity, assuring that rare POSs are always considered.

Disambiguation

Rule-based Constrain Grammar
Our disambiguation system consist of two sets of rules:

Lexical rules for every ambiguous word, stating the syntactic context for every POS:

Asign the tag T_j to word w_i when then preceding POS tag is T_k,
or
Asign the tag T_h to word w_i when the following POS tag is T_l.

Example:

– Asign the tag MD to 'hombre' (English 'man') when preceding tag is '#'
– Asign the tag N to 'hombre' when preceding tag is ART

These rules have been inferred automatically from the training corpus. For stating a lexical rule, a minimum of positive and no negative cases have to occur. These rules can be adjusted by hand. In addition, rules for very low frequency POSs can be written. The procedure is a combination of automatic and supervised learning.

Syntactic rules: these are general bigram tags ordered by frequency in the training corpus. In our experiment we have used 50 rules. The top five general rules are: 'ART N', 'P V', '# C', 'ADV #', and 'V PREP'.

Asign tag T_j to w_i if

either there is the rule T_jT_x and the next tag is T_x

or there is the rule T_xT_j and the previous tag is T_x

The disambiguation algorithm is:

apply the higher lexical rule that matches a syntactic context

in case of no lexical rule available, apply the higher general syntactic rule,

else, apply the most frequent POS for that word

Unknow words recognition

Four types of UW:

1. foreing words
2. missing words in the lexicon
3. mispelling in the transcription
4. neologims

GRAMPAL has been extended with derivation rules and morphemes
The Prefix rule is:

Take any prefix and any (inflected) word and form another word with the same features.

This rule is effective for POS tagging since in Spanish the prefixes never change the syntactic category of the base. The rule assings the category feature to the new word. With this information, the corresponding POS tag is assigned to the unknown word. 239 prefixes have been added to the GRAMPAL lexicon.

GRAMPAL has been also extended with the most productive suffixes in Spanish, including *-ble*, *-dero*, *-dizo*, *-dor*, *-ivo*, *-oso*, *-torio*, *-ante*, *-ción*, *-dad*, *-ez*, *-ista*, and *-ificar*.

Evaluation

	Complete corpus			
	Tokens	%	Types	%
One analysis	226507	75,1	13786	71,8
Ambiguous	65272	21,6	2180	11,4
Unknown	3132	1,0	1542	8,0
Names	6642	2,2	1698	8,8
TOTAL	301553	100	19206	100
	Training sub-corpus			
	Tokens	%	Types	%
One analysis	65124	75,4	4701	69,1
Ambiguous	18561	21,5	1048	15,4
Unknown	772	0,9	459	6,7
Names	1929	2,2	594	8,7
TOTAL	86386	100	6802	100
	Test sub-corpus			
	Tokens	%	Types	%
One analysis	17375	76,4	2791	74,9
Ambiguous	4693	20,6	584	15,7
Unknown	238	1,0	145	3,9
Names	441	1,9	205	5,5
TOTAL	22747	100	3725	100

Table 1 shows the initial results. First, the data for the whole corpus (160 texts); then the training sub-corpus (57 texts), and the initial figures for the test sub-corpus (10 texts).

For the disambiguation, 1446 lexical rules and 50 general syntactic rules have been inferred from training corpus. In a first evaluation with the 22747 words (4693 of them ambiguous) of the test sub-corpus, the system made 357 errors in assigning the proper POS tag, that is 1.5% of all the tokens, 7.7% of the ambiguous words.

	Unknown words in the test set			
	Tokens	%	Types	%
Initial results	238	1,0	145	3,9
Evaluation results	41	0,18	33	0,85

After passing the unknown words recogniser through the test sub-corpus, only 41 words remain unknown from the initial 238. The significant reduction from 1% of test set to 0.18% is due mostly due the derivative rules and new lexical entries added during the training.

The disambiguation method and the unknow words recognition module provide significant improvements against the initial scores. As a whole, the morpho-syntactic tagging system gives a success rate of 98.3%.