Tagging a spontaneous speech corpus of Spanish

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

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

Abstract

This paper describes the experience of tagging the Spanish corpus of the C-Oral-Rom project with morphological and POS information. This is a spontaneous speech corpus of 300.000 words that covers a great variety of language registers (formal and informal speech, media, telephone recordings). In total, there are 19.206 different words (types). Social features as sex, age or education have been taken into account for the sampling. All the recordings have been set in their real context without any restriction or scripted design.

In order to estimate the difficulty of the task, firstly we divide the performance of the tagging system into four classes: names, ambiguous words, non-ambiguous words, and unknown words. Then we present the training of the system, and finally the evaluation.

1 Introduction

Annotating spontaneous speech corpora presents some problems not found when annotating written corpora, but strategies originally developed for written texts may be adapted to the task. We present our experience in morphological and POS tagging of a spontaneous speech corpus of Spanish.

C-Oral-ROM is a multi-lingual corpus (Cresti et al., 2002) of the four main Romance languages: French, Italian, Portuguese and Spanish. Each sub-corpus has approximately 300.000 words that extend over a great variety of language registers, including informal speech (150.000 words), formal (65.000 words), media (60.000 words) and telephone recordings (25.000 words). In order to obtain a well-represented sampling we have taken into account balanced sociolinguistic features such as sex, age and education. Other criteria relevant for the corpus design have been: acoustic quality (most text are digital recordings), legal status (we got the written permission of the speakers) and spontaneity (no previous script nor restrictions to their opinions and way of expression).

Transcription is orthographic, not phonetic. However, no other convention from written texts, such as punctuation marks, sentences and paragraphs, has been followed. Instead, prosodic tags (tone units, retracting, overlapping, disfluencies) and dialog turns are annotated. Each transcription has been revised by three different linguists, and finally, a sound-text alingment of every utterance is made by hand.

With respect to the linguistic annotation, the main goal is to provide a complete morphological and POS tagging. These tasks will be performed automatically and validated by expert annotators.

For the morphological analysis we employ GRAMPAL (Moreno, 1991; Moreno & Goñi, 1995) which is based on a rich morpheme lexicon of around 40.000 lexical units, and morphological rules. This system has been successfully used in language engineering applications, ARIES (Goñi, González & Moreno, 1997). The tagging will be the most useful test for showing the ability of GRAMPAL to deal with a wide-coverage corpus of Spanish. We use this experiment for enhancing GRAMPAL with new modules: a POS tagger and an unknown words recogniser, both specifically developed for spoken Spanish.

Fortunately, proper names recognition, which is a hard problem for written corpora, is not a problem in our case: since it is an orthographic transcription of spontaneous speech, we do not consider it appropriate to follow the punctuation conventions of the written language. Only names are transcribed with a capital letter. Since each text is revised carefully by three different linguists, they resolve any possible ambiguity using the context and their communicative skills. As a consequence, name recognition is a trivial task: just find the words with a capital letter, since human transcribers did the job.

2 Morphological system performance

Tokenization in spoken corpora is slightly different to the same task in written corpora. No sentence or paragraph boundaries make sense in spontaneous speech. Instead, dialog turns and prosodic tags are used for identifying utterances boundaries.

After tokenization, we take the whole corpus and feed GRAMPAL, filtering all the proper names. In the output of the morpho-syntactic module four classes of words are distinguished:

1. Non-ambiguous words: those words that received a unique POS interpretation. Spanish is an inflective language: different syntactic categories use different sets of inflection morphemes.

2. Ambiguous words: those which got more than one morpho-syntactic interpretation.

3. Unknown words: those which are not recognised by the program, because they are not in the lexicon.

4. Names: this class includes also acronyms.

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).

As the figures show, we can split the POS tagging task into problematic and non-problematic words. The former ones are the ambiguous and unknown words (around 20% of the corpus). The non-problematic words constitute the remaining 80%.

In the following sections we will present our currently-under-construction POS tagger and unknown word recogniser. In the final section, we will provide the results of an evaluation against a fragment of the corpus, following the same criteria as the initial test shown in Figure 1.

3 Building a POS tagger for spoken Spanish

Our POS disambiguation method is a rule-based, constraint grammar (CG), applied successfully in other taggers (Brants & Samuelsson, 1995; Chanod & Tapannien, 1995). These taggers make extensive use of lexical rules extracted from a training corpus.

Different strategies are used in order to infer the relevant data for the disambiguation task. Some systems apply mechanical, statistical methods (decision trees, HMM). In others, like ours, the human expert supervises the extraction of rules from data, and sometimes writes the rules. The latter are called ”linguistic taggers” as opposed to ”statistical taggers.” Different evaluations show that both perform similarly when they are well-trained.

The main difference with usual taggers is that ours is applied to spoken language, instead of written texts. The syntax of spontaneous speech shows two important difficulties with respect to written language:

1. Shorter utterances: the typical syntactic and pragmatic units of the written texts, the **sentence** and the **paragraph**, do not apply in spoken language, especially in dialogues or conversations. Instead, we find smaller fragments based on prosodic (tone) units that correspond to a few words. As a consequence, in transliterated spoken corpus, prosodic tags are used in a simmilar way as punctuation marks are in written corpora.

2. Non-canonical grammatical phrases: when one writes, a written norm is always behind. For instance, regardless of what is grammatically correct or incorrect, no one writes twice the same word in Spanish. However, repetition is very frequent in most speakers: ”la la la medida que tomó el Gobierno” (lit.

	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: Initial results

”the the the measure taken by the Government”). This example produces this tagging: ART ART ART N... No grammar or tagger prepared for written text would accept this phrase, which is very frequent in any spoken context. In addition, we have also retracting (when the speaker changes his or her mind in the middle of the utterance production) or even concordance breaking (*la casa mío*, mixing feminine and masculine in the NP)[1]. Word order is more relaxed in spoken language.

Summing up, a tagger trained on a written corpus will not provide satisfactory annotation for a spoken corpus[2].

After training, 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.

  For context, the tag of the previous or posterior word is used, as convenience. Here is an example (MD means ”marcador discursivo”, Discursive Marker; ”#” stands for start or end of utterance)

  - 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

[1] In most cases, the difference is that the written text can be edited, while the spoken discourse cannot be edited in the same way, all the ”errors” remain in the transcription.

[2] (Uchimoto et al, 2002) provide a similar experience tagging a Japanese spontaneous speech corpus.

can be adjusted by hand. In addition, rules for very low frequency POSs, like la-N (musical note A) when preceding tag is ART, 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 $T_j T_x$ and the next tag is T_x
>
> or there is the rule $T_x T_j$ 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

This tagger will be evaluated against other types of taggers when the final version will be ready.

4 An Unknown Word Recogniser (UWR) for spoken Spanish

The rate of unknown words in the whole corpus is 8.02% (or 1542 word types). This rate is similar to the names (8.84%) and a little bit lower than that for ambiguous words (11.35%.) First step was to classify the unknown words in the training corpus (772 word tokens, 459 word types) into different classes:

1. Foreign words: *walkman*, *parking*, etc.
2. Missing words in the lexicon: *caramba*, *hijoputa*, *yuanes*, etc.
3. Errors in the transcription: names in lower case and misspellings.
4. New words or neologisms: mostly formed by derivative morphemes.

We estimate that more than fifty percent of the cases are neologisms or innovations produced by the speakers. In spontaneous spoken language it is especially frequent the use of emphatic and expressive affixes, such as the prefixes *super-*, *mega-*, diminutive suffixes like *-ito, -ico, -illo*, or the superlative suffix *-ísimo*. Resolving those cases, we can provide a POS tag to the most important group of unknown words in our corpus.

In order to analyse them, GRAMPAL has been extended with derivation rules and morphemes[3]. Here we only provide some examples; a complete account of the derivative rules will be reported in another paper.

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.

The Diminutive rules are similar to the Prefix one rule:

> A given suffix is concatenated to either a nominal root, like *gat-* ('cat'), *abuel-* ('grandfather') or an inflected word, like *azul* ('blue') to generate a new word with the `Lexeme` and `Category` features from the base and the `Gender` and `Number` features from the suffix.

This rule only applies to those suffixes that do not change the POS of the base (nominal lexeme or inflected word).

For changing-category suffixes, particular rules are needed:

[3] The original GRAMPAL morphological processor only deals with inflectional morphology, including clitics.

> Form an Adjective or a Noun from the concatenation of a verb root and a nominal suffix. The syntactic category and the agreement features are transferred from the suffix.

There are also rules for other types of derivation, Verbs derivate from Nouns or from Adjectives.

Currently we have extended GRAMPAL with the most productive suffixes in Spanish, including *-ble, -dero, -dizo, -dor, -ivo, -oso, -torio, -ante, -ción, -dad, -ez, -ista,* and *-ificar*. It must be noticed that a simple suffix stripping will not provide the same good results comparable to those obtained using derivation rules, since we make use of a lexicon that reduces over-generation.

In order to resolve the remaining three classes of unknown words we need a different, simpler approach than using rules:

- Foreign words are included in a list. This list has been extracted by hand from writing manuals and from corpora. It is regularly updated.

- Any word which appear in the training corpus but not in the lexicon is added, expanding the base resource.

5 Evaluation results

After passing the unknown words recogniser through the test sub-corpus, only 41 words remain unknown from the initial 238.

For 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.

6 Conclusions and future work

This paper has reported the promising results of an experiment to tag a spontaneous speech corpus. 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%.

As other authors have previously pointed out (Uchimoto et al, 2002), spontaneous speech corpora need special tools. Some remarking features of spontaneous speech are:

- In syntax: free, relaxed word order, retracting, word repetition, sub-sentential fragments, absence of punctuation. Prosodic tags are used for tokenization and disambiguation.

- In the lexicon: absence of the proper names recognition problem, low presence of new terms, importance of the derivative suffixes that do not change the syntactic category (mostly appreciative morphemes).

7 Acknowledgment

This research has been partially funded by the EU Commission under the C-Oral-Rom project (IST 2000-28226).

References

T. Brants & C. Samuelsson. 1995. Tagging the Teleman Corpus. In *Proceedings of the 10th Nordic Conference of Computational Linguistics, NODALIDA-95*

J-P Chanod and P. Tapanninen 1995. Tagging French: comparing statistical and a constraint-based method. In *Procs 7th Conference of the EACL*, pp 149-157.

E. Cresti, M. Moneglia, F. Bacelar, A. Moreno, J. Veronis, P. Martin, K. Choukri, V. Mapelli, D. Falavigna, A. Cid, and C. Blum. 2002. The C-ORAL-ROM Project. New methods for spoken language archives in a multilingual romance corpus In *Procs. of LREC-2002* pp. 2-9.

José M. Goñi, José C. González and Antonio Moreno 1997 ARIES: A lexical platform for engineering Spanish processing tools. *Natural Language Engineering* 3(4): 317–345.

Antonio Moreno 1991. *Un modelo basado en la unificación para el análisis y generación de la morfología del español.* Ph.D. dissertation. Universidad Autónoma de Madrid.

Antonio Moreno and José M. Goñi 1995. GRAMPAL: A morphological model and processor for Spanish implemented in Prolog. In *Procs of the Joint Conference on Declarative Programming (GULP-PRODE'95)*, pp. 321-331.

K. Uchimoto, C. Nobata, A. Yamada, S. Sekine, and H. Isahara 2002. Morphological Analysis of The Spontaneous Speech Corpus. In *Procs. of COLING 2002.*