A proposal of a morphological tagger for Spanish based on Cuban corpora

Lisette García-Moya
Center of Pattern Recognition and Data Mining
Universidad de Oriente
Santiago de Cuba
lisette@csd.uo.edu.cu

Aurora Pons-Porrata
Center of Pattern Recognition and Data Mining
Universidad de Oriente
Santiago de Cuba
aurora@csd.uo.edu.cu

Leonel Ruiz-Miyares
Center for Applied Linguistic
Santiago de Cuba
leonel@lingapli.ciges.inf.cu

Abstract

In this paper we describe a morphological tagger for Spanish based on Cuban corpora. The tagger combines Hidden Markov Models with some heuristics and dictionaries to provide the appropriate part-of-speech tag for each word in a text document, according to the context in which it appears.

Moreover, a morphological analyser that provides all possible morphological interpretations of words is used. It allows us to reduce possible grammatical tags and to obtain not only the appropriate part-of-speech tag, but also its morphological information. The proposed tagger achieves 97.76 % accuracy for a legal corpus.

Keywords

Morphological tagger, Hidden Markov Model, Statistical Natural Language Processing.

1. Introduction

Part-Of-Speech (POS) tagging is an essential task for all Natural Language Processing activities, for example, Information Retrieval, which in turn helps managing the enormous amount of text documents available nowadays, such as: Web pages, news, scientific papers, emails, etc.

Many words in natural language are grammatically and semantically ambiguous. Grammatical disambiguation consists of assigning the appropriate part-of-speech tag to each word of a given textual document, according to the context in which word appears. This type of annotation is carried out by a morphological tagger.

Morphological taggers are classified into deductive systems based on knowledge, inductive systems based on machine learning approaches and hybrid systems.

In deductive systems –also known as linguistic approaches- the model is written by a linguist, generally in form of rules or constraints [14]. The linguistic models range from a few hundreds to several thousand rules, and they usually require years of hard work.

Inductive methods consider that linguistic knowledge may be inferred by experience. This experience is obtained by textual corpora. Inductive methods build a computational model from a set of examples which may be annotated with linguistic information or not, using learning or statistical methods. These methods could be supervised or unsupervised, depending whether training data contains linguistic information or not, respectively. Many inductive techniques have been developed to solve the problem of grammatical disambiguation, such as: *n*-grams models [1], memory-based learning [5], transformation-based error-driven learning [3], Hidden Markov models (HMM) [11], maximum entropy [12] and decision trees [16]. Markov models combined with a good smoothing technique and with handling of unknown words perform at least as well as other current approaches [2,6].

Finally, hybrid models [10] combine statistical information with automatically extracted rule-based information trying to join the advantages of both approaches.

Most of the taggers have been developed for the English language. Nevertheless, several hybrid POS taggers for the Spanish language have been proposed, such as *Freeling* [4] and the Spanish version of *TreeTagger* [16]. TreeTagger is based on decision trees, whereas FreeLing is a trigram HMM tagger. Although these taggers achieve good results, they have some limitations.

TreeTagger tends to assign the proper noun tag to words beginning with a capital letter, even when that word is, in fact, a common noun, as in *Banco Popular de Ahorro*. The tagset of TreeTagger is very basic. As a consequence, a lot of potentially useful morphological information (including, for example, gender, number, verb person, etc.) is not included in the tags. Numbers were also problematic. They were generally treated as CARD (Cardinal), but in some cases they were tagged as CODE (Alphanumeric code). Verb forms with enclitic pronouns are tagged as verbs only, resulting in loss of information on such pronominal particles.

Freeling is unsuccessful when encountering certain words not present in its vocabulary, such as unknown place names (Azerbayán and Tampere, tagged as a verb,

etc.). For unclear reasons, in some cases Freeling is not able to find the lemma of certain plural nouns and adjectives and left them in the plural [15].

Both TreeTagger and Freeling are neither able to recognise pronominal verbs that are reflexive of form, for example, *me abstengo*. Moreover, they do not recognise dates or times in short format, such as 25/12/2007, 25-12-2007 or 12:45.

On the other hand, there are some differences between the Spanish spoken in Cuba and the Spanish spoken in other Spanish-speaking countries, basically from a lexical point of view. The Spanish language together with Sub-Saharan African and Indocuban languages were three linguistic trends that strongly determined the own characteristics of the Spanish spoken in Cuba. Words of african origin such as *quimbombó*, *sambumbia* and *conga* and indigenous words (e.g. *hayaca*, *caguairán*, *fotuto*) enrich this language. At morphological level, there are no notable differences between the Spanish language spoken in Cuban and the Spanish of other countries. However, a morphological peculiarity could be mentioned: *vos* does not exist in Cuba; *tú* is used instead on informal environments and *usted* when the relation requires a polite form.

Ruiz-Miyares presented ETIPROCT [13], a morphological tagger for the Spanish spoken in Cuba, which achieves satisfactory results. However, the tagset of this tagger is limited and it does not allow annotating the text with morphological information and lemmas.

In this paper, we propose a morphological tagger with a greater tagset and broader morphological information than ETIPROCT. It combines HMM with a morphological analyser, heuristics and dictionaries. This tagger is considered as a hybrid one. The morphological analyser we used is based on two-level morphology from Kimmo Koskenniemi [8]. This paper is focused on the morphological tagger.

2. The morphological tagger

The architecture of the proposed tagger is shown in Fig. 1.

The tokeniser divides the raw text into atomic items and identifies the sentence boundaries. It is able to recognise words, punctuation marks, symbols and identifiers. We understand as *identifier* any sequence of characters that is not a word in the language, such as: email addresses, URLs, expressions like: 2+5*4=22, and others. Tokeniser also identifies acronyms, measurement units, abbreviations, phrases (nominal, adjectival, adverbial, prepositional, conjunctive and Latin phrases), dates, times and numbers by using several dictionaries and heuristics.

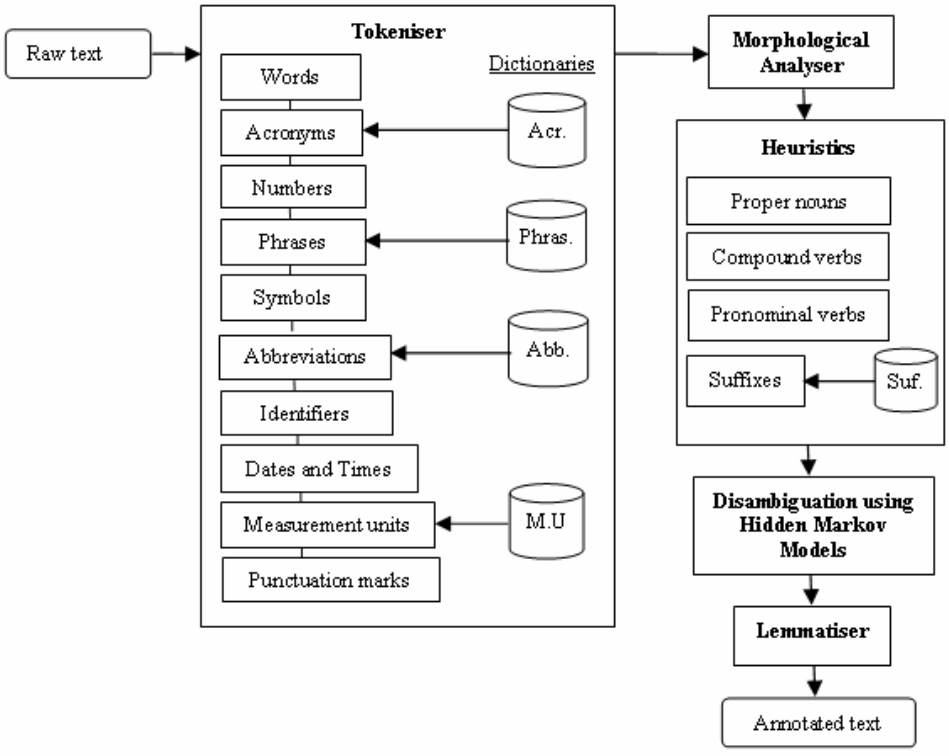

Figure 1. Architecture of the morphological tagger

The morphological analyser provides all possible interpretations of a given word and its morphological features. These features include gender, number, verb person, verb time, lemma information and so on. A *lemma* is defined as the canonical form of a word. A word could have different lemmas, for example, the lemma of the word *camino* is *camino* if it is a noun, and *caminar* if it is a verb.

Instead of regarding all possible tags for each word in the test data, we only consider the possible tags given by the morphological analyser. Thus, it allows us to constrain the set of possible grammatical categories of a given word and reduces the number of computations. For example, each word of the sentence *Reclamo una sentencia penal* has different possible tags given by the morphological analyser (see Figure 2).

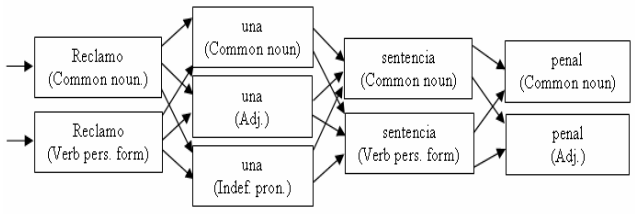

Figure 2. Possible tags are taken from morphological analyser

The morphological analyser and dictionaries of abbreviations, acronyms and phrases include the proper characteristics of Spanish spoken in Cuba on the basis of the Cuban corpora.

The proposed tagger is also able to recognise proper nouns, compound verbs (e.g. *he votado*) and pronominal verbs (e.g. *me abstengo*) by using some heuristics such as capital letters, the presence of verb *haber* and certain pronouns, etc. The set of possible part-of-speech tags of each word is obtained as a result of tokenization process, morphological analysis and heuristics.

Disambiguation process is carried out by applying Hidden Markov Models from the set of possible part-of-speech tags obtained before. Finding the appropriate lemma of each word is trivial by using the part-of-speech tag obtained by HMM model and the information provided by the morphological analyser. The lemmatiser considers acronym meaning and the expanded form of abbreviations and measurement units as lemmas.

The tagset used in our proposal is shown in Table 1. It is important to mention that tags include not only information about the major parts of speech but also other morphological information, such as number and gender for nouns, tense for verbs, and superlative, diminutive and despective forms for adjectives.

Table 2 shows the morphological features for each POS tag provided by the morphological analyser.

Table 1. Used tagset

Proper noun	Adjectival phrase
Common noun	Verbal phrase
Personal pronoun	Adverbial phrase
Demonstrative pronoun	Prepositional phrase
Possessive pronoun	Conjunctive phrase
Indefinite pronoun	Latin phrase
Relative pronoun	Article
Interrogative and exclamative pronoun	Preposition
Verb in personal form	Conjunction
Verb infinitive	Interjection
Verb gerund	Contraction
Verb participle	Adjective
Verb in personal form with enclitic	Adverb
Verb infinitive with enclitic	Acronym
Verb gerund with enclitic	Number
Multiple numeral	Measurement unit
Cardinal numeral	Date and Time
Ordinal numeral	Identifier
Collective numeral	Symbol
Fractional numeral	Punctuation mark
Nominal phrase	

Table 2. Morphological features for each POS tag

POS tag	Morphological features
Common noun	gender, number, degree
Personal pronoun	gender, number, person, politeness
Demonstrative pronoun	gender, number
Possessive pronoun	gender, number, person, politeness
Indefinite pronoun	gender, number
Relative pronoun	gender, number
Interrogative and exclamative pron.	gender, number
Verb in personal form	transitivity, pronominality, mode, tense, number, person, politeness
Verb participle	gender, number
Verb in personal form with enclitic	transitivity, pronominality, mode, tense, number, person, politeness
Article	gender, number
Adjective	gender, number, degree

2.1 Hidden Markov Model

As we mentioned above, the proposed morphological tagger is based on Hidden Markov Models. HMM is a widely used probabilistic finite state machine having a set of states, an output alphabet, transition probabilities, observation probabilities and initial state probabilities. In

our HMM model, states correspond to part-of-speech tags and observations correspond to words.

The HMM will be used to assign the most probable tag to the words of an input sentence. As we use a bigram model, output probabilities only depend on the most recent category, that is,

$$\arg\max_{c_i \in \{T_1,...,T_n\}} \left\{ P(w_k \mid c_i) \cdot P(c_i \mid c_j) \right\} \quad (1)$$

where w_k is the word to be disambiguated, $\{T_1,...,T_n\}$ is the possible tagset for w_k and c_j is the tag assigned to the previous word. Transition and observation probabilities are estimated from a tagged corpus.

When w_k is at beginning of a sentence, the probability of c_i being the grammatical category of the first word in the sentence (π_i) is estimated, instead of the transition probability $P(c_i \mid c_j)$, that is:

$$\arg\max_{c_i \in \{T_1,...,T_n\}} \left\{ P(w_k \mid c_i) \cdot \pi_i \right\} \quad (2)$$

2.2 Handling unknown words

The words that were not seen during the training are known as *unknown words*. Currently, the method of handling unknown words that seems to work best for inflected languages is a suffix analysis.

As Spanish is an inflected language, we use this method to predict the possible tags of an unknown word. In order to do that, we built a dictionary of frequent suffixes and its possible POS tags. For example, the *–ería* suffix is an indicator that word could be a common noun (e.g. *extranjería*) or a verb in personal form (e.g. *aparecería*).

In addition to the possible tags of the unknown word, an observation probability is required to applied equations (1) or (2).

To overcome data sparseness we apply the *Adding One* smoothing method, also known as Laplace's law [7], which adds one to all frequencies, thus avoiding zeroes and reducing the proportion between rare happening events. The observation probability is defined as follows:

$$P^{smoothing}(w_k \mid c_i) = \frac{f(w_k, c_i) + 1}{f(c_i) + |V|}$$

where *V* is the vocabulary in the training corpus, $f(w_k, c_i)$ is the number of times that word w_k is tagged as c_i and $f(c_i)$ is the number of words tagged as c_i in the training corpus. Then, the observation probability for unknown words is:

$$P^{smoothing}(w_k \mid c_i) = \frac{1}{f(c_i) + |V|}$$

If suffix analysis does not provide any possible grammatical category for the unknown word, Hidden Markov Model is applied assuming that unknown words may potentially have all tags, excluding those tags corresponding to closed categories (preposition, conjunction, article, etc.), which are considered to be all known. For unknown words, we consider as lemma the own word.

3. Experimental results

In order to evaluate our approach, a legal corpus containing 231634 words is built. This corpus was manually annotated by human experts.

We perform a 10-fold cross validation using 90% of the combined data set as training data and the remainder as test data. In the experiments, we use accuracy as our evaluation measure. It is defined as the ratio of the number of correctly tagged words to the total number of words.

The obtained results are shown in Table 3. Second and third columns contain the number of words in the training and test sets, respectively. As it can be appreciated, we obtained a similar accuracy to that of the current state-of-the-art taggers [2,4,9,12].

Table 3. 10-fold cross validation results

Subset	Training set	Test set	Accuracy (%)
1	208360	23274	98.02
2	209102	22532	97.71
3	208344	23290	97.70
4	209117	22517	97.69
5	207428	24206	97.70
6	209356	22278	97.94
7	207965	23669	97.69
8	208523	23111	97.77
9	208262	23372	97.69
10	208249	23385	97.73
Average			97.76

Table 4 summarizes the averaged accuracies obtained over different POS tags. As shown in the table, the tagger performs the worst in the verbs. The most common problems have been labelling as common noun words that should be infinitive verb, or labelling as adjective words that should be verb participle. In all other part-of-speech tags the accuracy values are similar. Thus, it seems that the effectiveness is not affected with different tags.

4. Conclusions

In this paper, a morphological tagger for texts written in Spanish with a particular emphasis on the Cuban variant has been presented. However, this tagger is able to process any text written in Spanish.

Table 4. Accuracy over different POS tags

POS tag	Averaged Accuracy
Proper nouns	100
Common nouns	97.69
Pronouns	98.01
Verbs	92.68
Numerals	98.97
Phrases	96.97
Article	98.23
Preposition	99.95
Conjunction	96.57
Contraction	99.98
Adjective	97.12
Adverb	98.06
Acronym	99.73
Number	98.09
Date and Time	100
Punctuation mark	99.96

The proposed tagger combines bigram-based HMM with a set of heuristics and dictionaries. Besides, it uses a morphological analyser which allows us to constrain the set of candidate grammatical categories to be considered for each word and provides richer morphological information.

In the experiments carried out on a legal corpus, we obtained a satisfactory accuracy (97.76%) that is similar to that of other taggers reported in the literature. The most common errors have been labelling words that should be verb infinitive as common noun, or words that should be verb participle as adjective. The proposed tagger becomes a high-quality tool for the annotation of Cuban corpora with part-of-speech information.

As future work, we plan to evaluate our morphological tagger on corpora from other knowledge domains. Also, we want to integrate it into other natural language processing tools, such as a named entity recogniser.

5. References

[1] L. R. Bahl, F. Jelinek and L. R. Mercer. A Maximum-Likelihood Approach to Continuous Speech Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI, pp. 179-190, 1983.

[2] T. Brants. TNT-A Statistical Part-of-Speech Tagger. In Proc. of the 6th Applied Natural Language Processing Conference ANLP-2000, Seatle, pp. 224-231, 2000.

[3] E. Brill. A simple rule-based Part of Speech tagger. In Proc. of the 3rd Conference on Applied Natural Language Processing of the Association for Computational Linguistics, Trento, pp. 152-155, 1992.

[4] X. Carreras, I. Chao, L. Padró and M. Padró. Freeling: an Open-source Suite of Language Analyzers. In Proc. of the 4th International Conference on Language Resources and Evaluation, Lisbon, pp. 239-242, 2004.

[5] W. Daelemans, J. Zavrel, P. Berck and S. Gillis. MBT: A Memory-Based Part of Speech Tagger-Generator. In Proc. of the 4th Workshop on Very Large Corpora, Copenhagen, pp. 14-27, 1996.

[6] S. Dandapat, S. Sarkar, and A. Basu. A Hybrid Model for Part-of-Speech Tagging and its Application to Bengali. International Conference on Computational Intelligence, pp. 169-172, 2004.

[7] H. Jeffreys. Theory of Probability, Second Edition, Section 3.23, Oxford, Clarendon Press, 1948.

[8] K. Koskenniemi. Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production. PhD Thesis. University of Helsinki, 1983.

[9] L. Márquez. Part-of-speech Tagging: A Machine Learning Approach based on Decision Trees. PhD Thesis. Polytechnic University of Catalonia, Barcelona, 1999.

[10] L. Márquez and L. Padró. A flexible POS tagger using an automatically acquired language model. In Proc. of ACL-97, Madrid, pp. 238-245, 1997.

[11] A. Molina. Disambiguation in Natural Language Processing by using machine learning techniques. PhD Thesis. Polytechnic University of Valencia, (In Spanish) 2004.

[12] A. Ratnaparkhi. A Maximum Entropy Model for Part-of-Speech Tagging. In Proc. of the 1st Conference on Empirical Methods in Natural Language Processing, EMNLP, Pennsylvania, 1996.

[13] L. Ruiz-Miyares. Development of a computational model based on tagging for processing textual corpora. PhD Thesis. Universidad de Oriente, Santiago de Cuba - University of Twente, Holanda, (In Spanish) 2001.

[14] C. Samuelsson and A. Voutilainen. Comparing a Linguistic and a Stochastic Tagger. In Proc. of 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics, ACL, Madrid, pp. 246-253, 1997.

[15] A. Sandrelli and C. Bendazzoli. Tagging a Corpus of Interpreted Speeches: the European Parliament Interpreting Corpus (EPIC). In Proc. of the 5th International Conference on LREC, Genoa, pp. 647-652, 2006.

[16] H. Schmid. Probabilistic Part-of-speech Tagging Using Decision Trees. In Proc. of the Conference on New Methods in Language Processing, Manchester, UK, pp. 44-49, 1994.