Investigating the Best Configuration of HMM Spanish PoS Tagger when Minimum Amount of Training Data Is Available*

Sergio Ferrández and Jesús Peral

Grupo de Investigación en Procesamiento del Lenguaje y Sistemas de Información
Departamento de Lenguajes y Sistemas Informáticos
University of Alicante, Spain
{sferrandez,jperal}@dlsi.ua.es

Abstract. One of the important processing steps for many natural language systems (information extraction, question answering, etc.) is Part-of-speech (PoS) tagging. This issue has been tackled with a number of different approaches in order to resolve this step. In this paper we study the functioning of a Hidden Markov Models (HMM) Spanish PoS tagger using a minimum amount of training corpora. Our PoS tagger is based on HMM where the states are tag pairs that emit words. It is based on transitional and lexical probabilities. This technique has been suggested by Rabiner [11] –and our implementation is influenced by Brants [2]–. We have investigated the best configuration of HMM using a small amount of training data which has about 50,000 words and the maximum precision obtained for an unknown Spanish text was 95.36%.

1 Introduction

Tagging is the task of classifying words in a natural language text with a respect to a specific criterion. PoS Tagger is the basis of many higher level natural language processing task. There are some statistical [2, 9, 10, 12] and knowledge-based [4, 7] implementations of PoS taggers, also there are some systems that combine different methods with a voting procedure [8]. Our implementation follows Brants [2].

One of the main sources of errors in Natural Language Systems is the incorrect resolution of PoS ambiguities (lexical, morphological, etc.). HMM as presented by Rabiner [11] are the standard statistical approach to try to properly resolve such ambiguities.

Unambiguously tagged corpora are expensive to obtain and require costly human supervision. Consequently, the main objective of this paper is to study the behavior of HMM applied to PoS tagging using a minimum amount of training data.

The next section shows our approach. Section 3 presents the evaluation of our approach and Section 4 shows some conclusions and further work.

* This research has been partially funded by the Spanish Government under project PROFIT number FIT-340100-2004-14.

A. Montoyo et al. (Eds.): NLDB 2005, LNCS 3513, pp. 341–344, 2005.

2 Our Approach: The HMM PoS Tagger

Tagging is the task of marking words in natural language text, the tagger has to choice a tag to unknown word from a defined finite tag set.

The forward-backward algorithm is used for unsupervised learning and relative frequencies can be used for supervised learning . The Viterbi algorithm [14] is used to find the most likely sequence of states for a given sequence of tags. Our implementation follows Brants [2].

The parameters of our model are initial state probabilities, state transition probabilities and emission probabilities: (a) π_i is the probability that a complete sentence starts at state s_i, $\pi_i = P(q_1 = s_i)$, where q_1 is an initial state. (b) $g_i(k)$ is the emission probability of the observed word w_k from state s_i, $g_i(k) = P(w_k = s_i)$. (c) a_{ij} is the transition probability from state i to state j, $a_{ij} = P(q_{t+1}|q_t = s_i)$, where q_t is the state in time t.

We have to compute this probability estimation for each initial state, transition probability or emission probability in the training corpora in order to generate the HMM.

Given a complete sentence (sequence of words $w_1...w_n$), we want to look for the sequence of tags T^* that maximizes the probability that the words are emitted by the model, we want to select the single most probable path through the model. This is computed using the Viterbi algorithm [14].

Mérialdo [9] shows that the Viterbi criterion optimizes sentence accuracy while the maximum likelihood criterion optimizes word accuracy. The maximum likelihood criterion selects the most probable tag for each word individually by summing over all paths through the model.

The next equation uses the Markov assumption that the transition and output probabilities only depend on the current state but not on earlier states.

$$T^* = arg \max_{t_1,...,t_n} \prod_{i=1}^{n} P(t_i|t_{i-1}, ..., t_1)P(w_i|t_i, ..., t_1) \tag{1}$$

$$\approx arg \max_{t_1,...,t_n} \prod_{i=1}^{n} P(t_i|t_{i-1}t_{i-2})P(w_i|t_i)$$

It is necessary to look for estimations that assign a part of the probability mass to the unseen events. To do so, there are many different smoothing techniques, all of them consisting of decreasing the probability assigned to the seen events and distributing the remaining mass among the unseen events. In our tagger the transition probabilities are smoothed using deleted interpolation [2].

$$P(t_i|t_{i-1}t_{i-2}) = \lambda_1\widehat{P}(t_i) + \lambda_2\widehat{P}(t_i|t_{i-1})\lambda_3\widehat{P}(t_i|t_{i-2}t_{i-1}) \tag{2}$$

For unknown words, a successive abstraction scheme is employed which look at successively shorter suffix. We borrow the calculation of the parameter θ from Brants and use a single context-independent value.

$$P(t|c_{n-i+1}, ..., c_n) = \frac{\widehat{P}(t, c_{n-i+1}, ..., c_n) + \theta P(t, c_{n-i+2}, ..., c_n)}{1 + \theta} \tag{3}$$

3 Evaluation

To realize the implementation of a PoS Tagger based on HMM we have to design tag set that show the syntactic category of a word in the context of a sentence and outstanding morphological information (our approach has 39 tags). We assume a trigram model, the states represent pairs of tags.

In order to ascertain the best configuration of HMM Spanish PoS Tagger, we have carried out different experiments when a minimum amount of training corpora is available.

3.1 Training Phase

Our system has been trained using a fragment of the *Lexesp* (CLiC- TALP Corpus) corpora [6] which contains 43 Spanish fragments (about 50,000 words) from different genres and authors. These corpora have been supervised by human experts and they belong to project of "Departamento de Psicología de la Universidad de Oviedo" and have been developed by "Grupo de Lingüística Computacional de la Universidad de Barcelona" with the collaboration of "Grupo de Procesamiento del Lenguaje de la Universidad Politécnica de Cataluña". Having worked with different genres and disparate authors, we felt that the applicability of our HMH PoS Tagger to other texts is assured.

In the experiments done using ten fold cross-validation a precision of **94.67%** was obtained. In order to find out the best precision of HMM Spanish PoS Tagger, we have done experiments using different values of parameters of equations used to calculate the probabilities of unknown words (see equation 3).

The best results for the different experiments, with a precision of **96.03%**, have been obtained when using 4 as a maximum suffix length and 0.5 as suffix backoff theta.

3.2 Evaluation Phase

In this section we have evaluated our approach. This task has been done using an unknown Spanish text which has about 10,000 words obtaining a precision of **95.36%**.

We have done experiments expanding the tag set until 259 which contains more morphological information that we will be useful for many Natural Language Systems. Obviously, we have observed that the precision diminish until 94.86%.

Generally, the obtained precision of the statistical models is between 95% and 97% (for example Freeling [5], Brill's Tagger [3], and Padró [10] for Spanish and TreeTagger [13] for English); on the other hand, the knowledge-based implementations (MACO [1] for Spanish) have a precision higher than 97%. Therefore, our approach proves to obtain a competitive result with a minimum corpora.

4 Conclusions and Further Work

We have proposed a Spanish Pos Tagger based on HMM that obtains competitive results using a minimum amount of training corpora which has 50,000 words.

Our computational system can be integrated inside other Natural Language Applications, due to PoS tagging is the basis of many higher level NLP tasks.

The main advantage of our Spanish tagger is to be able to obtain admissible results using a small training data. The evaluation result has been a precision of 95.36% using an unknown Spanish text.

As a future aim, we want to perform PoS Tagger of other languages (English, Italian, German, etc.) using the same core of HMM. Also, we plan to expand the morphological information of tags (with the use of dictionaries and rules).

References

1. J. Atserias, J. Carmona, I. Castellón, S. Cervell, M. Civit, L. Márquez, M.A. Martí, L. Padró, R.Placer, H. Rodríguez, M. Taulé, and J. Turmo. Morphosyntactic Analysis and Parsing of Unrestricted Spanish Text. *First International Conference on Language Resources and Evaluation, LREC'98*, pages 1267–1272, 1998.
2. T. Brants. Tnt- a statistical part-of-speech tagger. *Proceedings of the 6rd Conference on Applied Natural Language Procesing, ANLP*, pages 224 – 231, 2000.
3. E. Brill. Transformation-based error-driven learning of natural language: A case study in part of speech tagging. *Computational Linguistics*, 21:543 – 565.
4. E. Brill. A corpus-based Approach to Language Learning. 1993.
5. X. Carreras, I. Chao, L. Padró, and M. Padró. Freeling: An open-source suite of language analyzers. *Proceedings of the 4th International Conference on Languge Resources and Evaluation, LREC'04*, pages 1364 – 1371, 2004.
6. M. Civit. Criterios de etiquetación y desambiguación morfosintáctica de corpus en español. *PhD thesis, Linguistics Department,Universitat de Barcelona*, 2003.
7. W. Daelemans, J. Zavrel, P. Berckand, and S. Gillis. A memory-based part-of-speech tagger generator. *Proceedings of the 4th Workshop on Very Large Corpora*, pages 14 – 27, 1996.
8. G. Figuerola, F. Zazo, E. Rodríguez, and J. Alonso. La Recuperación de Información en español y la normalización de términos. *Revista Iberoamericana de Inteligencia Artificial*, VIII(22):135 – 145, 2004.
9. B. Mérialdo. Tagging English text with a probabilistic model. *Computational Linguistics*, 20(2):155 – 171, 1994.
10. M. Padró and L. Padró. Developing Competitive HMM PoS Taggers Using Small Training Corpora. *ESPAÑA for NATURAL LANGUAGE PROCESSING, EsTAL*, pages 127 – 136, 2004.
11. L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257 – 286, 1989.
12. A. Ratnaparkhi. A maximum entropy part-of-speech tagger. *Proceedings of the 1st Conference on Empirical Methods in Natural Language Procesing, EMNLP*, pages 16 – 19, 1996.
13. H. Schmid. TreeTagger — a language independent part-of-speech tagger. *Institut fur Maschinelle Sprachverarbeitung, Universitat Stuttgart*, 1995.
14. A.J. Viterbi. Error bounds for convolutional codes and asymptotically optimal decoding algorithm. *IEEE Transactions on Inf. Theory*, pages 260 – 269, 1967.