Spanish POS Tagger by Means of Hybrid Methods for the Accentuation of Words

Raymundo Montiel, Blanca E. Pedroza, Ma. Guadalupe Medina

Instituto Tecnológico de Apizaco
División de Estudios de Posgrado e Investigación
Avenida Instituto Tecnológico s/n, C.P. 90300, Apizaco, Tlaxcala México
mlirary@yahoo.com.mx, thelismedina@hotmail.com

Abstract. The task of Word Sense Disambiguation has as aim to identify the correct sense of a word in a context. The resolution of the ambiguity is a complex and useful task for many applications of natural language processing, for example: text categorization, machine translation, restoration of accents and, in general, in information retrieval. The process to accentuate words is also an ambiguity problem, because many words within the Spanish Language that can be marked or not depending on the context in which they are being used. This paper proposes a model that allows to obtain a correct accentuation of words with diacritical accent based on part of the speech tagging by means of the application of hybrid methods (supervised and unsupervised algorithms).

1 Introduction

One of the most difficult tasks and that has reached great interest within the *Natural Language Processing (PLN)*, takes place when several senses or meanings are associated with a word; this phenomenon of language is known as *polysemy*. The task of Word Sense Disambiguation consists of identifying the correct sense of a word in a context [10]. The problem of polysemy is closely related to the problem of the assignment of grammatical categories, which consists of saying if a word is a verb, an article or a noun [6], depending on the meaning that corresponds to that word in agreement to the context of the sentence.

The process of accentuate words is also a problem of ambiguation, because many words within the Spanish that they can be accentuated or not depending to different situations, such as the context, the time of action of the sentences, etc. The diacritical accent allows to distinguish words with identical form, that is, words written with the same letters, but that belong to different grammatical categories.

The lack of accents marks in some words within the sentences, is due to problems of ambiguation. The most common ambiguities in the accentuation of words are met between the words with endings in "o", as is the case of "completo" vs. "completó. They

© J. A. Carrasco-Ochoa, J. F. Martínez-Trinidad and H. Sossa (Eds.)
Special Issue in Advances in Pattern Recognition
Research in Computing Science 44, 2009, pp. 61-68

are the present and past tense, related to verbs with endings in "ar". There are other ambiguities purely semantic, including the nouns: "secretaria" (Person who takes charge writing the correspondence [8]) and "secretaría" (Section of an organization, institution or company[8]).
For this, we propose to develop a computer tool of semantic disambiguation for the Spanish Language, designed for the correct accentuation of words in written texts. This tool is based on the tagging of the sentences. For this goal, we have implemented a hybrid method.

2 Model for tagging of words

The purpose of this work consists of determining if a word with ambiguity on accent has to take an accent mark or has not, which is determined by the context in the one that word is dealing, with help of the assigned tags. For this part, we use a Hidden Markov Model (HMM).
The Models of Markov describe a process of probability which produces a sequence of events or not observable symbols. They are called "hidden" because of a process of underlying probability that is not observable, but it affects the sequence of observed events [7].
A HMM is characterized by a 5-tupla (Q, V, π, A, B), where:
Q is the set of states of the model. Though the states remain hide, they are known previously for the most of the practical applications. For the case of the labeling word, every label would be a state. Generally all the states are connected in such a way that any of them can be reached from any other one in an alone step. The states are labeled as {1, 2, ..., N}, and the current state in time t is denoted as q_t. In the case of the labeling word, we will not speak about the instants of time, but about the positions of every word inside the sentence.
V is the set of the different events that can be observed in each of the states. Each of the individual symbols that a state can emit is denoted as $\{v_1, v_2, ..., v_M\}$. In case of tagging word, M is size of dictionary and every v_k, $1 \leq k \leq M$, is a different word.
$\pi = \{ \pi_i \}$, is distribution of probability of initial state. Therefore,

$$\pi_i = P(q_1 = i), \qquad \pi_i \geq 0, \qquad 1 \leq i \leq N,$$
$$\sum_{i=1}^{N} \pi_i = 1 \tag{1}$$

$A = \{a_{ij}\}$, is distribution of probability of transitions between states, that is to say,

$$a_{ij} = P(q_t = j|q_{t-1} = i) = P(j|i), \qquad 1 \le i,j \le N, \qquad 1 \le t \le T,$$

$$\sum_{j=1}^{N} a_{ij} = 1, \qquad \forall\ i. \tag{2}$$

$B=\{b_j(v_k)\}$, is the distribution of probabilities of observable events, that is to say,

$$b_j(v_k) = P(o_t = v_k|q_t = j) = P(v_k|j),$$

$$b_j(v_k) \ge 0, \quad 1 \le j \le N, \qquad 1 \le k \le M, \qquad 1 \le t \le T \tag{3}$$

$$\sum_{k=1}^{M} b_j(v_k) = 1, \qquad \forall i.$$

3 Description of proposed solution model.

The general model proposed for the solution of the problem is showed in figure 1. The first task of the model is the lexical analysis, which consists of removing the accents of the words in the sentence of entry, thinking that for the calculation of the parameters of the model, words are needed without accents.

In addition, the lexical analyzer identifies and separates the signs of punctuation of the words, and then it identifies if ambiguous words exist inside the sentence of entry to the model and at the same time it indicates the position of every ambiguous word.

To identify the ambiguous words a comparison is done of each one of the words in the sentences, relating them to the words with ambiguity in the dictionary, which was constructed before.

For the phase of tagging in the model of solution (**Fig. 1**), we applying the modified Viterbi Algorithm [14], since all the posible states are not considered, that is to say, all the labels of the set of labels used, but only the most probable labels assigned to each of the ambiguous words.

To calculate the parameters of Hidden Markov Model we conducted a supervised and not supervised training (figure 2). For the phase of supervised training, we use the tagged corpus CONLL, it is collection of news articles by the Agency of News EFE in the year 2000. The parameter's model can be estimated by *maximum likelihood*, from the relative frequencies of appearance of the events in the corpus.

The parameters in matrix A, matrix B, and vector π, there were calculated taking the words without accents, since this is determined by the tags.

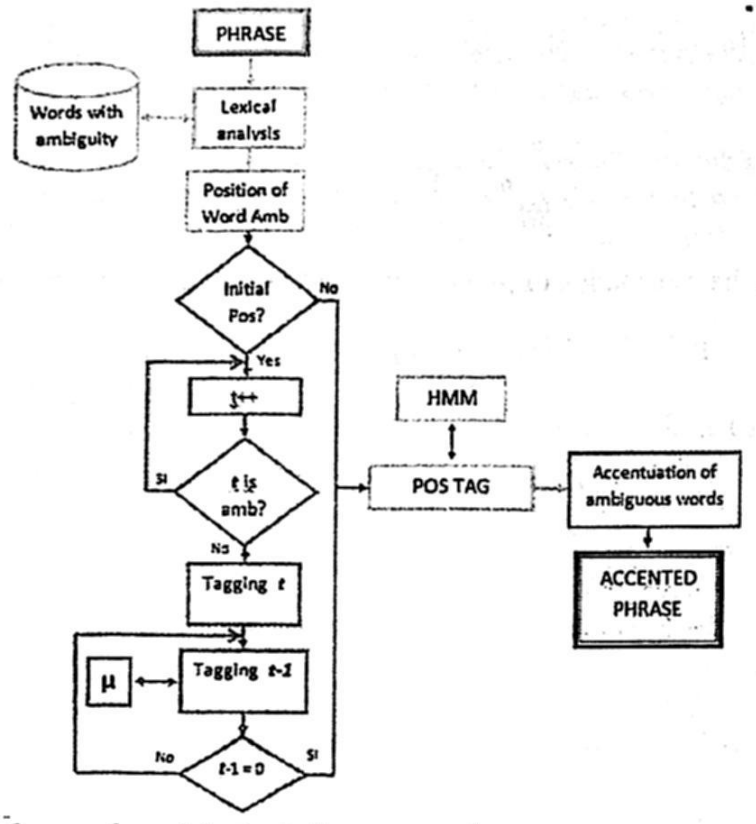

Fig. 1. Diagram of general model of solution proposed

For the matrix A, the probabilities of transition a_{ij} of the equation 2, are obtained counting how many transitions there are from state s_i to state s_j, and dividing by how many transitions there are by state s_i.
For the matrix B, the probabilities of emission, of the equation 3, are obtained counting how many emissions of symbol (v_k) are produced from state (s_j), and dividing by how many times that symbol has passed along by state s_j.
Once we have the initial parameters of μ (equation 1), we appy the algorithm Baum-Welch [1], which increases the probability of the transitions between the states and his symbols, so the probability is improved for the given sequence of observations.

3.1 Example applying the model

Considered the following sentence of entry:

"El jugo frío esta sobre la mesa"

In this sentence several words exist with diacritical accent, even to the beginning of the same one. In this case t words travels until finding a non ambiguous word, "frío". Once opposing the first non ambiguous word, it is assigned the most probable Tag.

After this, the label is assigned but probable t-1 from the already known tag (t) (fig.3), until arriving to t=1.

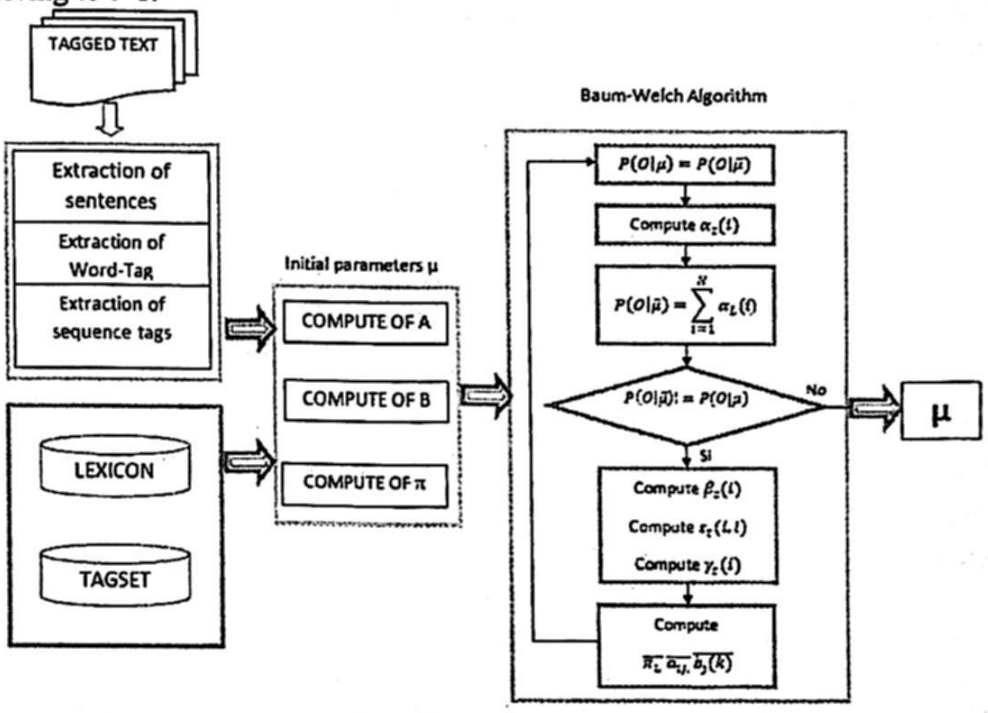

Fig. 2. Supervised and not supervised training

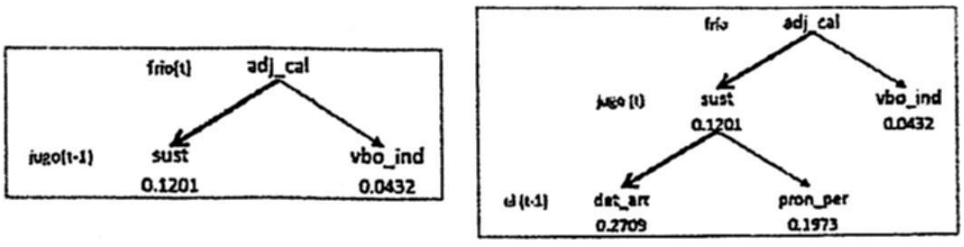

Fig. 2. Allocation of tags

Obtained the first tag of the sentence (if it is the case), the calculation to obtain the tag of the ambiguous words inside the sentence, it is realized only on the tags assigned manually to these words. Table 1 shows the sentence with assigned tag.

Table 1. Result of tagging

El	jugo	frío	esta	sobre	la	mesa
article	noun	adj_qua	verb	prep	article	noun

After having the tags of each one of the words of the sentence, we decide if the ambiguous words will have to or not to take accent, in agreement to our dictionary of ambiguous tagged words (Table 2).

Table 2. Example of ambiguous words together with its tags

Ambiguous word	Tag	Word
El	Article Pronoun	El Él
Esta	Det_dem Verb	Esta Está
jugo	Noun Vbo_ind	Jugo Jugó

Finally, we do a comparison of the words of the sentence together with the assigned tag and we determine if the word has to or not to take accent. Table 3 shows the result of our example.

Table 3. Result of the example.

El	jugo	frío	está	sobre	la	mesa

Besides accentuating the words with diacritical accent, the aim is to accentuate the words that always are accentuated, that is to say those words that must take accent mark. For this a database of accentuated words was realized.

4 Results

For the evaluation of the model, we met a collection of 63 texts of different contexts (Sports, Science, Health, etc.) obtained from the digital newspaper "El Universal[1]", these articles were evaluated by the model from complete phrases. Table 4 shows a summary of the obtained results. Here we have:

1 Number of Text
2 Number of words for Text
3 Number of words with diacritical accent
4 Number of words with diacritical accent, without accent.
5 Number of words with diacritical accent, without accentuating.
6 Number of words with diacritical accent, with accent.
7 Number of words with diacritical accent, accentuated.
8 Number of words without diacritical accent
9 Number of words without diacritical accent, accentuated.

[1] http://www.eluniversal.com.mx/noticias.html

Table 4. Summary of obtained results

1	2	3	4	5	6	7	8	9
1	164	26	26	26	0	0	12	12
7	166	24	22	22	2	2	14	8
10	214	30	26	26	4	3	22	21
19	376	54	49	49	5	3	26	23
24	1180	215	191	170	22	19	125	101
47	322	69	67	66	2	2	26	23
48	384	87	79	78	8	7	23	20
53	898	187	164	162	23	21	78	74
58	114	78	68	63	10	8	34	30
63	165	34	28	26	6	6	21	21
RESUL	22883	4238	3769	3611	472	397	1721	1526

With this information we deduce that:

$$Ambiguous\ words = \frac{4238*100}{22883} = 18.52\%$$

Besides these results, the information in the table 5 is evaluated using the metric of efficiency and obtaining the percentages of Recall (r), Precision (p), Mistake (m), Accuracy (a), and F_m.

Table 5. Results

Based on accented words		Based on non-accented words
r = .7153	m = 0.0549	r = .9796
p = .8411	a = .945	p = .9580
f-measure = .7731		f-measure = .9687

In the first part of the table 5, we show the results obtained from the words accentuated by our model, where the *precision* (number of words accentuated divided by the total number of words with diacritical accent) takes 84.11% as a result.

We have identified that one of the problems is in the tagging of the word "el", due to the fact that if the system finds this word at the beginning of the sentence, it assigns to the word the tag of "personal pronoun" and accentuates it, nevertheless, also it can take the tag of "article", in which case does not take accent.

As we mentioned previously, this model also determines when a word with diacritical accent must not take orthographic accent. The second part of the table 5 shows the results obtained from the words that must not take orthographic accent, in this case the *precision* (number of words that the model decide not to accentuate, divided by the total number of words with diacritical accent) takes 95.8% as a result.

Nevertheless, bearing in mind the result of *accuracy* (number of correct decisions done by the model) is 94.5%, having a mistake of 5.49%.

5 Conclusions

Tagging of words is a technique that can help to identify ambiguity in the accentuation of words of a certain text and it can generate computational tools that can help to the correct accentuation of texts in Spanish, being useful to make easier the writing of documents in Spanish or to use it as an assistant to teach grammatical rules.

References

1. Baum, L. Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Stat.* vol. 37, 1554-1563, 1966.
2. Bobiceva, V. (2008). O altă metodă de restabilire a semnelor diacritice. In Pistol I., Cristea D. Tufiş D. (eds.): Resurse Lingvistice şi Instrumente pentru Prelucrarea Limbii Române, pp. 179-188.
3. Dempster A., Laird, N., et al. Maximum likelihood from incomplete data via the EM algorithm. Journal of Royal Statistical Society. Series B (Methodological), vol. 39, No. 1, 1-38, 1977.
4. Marty, F. (1992). Trois systémes informatiques de transcription phonétique et graphémique. En Le Francais Moderne, pp. 179-197.
5. Mihalcea, R. (2002). Diacritics Restoration: Learning from Letters versus Learning from Words. In Proceedings of CICLing, pp. 339-348.
6. Perea Saedón, José Ignacio. (2005). Etiquetado de textos y su aplicación a la traducción. Unplublished monitired research. University of Granada.
7. Rabiner, L. R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. IEEE, vol. 77, no. 2, pp. 257-286.
8. REAL ACADEMIA ESPAÑOLA: Banco de datos (CORDE) [en línea]. Corpus diacrónico del español. <http://www.rae.es> [29 Abril 2009]
9. Rivest, R. L., Learning decision list, Machine Learning, 2, 229-246, 1987.
10. Stevenson, M; Y. Wilks. Combining independent knowledge sources for word sense disambiguation. En R. Mitkov (ed), Recent Advances in Natural Language Processing, John Benjamins Publisher, 2000.
11. Wagacha, P., De Pauw, G., et al. (2006). A Grapheme-Based approach for accent restoration in Gikuyu. In Proceedings of the Fifth International Conference on Language Resources and Evaluation, pp. 1937 – 1940.
12. Yarowsky, D. Decision list for lexical ambiguity resolution: Application to accent restoration in Spanish and French. In Proceeding of the 32th Annual Meeting of the Association for Computational Linguistics, (ACL'94). 1994. 88-95.
13. Yarowsky, D. A comparison of corpus-based techniques for restoring accents in Spanish and French text. To appear in Proceedings, 2nd annual Wordshop on Very Large Text Corpora, Kyoto, Japan. 1994
14. Viterbi, A. (2006). A personal history of the Viterbi algorithm. Signal Processing Magazine, IEEE, vol. 23, no. 4, pp. 120-142.