Sepe: A Spanish Texts POS Tagger

Héctor Jiménez[1] and Guillermo Morales[2]

[1] Facultad de Ciencias de la Computación, Universidad Autónoma de Puebla
C.U. 72570, Puebla, México
hjimenez@fcfm.buap.mx
[2] Programa de Simulación Molecular, Instituto Mexicano del Petróleo,
on leave of absence from Computer Science Section, CINVESTAV, México.
gmorales@cs.cinvestav.mx

Abstract. We describe a part-of-speech tagging system specially designed to tag Spanish texts using small linguistic resources. Nevertheless, the tagger obtains encouraging results. We have found and exploited useful contextual parameters to tag ambiguous and unknown words. Our tagger is mainly supported by word lists and one corpus with around 10^4 words. The system has been tested for texts of the so called "news" genre and is still on continuous development.
Keywords: Spanish language, part-of-speech tagging.

There are several part-of-speech (POS) taggers for the Spanish language ([6], [9], [11]). Even when the tagging performance is the most important matter, when building a tagger the size of available linguistic resources and the complexity of all involved parameters are quite relevant. In case of restricted domains it might be convenient to develop more precise taggers, as is suggested in [7]. We think that our experience is useful in new taggers development. Our tagger, named Sepe, has been easily implemented since it does not require neither copious resources at the beginning nor so much programming effort. The available corpus influenced strongly Sepe's implementation. We have started from a set of small resources and we followed simple criteria. The first design criterion for Sepe was to tag well-known words; the tagging of uncertain words was considered later. From the right beginning we collected a list of relevant words, a list of suffixes, a set of conjugation rules and a corpus composed by texts of "news" genre with around 10^4 words, extracted from *Corpus del Español Mexicano Contemporáneo* (CEMC) [5]. Sepe refines its criteria at each step. In the last steps, Sepe is strongly supported by a supervised learning method, applied to word contexts. The corpus is essential to determine the most important features of ambiguous and unknown words contexts. By aid of the corpus some patterns leading to additional morphosyntactic rules are identified. Also, word endings alleviate the small number of possible contexts in our corpus without restricting word tags.

This paper is divided in five sections. In section 1 some notation and the learning algorithm to choose the POS tag for some ambiguous and unknown words is introduced. Section 2 describes the resources used in the tagger system.

A tagging example is presented in section 3. In section 4 a performance test is shown. At the end the conclusions appear.

1 Background

The POS of a word is a tag on the set {`VERB, NOM, ADJ, ADV, CONJ, ART, PRON, CONT, NP, NUM`}, corresponding to *verbal form, noun, adjective, adverb, conjunction, article, pronoun, contraction, proper noun*, and *number* respectively, or punctuation signs as *comma, period*, etc.: `COM, PTO, PTC, DPT, INT, AIN, ADM, AAD, GUI,` and `SUS`. A text $T = [w_i]_i$ is a sequence of words pertaining to a vocabulary: for each i, $w_i \in \mathcal{V}$. A representative text for certain linguistic phenomena is called *corpus*. For each word w, we will denote its *ending* of length k as $w_{\_k}$. A *context* for w occurring in a text T, say at position j, is a subsequence $\bar{w}_{j-p} \dots \bar{w}_{j-1}\ \bar{w}_{j+1} \dots \bar{w}_{j+q}$, with $p, q > 0$, where each $\bar{w}_i$ is either a word in T, or a feature as an ending or a tag. A *dictionary* is a collection of words or suffixes. For each element in a dictionary a POS is assigned.

We distinguish several types of words with respect to a given corpus: a *definite word* has an invariable POS in all contexts, e.g. the article "el"; an *ambiguous word* has at least two contexts with different POS, e.g. "la" can have POS `PRON` (as in "*yo la amo*") or `ART` (as in "*la novia*"); an *unknown word* neither appears on the dictionary nor satisfies any rule related to its endings. The 100 *most frequent words* (MFW), taken from the analysis carrying out in the CEMC [5] are furtherly classified: the *definite frequent word* are the MFW that are definite words, the *ambiguous frequent word* are MFW ambiguous words, and the *frequent verbal forms*, which are conjugations of verbs in MFW. If a frequent verbal form is also ambiguous, then it will be taken as an ambiguous frequent word.

1.1 Learning Algorithm

Contexts, words and features The contexts around an ambiguous or unknown word help us to determine the word tag. In order to face this task, we will consider a *training set* of contexts S, whose elements are thus *training instances*, and their features. Let us go into some technicalities in order to introduce the measures that will allow us to choose the attributes in contexts useful in determining word tags.

Let $\mathcal{A} = \{A_1, \dots, A_m\}$ be a collection of attributes, for each $A \in \mathcal{A}$ let $D(A)$ be its domain (set of features) and let $U = \prod_{A \in \mathcal{A}} D(A)$ be the universe of instances. Given any set of training instances $S \subset U$, for any $A_i \in \mathcal{A}$ and any $a \in D(A_i)$, let $S_{A_i \leftarrow a} = \{X = (x_1, \dots, x_m) | x_i = a\}$. With respect to a given collection of classes $\mathcal{C} = \{C_1, \dots, C_n\}$, where $\forall j,\ C_j \subset U$, the *entropy* of S is

$$info_{\mathcal{C}}(S) = -\sum_{j=1}^{n} \text{freq}_j(S) \cdot \log_2 \text{freq}_j(S) \tag{1}$$

where $\text{freq}_j(S) = \frac{\#(S \cap C_j)}{\#(S)}$ is the "relative frequency of elements in S that fall in class C_j". For each $A \in \mathcal{A}$ let $N_{S_A} = \frac{\#(S_{A \leftarrow a})}{\#(S)}$ and

$$\textit{information gain}: \ \textit{gain}_{\mathcal{C},A}(S) = \textit{info}_{\mathcal{C}}(S) - \sum_{a \in D(A)} N_{S_A} \textit{info}_{\mathcal{C}}\left(S_{A \leftarrow a}\right) \quad (2)$$

$$\textit{split info}_A(S) = -\sum_{a \in D(A)} N_{S_A} \log_2 N_{S_A} \quad (3)$$

$$\textit{gain ratio}_{\mathcal{C},A}(S) = \frac{\textit{gain}_{\mathcal{C},A}(S)}{\textit{split info}_A(S)} \quad (4)$$

($\textit{split info}_A(S)$ does not depend on $\mathcal{C}$). The weights to be used as contextual features are given as $p_i = \textit{gain ratio}_{\mathcal{C},A_i}(S)$.

MBL We use Memory-Based Learning (MBL) [1] to classify words. Since it is a supervised learning method, it requires a collection of instances in order to classify any new instance: MBL assigns the new instance to the class of the most likely instance from the training set, or equivalently, to the class of the "closest" training instance towards the new instance. Hence, a distance function should be used in the classification. Let us introduce suscintly the formal details:

Suppose fixed a set of training instances S and a current partition $\mathcal{C}$ of classes. Given two instances $X = (x_1, \ldots, x_m)$, $Y = (y_1, \ldots, y_m) \in U$ let $\Delta(X,Y) = \sum_{i=1}^m p_i \cdot \overline{\delta}(x_i, y_i)$, where $\mathbf{p} = (p_1, \ldots, p_m) \in (\mathbb{R}^+)^m$ is a vector of weights and $\overline{\delta}$ is a *complementary Kroenecker delta*: $\overline{\delta}(a,b) = \begin{cases} 0 & \text{if } a = b, \\ 1 & \text{if } a \neq b. \end{cases}$ ($\Delta : (X,Y) \mapsto \Delta(X,Y)$ is a distance function realized as a *weighted average* of the "discrepancies" on attributes.) For any $X \in U$ let $\text{argmin}_{Y \in S} \Delta(X,Y)$ be any element in S that minimizes the map $Y \mapsto \Delta(X,Y)$ on S:

$$Y_0 = \text{argmin}_{Y \in S} \Delta(X,Y) \ \Leftrightarrow \ Y_0 \in S \ \& \ \forall Y \in S: \ \Delta(X,Y_0) \leq \Delta(X,Y).$$

Hence, any new instance X will be classified in the class of $\text{argmin}_{Y \in S} \Delta(X,Y)$, denoted from now on as $\text{Class}(\text{argmin}_{Y \in S} \Delta(X,Y))$.

In fact we may analyze also the context around an unknown feature: Given $X = (x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_m)$ and an attribute index $i \in \{1, \ldots, m\}$, a *context of* x_i is a substring c_i of X including x_i: $X = X_1 * c_i * X_2$, for some possibly empty strings X_1, X_2. For any possible value y_i of the i-th attribute let $c_i(y_i)$ be the string obtained from c_i substituting x_i by y_i. For any $Y = (y_1, \ldots, y_{i-1}, y_i, y_{i+1}, \ldots, y_m) \in S$ we shall estimate the probability that y_i appears in the context c_i of x_i. A measure of likeness of X to Y is

$$\overline{\delta}'(x_i, y_i) = \begin{cases} 0 & \text{if } x_i = y_i \\ 1 - \Pr(c_i(y_i)|c_i) & \text{if } x_i \notin D(A_i) \\ 1 & \text{otherwise} \end{cases} \quad (5)$$

MBL implementation A natural way to implement MBL is using a *trie*. The IGTree method [2] has the following characteristics:

- It compresses the training instances into a decision tree saving thus both search time and memory space.
- It faces the problem of exact matching failure –a feature value on the new instance which is not contained in any training instance– by using default information on the last non-terminal matching node. The default can be taken from "the most probable class for exact matchings", i.e. the feature values minimizing the gain-ratio values.

Several taggers following this approach have been reported in the literature. In [3] it is reported 97.8% in accuracy when tagging a text of 89×10^3 words using 711×10^3 training instances.

Certainly, MBL is an efficient and simple method quite adequate for natural language processing classification tasks (e.g. [12] [13]). In our approach, we have implemented a modification of IGTree [4]. We made the modification by means of a different distance function (represented by $\bar{\delta}'$), that was able to manage the unknown features into an instance. Now, the IGTree method is slightly modified:

1. If the current instance has an unknown feature, respect to the training set, with "high" gain ratio then using $\overline{\delta}'$ we select the closest class,
2. Else we proceed the classification as in IGTree.

$\bar{\delta}'$ improves the tagging of unknown words, and therefore the global tagging too.

2 Resources

The tagging system uses three types of resources: dictionaries, morphosyntactic rules and context instances. Some resources are corpus-independent, they are composed by conjugation rules, suffix lists and dictionary parts. The other resources are corpus-dependent. The main design criteria look to point the tagger to solve the most difficult problems, e.g. unknown words tagging. With this in mind, first of all it was carried out the tagging of simple words: definite words and words whose taggings were supported by morphosyntactic rules.

2.1 Dictionaries

As we have mentioned our tagger is based on a part of the CEMC. This text collection has 9,224 words, 2,644 distinct signs and an average of 2.02 tags per word. Using the corpus ambiguous, definite or unknown words were identified and put in corresponding dictionaries: DICCD, for frequent definite words[1]; DICCA, for frequent ambiguous words; DICCV, for frequent verbal forms; and DICCS joins the punctuation signs. Furthermore, an auxiliary source[2] to process manually a list of proper names, NOMP, was used. Table 1 lists the used dictionaries.

[1] DICCD was enriched with additional frequent definite words of the corpus with average rank 46.71.

[2] A collection of 127 articles from the mexican magazine **PROCESO** of 1998.

Dictionary	Size	Word type	Occurrence (%)
DICCD	184	Frequent definite	28.38
DICCA	163	Frequent ambiguous	17.59
DICCS	11	Punctuation	11.20
DICCV	835	Frequent verbal form	5.20
NOMP	3,337	Proper noun	2.08
Total	**4,530**		**64.47**

Table 1. Known words and its appearing percentage in the corpus.

2.2 Morphosyntactic Rules

According to the criteria cited above, it was carried out a general exploration on the corpus in order to determine the tags for some ambiguous frequent words and unknown words using morphosyntactic rules. The main parameter was the probability of occurrence of the word in the context:

> If $\Pr(tag(w) = m|C) = 1$, where C is a predetermined context, then we may conclude $C \Rightarrow (tag(w) = m)$.

Two examples that satisfy the preceding assumption are the following:

$$\Pr\left((m_{i-1}, v_i, m_i) \in \{\texttt{PREP}\} \times loas \times \{\texttt{ART}\} \middle| (m_{i-1}, v_i) \in \{\texttt{PREP}\} \times loas\right) = 1$$
$$\Pr\left((v_{i-1}, m_i, v_{i+1}) \in \{el, al\} \times \{\texttt{NOM}\} \times ddel \middle| (v_{i-1}, v_{i+1}) \in \{el, al\} \times ddel\right) = 1$$

where $ddel = \{de, del\}$, and $loas = \{la, las, lo, los\}$.

The verbal endings set from the COES spelling system [10] was used. This is a function that maps a conjugation ending into several possible infinitive endings. If the non-ending part of a supposed verbal form concatenated with an obtained ending matches an infinitive verb, then the original word is a right verbal form. A list VERBO of verbs in infinitive form is required and is provided by COES. The noun endings were selected from the inventory contained in [8]. It is represented by TERD and contains the following Spanish endings: “acia”, “ad”, “amento”, “amiento”, “ancia”, “anda”, “ato”, “encia”, “icia” “idumbre”, “ón”, “tad”, “tura” and “ud”. It is assumed that no ending in this list coincides with an ending of a verbal conjugation. Let $\mathcal{M}$ be the set of verbal endings. Let TERC be the intersection of both sets TERD and $\mathcal{M}$. It contains the endings: “ías”, “ado”, “ido”, “ía”, “to”, “so”, “es”, “as”, “o”, “a”, “era”, “ijo”, “iño”, “ite” and “uelo”, and were collected to identify the ambiguity `NOM/VERB`. Actually, when a word has an ambiguous ending of this type it is provisionally tagged `VERP`: *probable verb*. Besides, a regular expression detects some clitics. Indeed, they are represented in reversal order:

```
^s?[oa]l(son|e[mts])r[\'a\'e\'{\i}]|
^s?[oae]lr[aei]|
^s?([oae]l|[oa]l(e[mts]|son))odn\'a|
^s?([oae]l|[oa]l(e[mts]|son))odn\'ei|
^s?([oae]l|[oa]l(e[mts]|son))odn\'e
```

2.3 Contexts

For ambiguous, unknown and the frequent ambiguity (NOM/VERB) words the method MBL is applied, using the training set from the corpus. The features to be considered should be selected. Fig. 1 shows the gain ratio curve corresponding to 19 endings of length 6 at each side of an ambiguous word and, consequently, to the same number of tags surrounding the ambiguous frequent word. After performing this analysis for each of the three types of words the following features were selected:

Ambiguous frequent words: Two word endings from words which are at each side of the ambiguous word, as well as the tags of those words.

Ambiguity NOM/VERB: Two word endings from words which are at each side of the word with ambiguity NOM/VERB, as well as the tags of those words.

Unknown words: Two word endings immediate before the unknown word as well as the tags of those words.

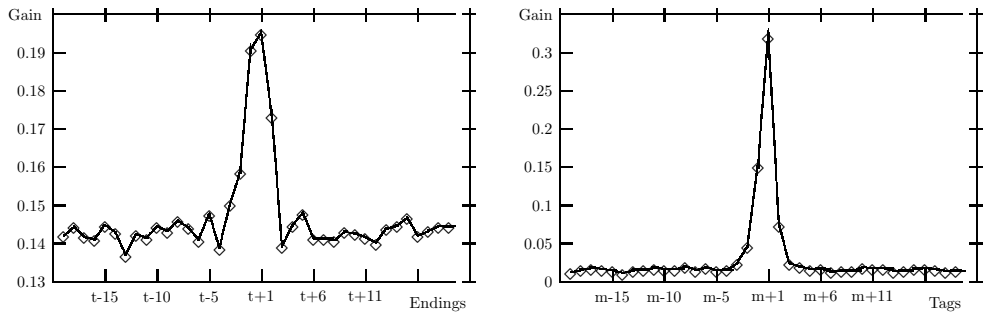

Fig. 1. Gain ratio curve for 19 elements around of an ambiguous word.

Let INSTA, INSTN and INSTD be the training sets for ambiguous, NOM/VERB resolution and unknown words, respectively. The required resources as a whole were:

1. Dictionaries
2. Morphosyntactic rules
3. Verb conjugation rules
4. Instances of ambiguous frequent words
5. Instances of NOM/VERB ambiguity
6. Instances of unknown words

The tagging steps followed the same order as listed above, see fig. 2.

3 Tagging Example

Figure 3 contains a text[3] used as a test for the tagger. Table 4 is the final result of the tagging. There is a row of text followed by a row of tags and an index (#)

Let inflex : $\mathcal{M} \rightarrow 2^{\mathcal{M}}$ be the function that gives a set of endings from an initial ending taken from a supposed verbal form, and let T be an untagged text. On each step we setup T:

step 1 For each member w_i of T:
If w_i is in DICCD, DICCV, DICCS, or NOMP, then define its tag m_i as the dictionary that contains w_i indicates.
step 2 For each non tagged member w_i of T:
If any morphosyntactic rule is able to be applied to a context of w_i, then define m_i according to that rule.
step 3 For each non tagged member w_i of T:
If there exists k such that $w_{i\_k} \in$ TERC, then $m_i =$ `VERP`.
Else if there exists k such that $w_{i\_k} \in \mathcal{M}$, $w_i = xw_{i\_k}$, and for some $y \in \text{inflex}(w_{i\_k})$ it holds $xy \in$ VERBO, then $m_i =$ `VERB`.
step 4 For each non tagged member w_i of T, and member of DICCA:
Let X be the context of w_i. Use INSTA to define $m_i = \text{Class}(\text{argmin}_{Y \in S} \Delta(X, Y))$.
step 5 For each member w_i of T tagged with `VERP`:
Use INSTN and apply IGTree to define m_i.
step 6 For each non tagged member w_i of T:
Use INSTD an apply IGTree with $\bar{\delta}'$ to define m_i.

Fig. 2. Main tagging steps.

to be used as a reference. Finally, table 2 summarizes the accuracy of each step applied to the example text, making reference to the resource used.

Step	**# tags**	**# right tags**	**%**
Dictionaries	103	102	99.02
Morphological	17	17	100.00
Verbal forms	13	10	76.92
Frequent ambiguous	16	14	87.50
`VERB/NOM`	9	7	77.77
Unknown words	40	34	85.00
Total	**198**	**184**	**92.92**

Table 2. Accuracy from each step of the example text tagging.

[3] This is a part of the news written by Carlos Acosta Córdova y Guillermo Correa, and published by the magazine **PROCESO**, on may 1999.

y el beneficiario directo de todo ello será , sin duda , el PRI , según reconoció el secretario de hacienda ante analistas , académicos e inversionistas en el consejo de las américas de Nueva York , horas antes de participar en la reunión de los organismos financieros internacionales . dijo , sin ambages , que si la economía sigue bien y se mantiene la disciplina , el partido revolucionario institucional tendrá buenas posibilidades no sólo en la contienda por la presidencia , sino en la elección del gobierno capitalino y , aun , en recuperar la mayoría absoluta en la cámara de diputados . ... en efecto , en su examen anual de la economía mexicana , que hizo público el jueves 29 , la OCDE - el club de los 29 países más ricos del mundo , al que México ingresó en 1994 - admite que el desempeño económico del país fue positivo en los últimos tres años , pero señala una serie de ineficiencias en la política económica - inestabilidad presupuestal , dependencia del petróleo y deficiente sistema tributario - que impiden sacar al país del subdesarrollo y contrarrestar los niveles de pobreza extrema .

Fig. 3. Example text.

4 Performance

With a text of 9,000 words a test was carried out. The input text was divided into ten parts. When processing each part, the training text was increased by adding the former part. This is represented in the x-axis of the graphs in fig. 5. The experiments results are shown in the graphs. The y-axis represents the accuracy, and was calculated as the number of right taggings divided by the number of text words. The first graph contains the average of accuracy and the minimum and maximum of the tagging process in an error bar graph. The second graph compares the average accuracy using $\bar{\delta}$ and $\bar{\delta}'$. In both graphs the performance is shown as the training corpus grows.

5 Conclusions

A part-of-speech tagger for Spanish language based on small resources and minimum programming effort has been built. Such conditions may be available to develop a new tagger and speed up the initial stage.

The tagging accuracy of Sepe is greater than 0.9. Of course, the behavior of our tagger can be improved. It is remarkable that the low verbs tagging accuracy and the low `VERB/NOM` ambiguity resolution accuracy at the test text can be increased by updating the TERC list. The word "serie" at position 165 on the example text is tagged as `VERB` by the `VERB/NOM` ambiguity solving procedure. However, this procedure only uses 515 instances (the least of the three training sets). Therefore, at the next stage we are considering:

- To increase the number of known words.
- To make use of derivational and inflectional lists to cope with partially known words. This should support the previous point.
- To increase the corpus to train the learning method as well as to grow the corpus valid-rules.
- To carry out performance test to compare to other methods. This requires to change the tag set and the corpus.

#	Text/Tags
1	y el beneficiario directo de todo ello será , sin
	CONJ ART NOM ADJ PREP ADJ PRON VERB COM PREP
11	duda , el PRI , según reconoció el secretario de
	NOM COM ART NP COM PREP VERB ART NOM PREP
21	hacienda ante analistas , académicos e inversionistas en el consejo
	VERB PREP NOM COM NOM CONJ ADJ PREP ART NOM
31	de las américas de Nueva York , horas antes de
	PREP ART NOM PREP NP NP COM NOM ADV PREP
41	participar en la reunión de los organismos financieros internacionales .
	VERB PREP ART NOM PREP ART NOM ADJ VERB PTO
51	dijo , sin ambages , que si la economía sigue
	VERB COM PREP NOM COM CONJ CONJ ART NOM VERB
61	bien y se mantiene la disciplina , el partido revolucionario
	ADV CONJ PRON VERB ART NOM COM ART NOM ADJ
71	institucional tendrá buenas posibilidades no sólo en la contienda por
	ADJ VERB NP NOM ADV ADV PREP PRON VERB PREP
81	la presidencia , sino en la elección del gobierno capitalino
	ART NOM COM CONJ PREP ART NOM CONT NOM ADJ
91	y , aun , en recuperar la mayoría absoluta en
	CONJ COM NOM COM PREP VERB ART NOM ADJ PREP
101	la cámara de diputados . / en efecto , en
	ART NOM PREP NOM PTO SUS PREP NOM COM PREP
111	su examen anual de la economía mexicana , que hizo
	ADJ NOM ADJ PREP ART NOM ADJ COM PRON VERB
121	público el jueves 29 , la OCDE - el club
	NOM ART NOM NUM COM ART NP GUI ART NOM
131	de los 29 países más ricos del mundo , al
	PREP ART NUM NOM ADJ NOM CONT NOM COM CONT
141	que México ingresó en 1994 - admite que el desempeño
	CONJ NP VERB PREP NUM GUI VERB CONJ ART NOM
151	económico del país fue positivo en los últimos tres años
	ADJ CONT NOM VERB VERB PREP ART NOM ADJ NOM
161	, pero señala una serie de ineficiencias en la política
	COM CONJ VERB ART VERB PREP NOM PREP ART NOM
171	económica - inestabilidad presupuestal , dependencia del petróleo y deficiente
	ADJ GUI NOM ADJ COM NOM CONT NOM CONJ NOM
181	sistema tributario - que impiden sacar al país del subdesarrollo
	ADJ ADJ GUI CONJ VERB VERB CONT NOM CONT NOM
191	y contrarrestar los niveles de pobreza extrema .
	CONJ VERB ART NOM PREP NOM VERB PTO

Fig. 4. Example text with tags.

In spite of the fact that there is a small difference between $\bar{\delta}$ and $\bar{\delta}'$ we expect a greater improvement with a greater corpus. Further, a greater corpus might help to debug the ambiguity grammem lists as well as the morphosyntactic rules.

References

1. Daelemans, Walter: Memory-based lexical acquisition and processing, *Lecture Notes in Artificial Intelligence*, 898, Springer Verlag, pp 85-98, 1995.
2. Daelemans, Walter; Durieux, Gert & van-den-Bosch, Antal: Towards inductive lexicon, *Proc. of LREC Workshop on Adapting Lexical and Corpus Resources to Sublanguages and Applications*, Granada, http://ilk.kub.nl/, 1998.
3. Daelemans, Walter; van-den-Bosch, Antal; Zavrel, Jakub; Veenstra, Jorn; Buchholz, Sabine & Busser, Bertjan: Rapid development of NLP modules with memory-based learning, *Proc. of ELSNET in Wonderland*, pp 105-113, 1998.

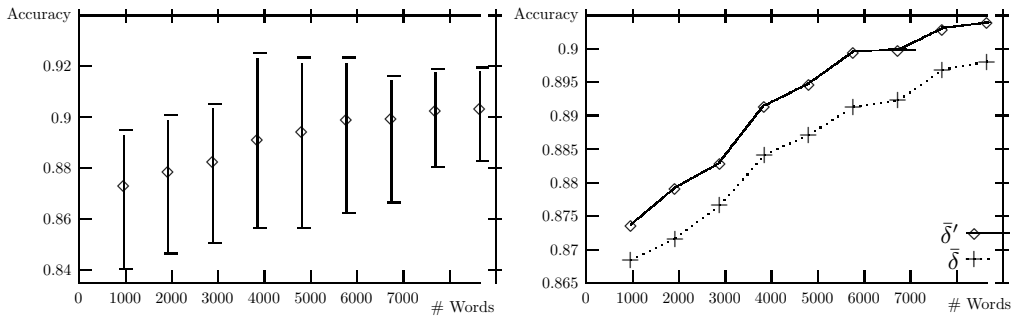

Fig. 5. Performance of the tagger.

4. Jiménez-Salazar, Héctor & Morales-Luna, Guillermo: Instance metrics improvement by probabilistic support, *Lecture Notes in Artificial Intelligence*, 1793, Springer Verlag, pp 699-705, 2000.
5. Lara, Luis Fernando; Ham-Chande, Roberto & García-Hidalgo, Ma. Isabel: *Investigaciones lingüísticas en lexicografía*, Jornadas 89, El Colegio de México, 1979.
6. Màrquez, Lluís & Rodríquez, Horacio: Part-of-speech tagging using decision trees, *Lecture Notes in Artificial Intelligence*, 1398, pp 25-33, 1998.
7. Marques, N. & Pereira, G.: A POS-tagger generator for unknown languages, *Procesamiento del Lenguaje Natural*, Rev. No. 27, SEPLN, pp 199-206, España, 2001.
8. Moreno de Alba, Jose G.: *Morfología derivativa nominal en el español de México*, UNAM, 1986.
9. Pla, F.; Molina, A. & Prieto N.: Evaluación de un etiquetador morfosintáctico basado en bigramas especializados para el castellano, *Procesamiento del Lenguaje Natural*, Rev. No. 27, SEPLN, pp 215-221, España, 2001.
10. Rodríguez, Santiago & Carretero, Jesús: Building a Spanish speller, http://www.datsi.fi.upm.es, 1997.
11. Ruiz, L.: *Desarrollo de un modelo computacional para el procesamiento de corpus textuales basado en la etiquetación automática*, Tesis doctoral, U. de Oriente, Cuba, 2001.
12. van-den Bosch, Antal; Daelemans, Walter; Weijters, Ton: Morphological analysis as classification: an inductive-learning approach, http://lib-www.lanl.gov/cmp-lg/9607021, 1996.
13. Zavrel, Jakub; Daelemans, Walter; Veenstra, Jorn: Resolving PP-attachment ambiguities with MBL, *CoNLL*, http://ilk.kub.nl/, 1997.