Instance metrics improvement by probabilistic support

Héctor Jiménez*[1] and Guillermo Morales[2]

[1] Facultad de Ciencias Físico-Matemáticas
Universidad Autónoma de Puebla
hjimenez@fcfm.buap.mx
[2] Centro de Investigación y Estudios Avanzados del IPN
gmorales@cs.cinvestav.mx

Abstract. The use of distance functions in order to determine nearest instance class at Memory Based Learning methods may be crucial when there are no exact matchings. We add relative information over unknown feature values to improve the information extract on the training instances. An experiment was carried out for Spanish Part-Of-Speech tagging of unknown words finding a better performance with our modified function.
Keywords. Memory Based Learning, inexact matching, Spanish Part-Of-Speech tagging, unknown words.

1 Introduction

Memory Based Learning (MBL) has had many applications on natural language processing [3]. This is a powerful and very simple method, that preserves the regularities and the exceptions on the training instances with a low cost of implementation. We have chosen MBL to deal the problem of determining the Part-Of-Speech (POS) tags of Spanish words from the contexts in which they occur.

Several high quality taggers for English [1, 4, 10] have been proposed. Most of them have contextual information as their main support. Nevertheless, on Spanish language it is very important to consider the morphosyntactic information. The POS tagger due to Màrquez and Rodríguez [7] has meant an important advance on Spanish POS tagging. In any natural language, the problem of unknown words tagging deserves a special treatment at tagging irrestricted text. Here, we present the results of a metric used for MBL, which improves the performance of unknown words tagging.

In our POS tagger for Spanish, we use MBL to treat the tagging that requires contextual information. In general, our tagger has three steps: It tags the most frequent unambiguous words, then it tags the most frequent ambiguous words and, finally, it tags the unknown words. At the first step, a dictionary

* This work was supported by Mexican CoNaCyt under grant nr. 50084 and the Posgrado de Matemáticas, Facultad de Ciencias Físico-Matemáticas, UAP.

with morphosyntactic information is necessary. The second step requires the morphosyntactic information and a bit more of contextual information. And the third step uses mainly contextual information, but this step is very sensitive to the unknown feature values on the test instance.

In order to tag an unknown word regarding its context, we use morphosyntactic information to determine an unknown feature value, and to achive a full distance computation. This improves the class selection whenever an exact matching fails. In an experiment, we tested the accuracy of determining POS for unknown words using the IGTree version of the MBL method and the enriched distance function.

The second section of the present paper is dedicated to analyse the metric used in MBL. In the third section, our tagger is described. An experiment carried out to compare the IGTree method with the IGTree using the new metric is presented in the fourth section. At the end, the conclusions of the present work are presented.

2 MBL and the Modified Metric

Daelemans proposed the use of MBL on natural language processing problems, and has reported important results [3]. The method can be summarized as follows: The method takes a new instance to be classified and, from a set of classified training instances, it selects the class of the nearest instance of the training set, assigning indeed this class to the new instance. The method is precise when the notion of distance between instances is defined. As an example of such distance consider the following: Given two instances $X = (x_1 \dots x_m)$ and $Y = (y_1 \dots y_m)$, let

$$\Delta : (X, Y) \mapsto \Delta(X, Y) = \sum_{i=1}^{m} \delta(x_i, y_i) \cdot w_i \tag{1}$$

where

$$\delta(a, b) = \begin{cases} 0 \text{ if } a = b \\ 1 \text{ otherwise} \end{cases} \tag{2}$$

and w_i is the *information gain* corresponding to feature i, thus, according to its importance, w_i contributes to increase the Δ value when the i-features differ. The IGTree version [2] of the MBL method compresses the training instances into a decision tree saving thus the search time and memory space. Additionaly, the MBL method 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 matching non-terminal node. This default can be taken from the most probable class given, i.e. the most probable class for exact matchings of feature values arranged in decreasing order according to information gain. We call *maximum substring* to the string composed of these feature values.

The proposed modification to the inexact matching consists in pursuing the search even if the matching fails. In our approach, we assign a known feature value to the unknown value that occurs in the used training instances. The known feature value is supplied by *n-grams* calculated from the corpus context. Namely, let X be a new instance, Y an instance from the training base and let $\mathbf{dom}(A_i)$ be the domain of the i-attribute from the training instances. The estimation of a known feature value uses the modified distance function:

$$\delta'(x_i, y_i) = \begin{cases} 0 & \text{if } x_i = y_i \\ p_i & \text{if } x_i \notin \mathbf{dom}(A_i) \\ 1 & \text{otherwise} \end{cases} \tag{3}$$

where $p_i = 1 - \Pr(c\, y_i|c)$, and c is a context of x_i. Strictly speaking, the resulting function Δ is not a metric since it lacks the "symmetric property" of metrics (if a pair is conmuted Δ may be undefined), nevertheless it can be extended to a metric. With this in mind we will call Δ a metric as well. Certainly, we use *contexts* around each word to be tagged, moreover we use contexts around the unknown feature values to estimate known feature values i.e. calculating the probability of y_i occurring in the context of x_i, see figure 1.

training instance	$y_1 \ldots$	$\ldots y_{i-1}\ y_i\ y_{i+1} \ldots$	y_m
new instance with an unknown feature value	$x_1 \ldots$	$\ldots x_{i-1}\ x_i\ x_{i+1} \ldots$	x_m
context of x_i in the new instance		$x_{i-k} \ldots x_{i-1}\ x_i\ x_{i+1} \ldots x_{i+k}$	

Fig. 1. Context of an unknown feature value taken from the new instance.

The use of p_i resembles the use of information gain. If the feature value y_i is the most probable one to be in the new instance context, then this value will provide a low weight for the distance computation. This new weight will modify the distance value according to the probability-complement of finding a known feature value in the new instance.

Now, the IGTree is modified making use of δ' as follows: Given a new instance X and a training base,

1. If any of feature value of X with the highest weights is unknown by the training instances, then we apply δ' in order to select the nearest instance class.
2. Otherwise, we continue the IGTree computation.

Due to the fact that $\Pr(c\ y_i|c)$ has an inverse relation with δ', the first step of the above procedure chooses the class of that instance which contains the feature value y_i with maximum probability, because this y_i is such that $\Pr(c\, y_i|c)$ minimizes the distance.

Table 1 shows a very simple example of the trigrams probability, in where `SUST`, `ADV`, `NUM`, `PRON`, `VERB`, `NP` and `CONJ`, stand for noun, adverb, numeric value, pronoun, verb, proper noun and conjunction, respectively. Suppose that:

- a new instance, X, has one of the highest weights attribute, say the second attribute,

– the second feature value in X is `ADV` and
– `ADV` $\notin \mathbf{dom}(A_2)$.

In this case, the IGTree method uses the most probable class to assign the corresponding POS tag to X, whereas the modified IGTree method with δ' makes the matchings with all instances as if the second attribute would be known but weighting it now with p_2. Thus, if

$$Y = (\texttt{dor SUST del CONT})$$

is a training instance to tag an unknown word, and the new instance is:

$$X = (\texttt{dor ADV , COM})$$

with the unknown second feature value context :

$$c = (\ldots\ \texttt{de dor ,}\ \ldots),$$

then we use table 1 because we have supposed `ADV` $\notin \mathbf{dom}(A_2)$, and we obtain $\delta'(x_2, y_2) = 0.333333$. Of course, this process is possible on taking into account the trigram information: whenever a feature value is not in the corresponding attribute domain we try to use the trigram probability according to the feature value context to compute the distance value.

end_{-1}	end_1	x	$\Pr(end_{-1}\ x\ end_1 \mid end_{-1}\ end_1)$
de	,	SUST	0.666667
de	,	ADV	0.025641
de	,	NUM	0.0512821
de	,	PRON	0.025641
de	,	VERB	0.0512821
de	,	NP	0.128205
de	,	CONJ	0.0512821

Table 1. Conditional probability to find some feature value between two word endings.

3 The Tagging Process

We have begun the task of building a POS tagger with very limited linguistic resources. A corpus[1] of 10,000 tagged words has been used at this first stage of development. The tagging process was divided on three parts according to the types of words: defined frequent words (words having just one tag in any context), ambiguous frequent words (words having at least two different contexts with different POS) and unknown words (all remaining words). The tagger should accomplish the following tasks:

[1] This is a part of *Corpus del Español Mexicano Contemporáneo* [6], kindly supplied by El Colegio de México.

1. To tag the defined words.
2. To disambiguate ambiguous words.
3. To determine the POS of unknown words.

At each step, the tagger uses a special type of resource. The defined words tagging is based on a dictionary and morphosyntactic rules. The same is used for ambiguous words tagging plus the context of these words. For unknown words, the major support is provided by the context, only a portion of morphosyntactic information is still used.

The involved resources are the dictionary, the list of morphosyntactic rules and the extracted contexts from the corpus. The dictionary is structured by sections: defined frequent words, ambiguous frequent words, very frequent verbal forms, punctuation signs and proper nouns. The criteria for different word types inclusion are determined by the analysis of the frequent words[2].

The list of morphosyntactic rules includes the rules based on the conditional probability of trigrams, verb conjugation [9] rules for regular and irregular verbs, and rules to determine the POS of a word [8] from a suffix.

The contexts used were formed by tags and endings of the neighboring words of the corpus. The MBL method was applied to three different problems: to disambiguate ambiguous words, to disambiguate the verb/nominal ambiguity, and to determine the corresponding POS to unknown words. Using around 9,000 words, we obtained an average near to 94% on accuracy.

4 The Experiment

After the step of disambiguating the verb/nominal ambiguity, two tests were carried out to tag unknown words: the first one applying the IGTree method and, the second one applying the IGTree method with δ'.

In order to define the context c (referred in eq. 3 by p_i) of an unknown feature value contained in a new instance, an ending word of left and right contiguous words for the words whose feature value is unknown was used.

Each of the two tests was carried out by dividing the corpus on ten parts. One of them is chosen as the test text. The training text is successively increased with the other nine parts. Once the unknown words are tagged the accuracy, of the IGTree and IGTree with δ', is determining which is shown at Fig. 2.

Indeed, the probability p_i is better determined as the corpus grows. Therefore, the estimation of an unknown feature value will have higher accuracy with a larger corpus because it increases the number of contexts for each unknown feature value, and the computation of the probability is thus better precised. Also, as the corpus grows, the behaviour of the IGTree method, from the point of view of the maximum substring introduced at the begining of section 2, adds

[2] Our treatment is based on frequent word and considers the one hundred most frequent words. Thus, the result provided by Guiraud [5] holds: the words with ranked up to position 100 cover 60% of the lexicon.

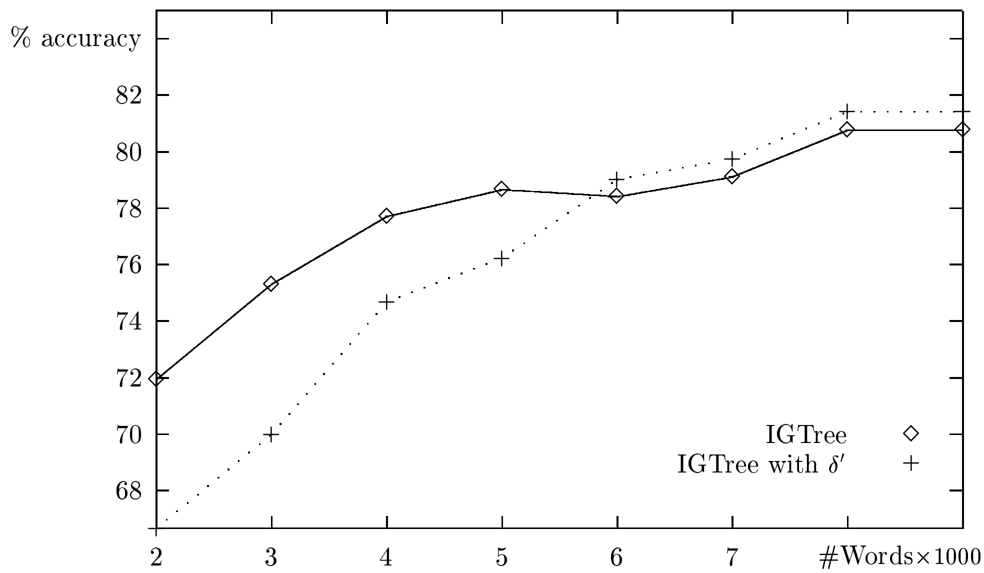

Fig. 2. Performance of IGTree and the use of δ' on unknown words tagging.

new substrings and precises the IGTree information gain. Nevertheless, no information on unknown features values like probability of trigrams is used.

The performance result in the experiment carried out shows that after 5,000 words of training corpus, the δ' modification of IGTree has a slightly higher accuracy than the IGTree pure method. It is necesary to carry out a new experiment with a larger corpus to determine a definitive significance of the δ'-IGTree on the IGTree method. Màrquez and Rodríguez show some performance data on unknown word problem, in POS tagging: their decision tree-based method reaches 83.3% on accuracy and the IGTree method 79.18%. But it is uncomparable with the δ'-IGTree method because this experiment was carried out under different conditions (e.g. tag-set and corpus).

5 Conclusions and Future Work

POS tagging of unknown words is a very important problem to solve in natural language processing. Also, it is necessary to offer a solution to unknown feature value problem, because learning procedures frequently faces this problem.

A modified metric to deal the problem of non exact matching into the MBL method has been presented. This metric was tested on unknown words POS tagging, using a POS tagger on development.

At the present time, we have an initial version of a POS tagger, still unpractical for real problems. The next step in this direction is to grow the corpus to 50,000 words. In spite of the fact that the corpus is very small, the present results allow us to explore about new metrics for the MBL method. The reason why the proposed metric improves the implicit metric on the IGTree method is due to the fact that the new metric is enriched with additional information.

In any particular case where some unknown feature value is met, the new metric requires a full computation of the distance, and this is clearly contrary to the more natural procedure of picking up one of the most probable classes from the IGTree method.

In the present application, the attributes to deal with an unknown feature value of any new instance were taken from the context in which the unknown feature value appears. This is a particular case. In general, research should be carried out in order to select the attributes depending on each purpose. Also, in order to reinforce the metric, a method should be used with a better efficiency with respect to the method offered by the probability of *n-grams*.

References

1. Eric Brill: Some advances in transformational-based part of speech tagging. In *Proceedings of 12th National Conference on Artificial Intelligence*, pp 722-727, 1994.
2. Walter Daelemans, Antal van den Bosch, Ton Weijters: IGTree: Using Trees for Compression and Classification in Lazy Learning Algorithms. In D. Aha (ed.) *Artificial Intelligence Review*, special issue on Lazy Learning, 1996.
3. W. Daelemans, A. van den Bosch, J. Zavrel, J. Veenstra S. Buchholz & B. Busser: Rapid development of NLP modules with memory-based learning. In *Proceedings of ELSNET in Wonderland*, Utrecht: ELSNET, pp 105-113 1998.
4. A. Franz: *Automatic ambiguity resolution in natural language processing, an empirical approach*, Springer Verlag, 1996.
5. Pierre Guiraud: *Les caractères statistiques du vucabulaire*, Mouton, 1954.
6. L.F. Lara, R. Ham Ch. & M.I. García: *Investigaciones lingüísticas en lexicografía*, Jornadas 89, El Colegio de México, 1979.
7. Lluís Màrquez and Horacio Rodríguez: Part-Of-Speech tagging using decision trees. In *Lecture Notes in Artificial Intelligence No. 1398*. ECML-98, Springer Verlag, pp 25-36, 1998.
8. José G. Moreno De Alba: *Morfología derivativa nominal en el español de México*. Ed. UNAM, México, 1986.
9. Santiago Rodríguez and Jesús Carretero: A Formal approach to Spanish morphology: the COES tools. From `http://www.datsi.fi.upm.es/ coes`, 1997.
10. R. Weischedel, M. Meteer, R. Schwartz, L. Ramshaw and J. Palmucci: Coping with ambiguity and unknown words through probablistic models. In *Computational Linguistics*, 1993.
11. Jakub Zavrel and Walter Daelemans: Recent Advances in Memory-Based Part-of-Speech Tagging. In *VI Simposio Internacional de Comunicación Social*, Santiago de Cuba, pp. 590-597, 1999.