Simple Features to Identify Tags in Named Entity Recognition for Spanish Texts

Marisol Fierro Ayón, Héctor Jiménez-Salazar &
José de Jesús Lavalle Martínez

Facultad de Ciencias de la Computación
B. Universidad Autónoma de Puebla
C.U. 72570, Puebla, México
Tel. (+52-222)2295500 Ext. 7212, Fax (+52-222)2295672
marisolfa@hotmail.com, hgimenezs@gmail.com, jlavalle@aleteya.cs.buap.mx

Abstract. In this work Named Entity Recognition (NER) using Memory-Based Learning (MBL) is presented. This application is based on several works that deal with this topic. Our contribution is the analysis of some feature sets, taken from POS, capitalized, context, and without external information sources, in order to constitute the training set for the learning method. In the experiments the corpus from the CoNLL-02 conference was used, and for tag identification 96.14% of precision was reached using just 14 features.

Keywords: named entity recognition, memory based learning.

1 Introduction

Named entities (NE) are phrases that contain persons (PER), organizations (ORG), locations (LOC), dates and quantity names (MISC)[1]. For example, in the following clause there are tags that identify the NE that occur in it:

> [PER Fernando Lozano], presidente de [LOC Valle Alto], llegó al [MISC XXI Torneo Universitario].
>
> *[PER Fernando Lozano], president of [LOC Valle Alto], arrived at the [MISC XXI Universitary Tournament].*

Named entity recognition enriches text representation, and it could be applied to tasks supporting Natural Language Processing. As an example, in Question-Answering Systems the responses to questions using pronouns such as where, who, etc. could be supported by NE.

In this work we are focussing on NE of the classes PER, ORG, LOC and MISC. A NE tagger using few linguistic resources and tools, but having a high degree of precision is of particular interest. Nevertheless, to recognize the whole NE occurring in a text is not very important to us. Since we need the building blocks to construct a NE database within of a journalistic navigational system; hence our interest on just precision.

There have been many works about NER, mainly in CoNLL meetings. Spedifically for Spanish, we can cite the works submitted to CoNLL-02 [2]. For example, Fien De Meulder & Walter Daelemans [3] analysed the influence in using external information sources for NER. They used the TiMBL [8] system for NER in English and German. Their system uses a training file and NE lists (*gazetteer*). Some of the features taken into account for the English language are: the context, parts of speech (POS), capital letter use, the first and last three letters of each word, and ten more features which indicate if a word belongs to some NE list. In the German language case, they also used the root of each word of the context.

Another work related with this is the one of Tjong Kim Sang [12], whom evaluated his strategies in language-independent NER tasks using the Memory-Based Learning (MBL) method. He considered as features: contextual words, POS, and morphological features (prefixes, suffixes, capital letters, etc.). The tests presented were obtained applying waterfall, feature selection and voting methods to Dutch and Spanish. The global performance for Spanish, measured with respect to F_1, was 75.78%.

Xavier Carreras *et al.* [11], also in CoNLL-02, got the highest efficacy, 79.38% for F_1. They employed many features: contextual (POS included), morphological, patterns of words, and predictor words. Also, they confirmed that external knowledge sources are not essential.

Thamar Solorio & Aurelio López [5] employed support vector machines (SVM). Unlike the last two works, the dimension of their representation space is very high, because of the combination of the labels that they used, speech parts, and label of the kind of NE (PER, LOC, etc.). Reclassifying, entities given by an extractor, using SVM together with the idea of combined attributes made possible an increase of 7.5% in F_1. Other works were also considered because they are Spanish oriented [6] [1].

In the following section the classification method is presented, after that the data and experiments are described, finally the conclusions are stated.

2 Classification method

The Memory-Based Learning method is trained from a set of resolved instances. The training consists of mapping each attribute with its entropy, in order to define a weighted metric. Such a metric allows us to calculate the "distance" between two instances based on the difference of their attributes; the higher difference between attributes with greater information, the biggest distance between the instances. Then, a new instance will have the same solution as the instance in the training set closest to it.

Formally, let S be an instance or training set, $\{A_1, \ldots, A_m\}$ a set of attributes, and $\mathcal{C}$ a set of classes, or solutions of each instance in S, which are represented in the m-th attribute (A_m). Each instance $X = (x_1, \ldots, x_{m-1})$ is assigned to the class of the instance:

$$Y_0 = \mathrm{argmin}_{Y \in S} \mathit{\Delta}(X, Y), \qquad (1)$$

where

$$\Delta(X,Y) = \sum_{i=1}^{m-1} p_i \cdot \overline{\delta}(x_i, y_i), \tag{2}$$

$p_i = Gn(A_i)$, and

$$\overline{\delta}(x_i, y_i) = \begin{cases} 0 & \text{if } x_i = y_i, \\ 1 & \text{if } x_i \neq y_i. \end{cases}$$

The entropy of S is

$$H_S = -\sum_{s_i \in S} \Pr(s_i) \log_2(\Pr(s_i))$$

Given an attribute A_i we can partition S in classes $S|_{x_{ij}}$ (instances with value $x_{i,j} \in A_i$). In this way, the entropy of S with respect to the attribute A_i is the weighting of the entropy for each partition done with the values of A_i:

$$H_{S(A_i)} = \sum_{x_{ij} \in A_i} H_{S|_{x_{ij}}} \frac{\#S|_{x_{ij}}}{\#S} \tag{3}$$

With this, the information gain of an attribute A_i is defined by, $G(A_i) = H_S - H_{S(A_i)}$ and the gain ratio [10] by

$$Gn(A_i) = \frac{G(A_i)}{H_{S(A_i)}}. \tag{4}$$

The algorithm requires an exhaustive search in the training set (equation 1), but it is possible to save computational resources (memory and processor time) using a *trie* tree to represent the instances, which makes it possible to prune the tree on visit and distance computing steps. This implementation is known as *IGTree* [13]. Also, the TiMBL [8] system offers alternative metrics to the one in equation 2.

Using MBL in NER takes up again the BIO tag scheme; B if it is the beginning of the NE, I if the NE continues, and O if it is out of the NE. Even more, if it is the case of a NE its classification (PER, LOC, ORG, MISC) is added. As an example:

> Fernando/B-PER Lozano/I-PER ,/O presidente/O de/O Valle/B-LOC Alto/I-LOC ,/O llegó/O al/O XXI/B-MISC Torneo/I-MISC Universitario/I-MISC .
>
> *Fernando/B-PER Lozano/I-PER ,/O president/O of/O Valle/B-LOC Alto/I-LOC ,/O arrived/O at /O the /O XXI/B-MISC Universitary/I-MISC Tournament/I-MISC ./O*

In this work we focused in identifying (BIO) tags for the (PER, LOC, ORG, MISC) classes in NER.

3 Data sets

The training file contains 273,037 lines, one tag and one word per line. The test file contains 53,049 lines. In our experiments we just took 35,000 (*train*) and 11,000 (*test*) lines of the respective files. These files were obtained from the CoNLL-2002 congress competition [7].

The next step was to select the features that provide more information in solving the problem. Some feature combinations used in other works were taken into account, with very good results. The most important features were intuitively chosen before the experiments were done. Moreover, other feature combinations were tried in order to check their efficacy in NER. The basic features are the following:

- The word context, taking into account three words before and after the word to tag [6].
- The part of speech corresponding to each word of the context.
- Capital letters used in context, if a word begins with a capital letter, it is represented by 1, in other cases by 0.

The files were produced tagging each word with the basic features, and the class to which belongs to. The modified value difference metric was used (MVDM), with $k = 3$ nearest neighbors [3] [4].

4 Experiments

An initial experiment considered to measure the gain ratio of each feature (see figure 1). Context (7), POS (7), using capital letters, and (7) features are arranged on the horizontal axis.

We can see on the graph that the features providing more information to identify tags are those that use capital letters and context; particularly, the fifth word (the one following the word to tag) and using a capital letter has the maximum gain ratio. In order to check the information shown on the graph, experiments about tag identification, using different feature combinations, were done. The following combinations were used:

1. 7 features: word context.
2. 14 features: word context, POS of each word in the context.
3. 21 features: word context, POS , and using capital letters.
4. 14 features: word context, and using capital letters.
5. 8 features: word context, and using capital letters at the fifth word of the context.

We use standard measures, i.e. precision P, recall R, and F_1 to evaluate our results:

$$P = \frac{\#\text{right tags gotten by the system}}{\#\text{tags gotten by the system}}, \tag{5}$$

$$R = \frac{\#\text{right tags gotten by the system}}{\#\text{right tags}}, \tag{6}$$

$$F_1 = (2 \cdot P \cdot R)/(P + R). \tag{7}$$

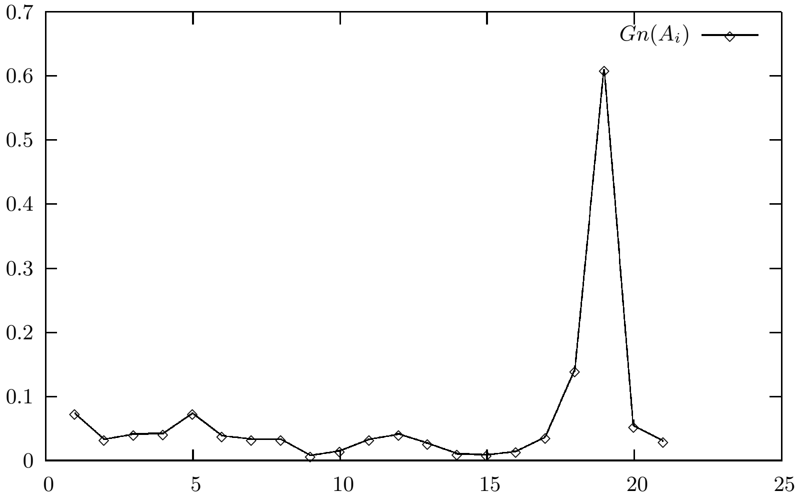

Fig. 1. Information gain of the whole feature set.

As we had said, our interest is on precision, rather than completeness. So, we obtained the recall measure in order to know how many tags will be lost using this method in a text. Recall in each experiment on the CoNLL-02 *test* set is shown on table 1.

Table 1. Recall measure for the different feature combinations.

Tag	**Experiment**				
type	1	2	3	4	5
O	98.2	97.9	99.6	99.5	98.2
B-ORG	41.3	44.6	66.1	57.5	34.0
B-PER	48.4	51.2	67.2	58.8	43.4
I-PER	49.7	58.7	65.5	67.2	44.9
I-ORG	19.5	26.1	40.9	39.9	16.6
B-LOC	57.4	59.7	68.6	63.2	50.8
I-LOC	34.1	38.9	34.1	34.1	28.0
I-MISC	9.0	15.6	12.0	11.2	8.2
Average	40.6	46.7	**52.5**	**49.7**	37.1

As we can see, the third experiment has the highest recall percentage 52.51%. In this case, the precision is 93.11%. However, this test uses the POS and, in our application, it is not possible to include the use of a POS tagger. So, for this system, precision is preferred rather than recall. We considered in these cases we must adopt the features of test 4, which has recall 49.72% and precision 92.48%. Average standard measurements for the tests are shown in Table 2.

Table 2. Performance for each feature set.

Experiment	P	R	F_1
1	89.53	40.69	55.95
2	90.07	46.77	61.57
3	93.11	52.51	67.15
4	92.48	49.72	64.67
5	88.76	37.15	52.37

In order to situate the results, we did 10 cross validation tests on the development set, and we shall cite the results, about tag identification, presented in the CoNLL-02. The purpose of this comparison is to know how much is lost when some features are omitted (for example, POS), because we are just taking features we are interested in 14 features (experiment 4), and do not need a POS tagger. The averaged measures were $R = 70.71$, $P = 95.89$, and $F_1 = 81.39$. The classification on the development set without cross validating got $R = 69.8$, $P = 96.14$, and $F_1 = 80.87$. Our classification is better than the one presented in [12] ($F_1 = 74.34$). Both took place under the same conditions (test set, learning method, and cross validating). Nevertheless, Carreras *et al.* result [11] ($F_1 = 91.66$) is better than ours.

5 Conclusions

We have shown the results on the efficacy of different sets of features, used in NER by the MBL method, on a collection from the CoNLL-2002 conference. The experiments are based on the combination of basic features: word context (three words before and after), the POS of words in the context, and the presence of a capital letter at the beginning of the words in the context.

We see that contextual and some morphological features are very helpful in classifying tags for NER. Other authors have referred that external information sources are almost useless. In this work we have seen that, omitting the POS of contextual words does not impact the precision in identifying tags for NER: the maximum precision gotten in CoNLL-02 [11] was 92.45 against our result 96.14 reached using just 14 features. Certainly, our recall is very poor (90.88 by them against 69.8 by us), because of the reduced number of features that we used.

References

1. Lluis Padró: Tendencias en el reconocimiento de entidades con nombre propio, en *Tecnologías del Texto y del Habla*, M. Antonia, Martí & Joaquim, Llisterri (Eds.) Edicions Universitat de Barcelona, Fundación Duques de Soria (2004).

2. Eric F. Tjong Kim Sang.: Introduction to the CONLL-2002 Shared Task: Language-Independent Named Entity Recognition, *CNTS - Language Technology Group*, University of Antwerp.
3. Fien, De Meulder & Walter, Daelemans.: Memory-Based Named Entity Recognition using Unannotated Data, *CNTS - Language Technology Group*, University of Antwerp.
4. Walter Daelemans., Jakub Zavrel, Ko van der Sloot & Antal van den Bosch.: TiMBL: Tilburg Memory Based Learner, version 4.3, Reference Guide (2002).
5. Thamar, Solorio & Aurelio, López López.: Adapting a Named Entity Recognition System for Spanish to Portuguese, Instituto Nacional de Astrofísica Optica y Electrónica,2004
6. Alvaro, Balbontín Gutiérrez & José Javier Sánchez Martín.: SPNER-Reconocedor de entidades nombradas para el español, Universidad Europea de Madrid.
7. Language Independent Named Entity Recognition, *Conference on Natural Language Learning 2002*, `http://lcg-www.uia.ac.be/conll2002/ner/`.
8. Machine Learning for Language Engineering and Linguistics, Tilburg Uniersity, `http://pi0657.uvt.nl/`.
9. X. Carreras & L. Padró: A flexible distributed architecture for natural language analyzers, In *Proc. of LREC'02*, Las Palmas de Gran Canaria, España, 2002.
10. C. E. Shannon: Communication Theory, *The Bell System Technical Journal*, (27) 379, p. 19-48, 1950.
11. Xavier Carreras, Lluis Màrquez & Lluis Padró: Named Entity Extraction using AdaBoost, Language Independent Named Entity Recognition, *Conference on Natural Language Learning 2002.*
12. Eric F. Tjong Kim Sang.: Memory-based Named Entity Recognition *Conference on Natural Language Learning 2002.*
13. Walter Daelemans, Antal van den Bosch, Jakub Zavrel, Jorn Veenstra, Sabine Buchholz, and Bertjan Busser: Rapid development of NLP modules with memory-based learning, In *Proceedings of ELSNET in Wonderland*, pp. 105-113. Utrecht: ELSNET, 1998.