Language Portable Detection for Spanish Named Entities

Zornitsa Kozareva, Oscar Ferrández, Andrés Montoyo and Rafael Muñoz

Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante, Spain
{zkozareva,ofe,montoyo,rafael}@dlsi.ua.es

Abstract. We propose a language portable Named Entity detection module developed and tested over Spanish and Portuguese. The influence of different feature sets over the classification task was studied and demonstrated. The differences in language models learned by three data-driven systems performing the same NLP tasks were examined. They were combined in order to yield a higher accuracy than the best individual system. Three NE classifiers (Hidden Markov Models, Maximum Entropy and Memory-based learner) are trained on the same corpus data and after comparison their outputs are combined using voting strategy. Results are encouraging since 92.96% f-score for Spanish and 78.86% f-score for Portuguese language portable detection were achieved. For Spanish the classification which is based on the language portable detection reached 78.59% f-score. Compared with the systems competing in CoNLL-2002 our system reaches third place.

1 Introduction

The increasing flow of digital information requires the extraction, filtering and classification of pertinent information from large volumes of texts. Information Extraction, Information Retrieval and Question Answering systems need Named Entity (NE) recognition and classification modules. For English the available resources and the developed systems outnumber, but in the case of languages as Spanish, Portuguese or eastern European ones where the resources as gazetteers[1], annotated corpora etc. are not sufficient but the need is still the same, the situation looks different. This fact motivated us to start the development of a language resource independent system during its NE detection and using less resources while classifying into LOC, PER or ORG classes.

In this paper we present a NE system developed for Spanish. Three machine learning algorithms were used in concrete: Hidden Markov Model, Maximum Entropy and the Memory-based learner. They were applied to the CoNLL-2002 shared task for Spanish. The language portable detection was also tested with Portuguese language. Both languages come from the Romance language group and have similar behavior so features valid for Spanish were directly adopted by Portuguese.

[1] catalogues of names of people, locations, organizations etc.

© A. Gelbukh, R. Monroy. (Eds.)
Advances in Artificial Intelligence Theory
Research on Computing Science 16, 2005, pp. 219-228

In order to improve overall performance feature selection and systems' combination were done. Aiming at minimal feature space, less processing time and gaining results while restraining from gazetteers, the obtained results are quite encouraging. For Spanish we reached 92.96% f-score for language portable detection and 78.59% f-score for classification. Portuguese was used to support our hypothesis for language portable detection and we gained 78.86%. A study of the occurred errors and proposals for resolving them was made, comparison with existing systems and future work are discussed. The paper is organized as following: in Section 2 we expose the features on which the classification methods are based and a brief description of the classifiers, the voting strategy and the data with which we worked is in Section 3, discussion of the obtained results and error correction during NE detection is demonstrated in Section 4, classification into classes is represented in Section 5, a comparison with CoNLL-2002 systems is exposed in Section 6, we conclude and mention about the future work in Section 7.

2 Feature description and Classification methods

For NE detection and classification task, the Memory-based learning and Maximum Entropy classifiers utilize the features described below, HMM takes only the three most informative attributes.

2.1 Features for NE detection

We use the well-known BIO model for NE detection, where a tag shows that a word is at the beginning of a NE (B), inside a NE (I) or outside a NE (O). For the sentence: *Paulo Suarez es mi amigo.* , the following tags have been associated, "B I O O O O ", where *Paulo* starts a named entity; *Suarez* continues this entity; while the words *es*, *mi*, *amigo* and the full stop are not part of a NE.

The original set for BIO is composed of 29 features as described in Figure 1 and we denote this set by A. For improving classifier's performance different feature combinations of the original set were constructed. The features represent words, position in a sentence, capitalization, suffixes, context information, lists of entity triggers for NE. The features c[1-6], C[1-7], d[1-3] refer to the words in a $\{-3, +3\}$, window of the anchor word **a**.

We extracted two, three and half substrings of the anchor word, knowing that some prefixes and suffixes are good indicators for certain classes of entities, taking into account the morphological structure of a word and its paradigm. In general suffixes are more informative, for Spanish endings as *-er,-or,-ista* imply person's occupation *pianista, futbolista, profesor, director* and can help during detection and classification phase. It is surprising the number of Spanish surnames that end in *-ez*, meaning "son of", like the suffix *-son* and *-sen* in many German and Scandinavian languages, *-ov,-ova,-ev,-eva* in Russian and Bulgarian, and-*es* in Portuguese. (e.g. Fernandez is the son of Fernando [Ferdinan]). Of course these examples have many exceptions, but the information they contribute is

- **a**: anchor word (e.g. the word to be classified)
- ***c[1-6]***: word context at position ±1, ±2, ±3
- ***C[1-7]***: word capitalization at position 0, ±1, ±2, ±3
- ***d[1-3]***: word +1,+2,+3 in dictionary of entities
- ***p***: position of anchor word
- ***aC***: capitalization of the whole anchor word
- ***aD***: anchor word in any dictionary
- ***aT***: anchor word in dictionary of trigger words
- ***wT***: word at position ±1, ±2, ±3 in a dictionary of trigger words
- ***aL***: lema of the anchor word
- ***aS***: stem of the anchor word
- ***aSubStr[1-5]***: ±2, ±3 and half substring of the anchor word

Fig. 1. Features for NE detection

significant when combined with other features. The lemma expands the search in the gazetteers' list we maintain, we can have the word "profesor" but not "profesora" and by the lemma which returns the base of the word, we are going to have a positive vote.

2.2 Features for NE classification

The tags used for NE classification are PER, LOC, ORG and MISC as defined by CoNLL-2002 shared task. For classification, the first seven features used by the BIO model (e.g. a, c[1-6], p) were used as well as the additional set described in Figure 2. The gazetteers for the attributes gP, gL and gO have been collected randomly from cites as yellow pages.

- ***eP***: entity is trigger PER
- ***eL***: entity is trigger LOC
- ***eO***: entity is trigger ORG
- ***eM***: entity is trigger MISC
- ***tP***: word ±1 is trigger PER
- ***tL***: word ±1 is trigger LOC
- ***tO***: word ±1 is trigger ORG
- ***gP***: part of NE in gazetteer for PER
- ***gL***: part of NE in gazetteer for LOC
- ***gO***: part of NE in gazetteer for ORG
- ***wP***: whole entity is PER
- ***wL***: whole entity is LOC
- ***wO***: whole entity is ORG
- ***NoE***: whole entity not in one of the defined three classes
- ***f***: first word of the entity
- ***s***: second word of the entity
- ***clx***: capitalization, lowercase, other symbol

Fig. 2. Features for NE classification

2.3 Classification methods

For NE detection we worked with Memory-based learning and Hidden Markov Model, while for NE classification we had also Maximum Entropy.

The memory-based software package we used is called TiMBL [3]. Its default learning algorithm, instance-based learning with information gain weighting (IB1IG) was applied. The Hidden Markov Models toolkit ICOPOST[2] developed by [8] has been functioning for POS tagging, but we adapted it for NER. The maximum entropy classifier we worked with was a very basic one with no smoothing or feature selection, implemented in C++ by [9].

3 Classifier combination and Data

3.1 Classifier combination

It is a well-known fact that if several classifiers are available, they can be combined in various ways to create a system that outperforms the best individual classifier. Since we had several classifiers available, it was reasonable to investigate combining them in different ways. The simplest approach to combining classifiers is through voting, which examines the outputs of the various models and selects the classifications which have a weight exceeding some threshold, where the weight is dependent upon the models that proposed this particular classification. It is possible to assign varying weights to the models, in effect giving one model more importance than the others. In our system, we assigned to each model the weight corresponding to the correct class it determines.

3.2 Data and its evaluation

The Spanish train and test data we used are part of the CoNLL-2002 [7] corpus. For training we had corpus containing 264715 tokens and 18794 entities and for testing we used Test-B corpus with 51533 tokens and 3558 entities.

The Portuguese corpus we used is part of HAREM[3] competition with 68597 tokens and 3094 entities for training, and 22624 tokens and 1013 entities for testing.

Scores were computed per NE class and the measures used were Precision (of the tags allocated by the system, how many were right), Recall (of the tags the system should have found, how many did it spot) and $F_{\beta=1}$(a combination of recall and precision). *Conlleval* evaluation script was used in order to have comparable results with the CoNLL-2002 systems.

[2] http://acopost.sourceforge.net/

[3] http://poloxldb.linguateca.pt/harem.php

4 NE recognition by BIO model

Our NER system is composed of two passages

1. detection: identification of sequence of words that makes up the name of an entity.

2. classification: deciding to which category our previously recognized entity should belong.

For NE detection we follow the BIO model described briefly in subsection 2.1. Our experiments with TiMBL started using set $C24 = A/\{aSubStr[1-5]\}$, which contained all attributes as lemma, dictionaries, trigger words etc. The obtained results have been satisfactory as can be seen in Table 1, but since we have been searching for an appropriate feature set F that maximizes the performance, minimizes the computational cost and being language portable, we made a study of the features and selected the most informative ones according to the information gain measure. Four candidate sets were formed and we denote them by $C24r = \{a, c[1-6], C[1-7], p, aC, wD, wT, aL, aS\}$ and $C17 = C24r/\{c[5-6], C[6-7]\}$; considered as language dependent (they use dictionaries, tools as lemmatizers, stemmers etc.) and $E12 = \{a, c[1-4], C[1-5], p, aC\}$ and $E17 = E12 \cup \{aSubStr[1-5]\}$, considered as language portable. The results of each individual set can be seen in Table 1.

Tags	B(%)			I(%)			BIO(%)		
Classifier	Prec.	Rec.	$F_{\beta=1}$	Prec.	Rec.	$F_{\beta=1}$	Prec.	Rec.	$F_{\beta=1}$
TMB-C24	94.42	95.19	94.81	87.25	85.67	86.45	92.51	92.61	92.56
TMB-C17	94.47	95.11	94.79	87.28	85.37	86.31	92.56	92.47	92.51
TMB-C24r	94.63	94.01	94.32	87.99	85.07	86.50	92.86	91.58	92.22
HMM-CD	92.18	93.82	92.99	83.94	81.98	82.95	90.01	90.60	90.31
HMM-CW	92.40	93.99	93.19	83.71	81.00	82.33	90.13	90.46	90.29
Vote 1 ld	95.31	95.36	95.34	88.02	87.56	87.79	93.34	93.24	93.29
TMB-E12	94.33	94.91	94.62	87.00	85.29	86.14	92.38	92.30	92.34
TMB-E17	94.17	95.28	94.72	87.62	85.37	86.48	92.44	92.59	92.51
HMM-CW	92.40	93.99	93.19	83.71	81.00	82.33	90.13	90.46	90.29
Vote 2 li	94.43	95.73	95.07	88.31	86.05	87.17	92.81	93.10	92.96

Table 1. *BIO for Spanish*

Initially to HMM we passed the NE and the tag associated with it. The obtained performance of 88.63% is less than each one of TiMBL's individual sets, however this difference is compensated with the number of features TiMBL uses. For the word *Don Simon*, which in one text can mean a name of a person or organization (e.g. company name), in order to determine its correct significance more information is needed. One advantage of HMM is its time performance of several minutes in comparison with the other methods, but fails in adding lots of features. As studied by Rössler [6] to HMM features can be passed by

Tags	B(%)			I(%)			BIO(%)		
Classifier	Prec.	Rec.	$F_{\beta=1}$	Prec.	Rec.	$F_{\beta=1}$	Prec.	Rec.	$F_{\beta=1}$
TMB-E12	82.50	83.32	82.91	72.77	64.77	68.53	79.59	77.26	78.41
TMB-E17	80.13	83.22	81.65	69.64	58.86	63.80	77.16	75.27	76.20
HMM-CW	77.83	68.61	72.93	61.02	58.66	59.81	72.01	65.36	68.53
Vote 3 li	82.35	84.30	83.32	72.75	65.78	69.09	79.47	78.26	78.86

Table 2. *Language portable BIO for Portuguese*

corpus or tag transformation. We studied both possibilities and saw that tag transformation gave better results. We took the two most informative attributes - word capitalization and whole word in capitals, plus the gazetteer list and passed them as features to the B and I tags. For *La Coruña* we have B-XX and I-XX tags, where the XX take binary features. With HMM-CD we denote the results after passing the attributes word capitalization and word in dictionary and with HMM-CW the results from word capitalization and whole word in capitals. Adding these attributes, HMM's performance increases with around 1.68%.

The obtained results from all BIO sets for Spanish can be observed in Table 1, there we mention the language dependent sets for comparison, but for further experiments (classification) we consider the results from the language portable sets. In Table 2 we demonstrate Portuguese language portable BIO detection using the same sets as for Spanish.

The coverage of tag O is high due to its frequent appearance, however its importance is not so significant as the one of B and I tags, who actually detect the entities. For this reason we demonstrate separately system's precision, recall and f-score for B and I tags. The best score for Spanish BIO was obtained by TiMBL considering the complete $C24$ set with f-score of 92.56%. Comparing this score with set $C17$ where he number of features is reduced, the word window diminished from ±3 to ±2, the difference of 0.05% is insignificant. Set $C24r$ was studied for reducing some noisy attributes from set $C24$ but still keeping the ±3 window. Its total BIO performance decreased but gained 86.50% - the highest f-score per I tag.

The language portable sets perform quite similar to the dependent ones. For tag B, set E12 with its 12 attributes performs better than C24r. The complete BIO for E12 is better than those of C24r. TMB-E17 improves slightly the overall results of E12 and has similar results to C17. For tag I it performs better than C24, C17 and has 0.02% less performance than C24r.

The classifiers used different feature sets and we noticed that one classifier detects an entity while the other doesn't. The classifiers used different feature sets and we noticed that one classifier detects an entity while the other doesn't. After obtaining the different results we applied voting techniques grouping the language dependent sets in vote one and the language portable sets in vote two. The difference of 0.33% between *Vote 1 language dependent* with 93.29% performance and *Vote 2 language portable* with 92.96% f-score shows how small

Tags	LOC(%)			MISC(%)			ORG(%)			PER(%)		
Classifier	Prec.	Rec.	$F_{\beta=1}$	Prec.	Rec.	$F_{\beta=1}$	Prec.	Rec.	$F_{\beta=1}$	Prec.	Rec.	$F_{\beta=1}$
ME-F24	81.16	74.72	77.81	69.29	49.12	57.49	74.21	84.07	78.83	82.95	88.03	85.41
TMB-F24	75.70	75.28	75.49	55.03	51.47	53.19	75.22	79.79	77.44	84.53	83.27	83.89
ME-F24clx	81.94	74.91	78.27	69.67	50.00	58.22	73.92	84.00	78.64	83.18	88.16	85.60
TMB-F24clx	74.84	75.46	75.15	55.88	50.29	52.94	75.88	79.79	77.79	85.42	85.31	85.36
TMB-R24	80.08	75.65	77.80	57.95	48.24	52.65	77.01	81.36	79.12	79.24	88.30	83.53
TMB-R24clx	79.20	75.18	77.14	63.20	50.00	55.83	76.14	81.36	78.66	80.15	88.44	84.09
HMM	74.85	67.80	71.15	44.66	46.76	45.69	72.06	73.86	72.95	66.11	74.83	70.20
VM24T24fclxH	81.16	75.92	78.46	66.80	49.71	57.00	75.06	83.21	78.93	83.72	89.52	86.52

Table 3. *NE classification*

feature set containing attributes independent from any tools, dictionaries or gazetteers can give good and similar results to the dependent sets.

Taking in mind that Spanish and Portuguese are languages having similar behavior, we studied and saw how attributes valid for Spanish were directly adopted by Portuguese. In Table 2 we to show the results for Portuguese after applying the same set of portable features as for Spanish. With voting 83.32% f-score for B tag and 78.86% for complete BIO were achieved. These results are acceptable since we didn't have sufficient training data and the annotated corpus we used had significant number of errors.

4.1 BIO error Analysis

After analyzing the obtained results, we saw that some of the occurred errors can be avoided by applying simple post-processing: when an I tag has been preceded by O tag we substituted it by B if the analyzed word starts with a capital letter and in the other case we simply put O; sequences such as *OBIBI*, have been transformed into *OBIII*. (see the example in subsection 2.1).

For Spanish around 2% of the errors came from the annotated corpus, sometimes quotation mark symbol was annotated as B and sometimes as O. Portuguese corpus was quite noisy having entity as *v+,n 12* annotated as organization or some names of people were even not annotated.

One of our attributes concerns word capitalization and had great impact over the detection task. We noticed how sometimes words starting a sentence but not belonging to any of the named entity classes were classified as B tags. A statistical study of word frequency, determines if a word at the beginning of a sentence should have a B tag or not. The word variants (e.g. writing of a word), their individual frequency and neighbors with which these words appear are studied. Thus we have been able to correct and avoid this kind of error.

Tags	LOC(%)			MISC(%)			ORG(%)			PER(%)		
Classifier	Prec.	Rec.	$F_{\beta=1}$	Prec.	Rec.	$F_{\beta=1}$	Prec.	Rec.	$F_{\beta=1}$	Prec.	Rec.	$F_{\beta=1}$
ourNE	81.16	75.92	78.46	66.80	49.71	57.00	75.06	83.21	78.93	83.72	89.52	86.52
WNC	79.15	77.40	78.26	55.76	44.12	49.26	74.73	79.21	76.91	80.20	89.25	84.48
CY	79.66	73.34	76.37	64.22	38.53	48.16	76.79	81.07	78.87	82.57	88.30	85.34
Flo	82.06	79.34	80.68	59.71	61.47	60.58	78.51	78.29	78.40	82.94	89.93	86.29
CMP	85.76	79.43	82.43	60.19	57.35	58.73	81.21	82.43	81.81	84.71	93.47	88.87

Table 4. *CoNLL-2002 NE classification*

Classifier	Prec. %	Rec. %	$F_{\beta=1}$ %
CMP	81.36	81.40	81.39
Flo	78.70	79.40	79.05
ourNE	78.09	79.10	78.59
CY	78.19	76.14	77.15

Table 5. *Complete system performance*

5 NE classification

After detection follows NE classification into LOC, MISC, ORG or PER class as defined by CoNLL-2002. For this task, we used the results obtained from the language portable detection.

For ME and TiMBL, we started the classification with a set composed of 24 features as described in subsection 2.2. Let us denote by $F24$ the set having features: a, c[1-6], p, eP, eL, eO, eM, tP, tL, tO, gP, gL, gO, wP, wL, wO, NoE, f and s. In Table 3 comparing the performance of ME and TiMBL with the same set can be seen how ME classifies better for each one of the classes.

Choosing the most informative attributes,{a, c[1], eP, gP, gL, gO, wP, wL, wO, NoE, f}, we create a set $R24$, where $R24 \subset F24$. In Table 3 we displayed only the results obtained by TiMBL, because ME needs a lot of time for training and testing. When both classifiers were compared on small random samples from the original set, we saw that TiMBL performs better with the reduced set. When $R24$ was tested with the complete data TiMBL achieved the highest result for ORG class of 79.12%. Two additional sets $R24clx = R24 \cup \{clx\}$ and $F24clx = F24 \cup \{clx\}$, where clx is the attribute described in Figure 2, were constructed. $R24clx$ lowered the performance for LOC and ORG class compared to the $R24$ set but performed better dealing with MISC and PER class. By adding clx attribute to $F24$, ME improved its performance with 0.46% for LOC and 0.19% for PER class and gained the maximum score of 58.22% for MISC class. Using the same set TiMBL decreased its score for LOC and MISC class and slightly improved ORG and PER classes. Among all classifiers, HMM has the lowest score per class. The voting we applied considers ME-F24, Timbl-F24clx and HMM results.

We have seen how the elimination or addition of features gave impact over given types of classes during classification. As a whole our systems perform well when classifying into PER,ORG and LOC class, but not when dealing with MISC class which is difficult to be detected due to its heterogeneity.

6 Comparison with CoNLL-2002 systems

We demonstrated the performance of NER considering different machine learning methods, where the advantages and disadvantages of each one of them being in time performance or feature maintenance was shown. Apart from this it is very interesting to make a comparative study with the systems participating in CoNLL-2002 shared task, since our system has been developed using the same data; we should take in mind that our classification has been based on language portable detection.

Table 4 represents the results per class for our system and the first four best performing systems in CoNLL-2002; WNC[4], CY[2], Flo[5], CMP[1]. When classifying into LOC class our system performed with 0.2% and 2.09% better than the one of Wu and Cucerzan and less with 2.22% and 3.97% from the systems of Florian and Carreras. Our classification into MISC class was better with 7.74% and 8.84% compared to the one of Wu and Cucerzan and less with 3.58% and 1.73% from Florian and Carreras. For ORG and PER classes we outperformed all systems except the one of Carreras. With Wu's system we have 2.02% and 2.04% better score per ORG and PER class, from Cucerzan's 0.06% and 1.18% and from the system of Florian 0.53% and 0.23%.

We separated the overall performance of the first three best performing systems in Table 5. Comparing the f-score our system performs with 1.44% better than the third one, with 0.46% less than the second and with 2.8% less than the first system.

7 Conclusions and future work

We presented a combination of machine learning methods (Memory-based learning, Maximum Entropy and HMM) for performing NE detection and classification task for Spanish. Aiming at minimal feature space and restraining from dictionaries or other language dependent tools, we demonstrated one language portable detection for Spanish and Portuguese. The Portuguese system served as proof for our experiments and hypothesis. At present we didn't study the achievement of language portable classification over Spanish and we depend on gazetteers but in future we intend to work on this task. Comparing our results with CoNLL-2002 participants the f-score results of 78.46% for LOC, 57.00% for MISC, 78.93% for ORG and 86.52% for PER are quite encouraging and are among the second and third system.

As future work we intend to develop and use specific dictionaries for NEs, to apply the same method for languages as Catalan, Italian and French. We are interested in dividing the original three base tags into more detailed ones, for example: ORG class into administration, institution, company etc. A Word Sense Disambiguation module is going to be included and the rule based system that was separately developed and deals with weak entities such as *El presidente del Gobierno de La Rioja* is going to be merged with the machine learning module we have developed.

Acknowledgements

This research has been funded by the Spanish Government under project CICyT number TIC2003-07158-C04-01 and PROFIT number FIT-340100-2004-14 and by the Valencia Government under project numbers GV04B-276 and GV04B-268.

References

1. Xavier Carreras, Lluís Màrques, and Lluís Padró. Named entity extraction using adaboost. In *Proceedings of CoNLL-2002*, pages 167–170. Taipei, Taiwan, 2002.
2. Silviu Cucerzan and David Yarowsky. Language independent ner using a unified model of internal and contextual evidence. In *Proceedings of CoNLL-2002*, pages 171–174. Taipei, Taiwan, 2002.
3. Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. TiMBL: Tilburg Memory-Based Learner. Technical Report ILK 03-10, Tilburg University, November 2003.
4. Marine Carpuat Jeppe Larsen Dekai Wu, Grace Ngai and Yongsheng Yang. Boosting for named entity recognition. In *Proceedings of CoNLL-2002*, pages 195–198. Taipei, Taiwan, 2002.
5. Radu Florian. Named entity recognition as a house of cards: Classifier stacking. In *Proceedings of CoNLL-2002*, pages 175–178. Taipei, Taiwan, 2002.
6. M. Rössler. Using markov models for named entity recognition in german newspapers. In *Proceedings of the Workshop on Machine Learning Aproaches in Computational Linguistics*, pages 29–37. Trento, Italy, 2002.
7. Tijong Kim Sang. Introduction to the conll-2002 shared task: Language independent named entity recognition. In *Proceedings of CoNLL-2002*, pages 155–158, 2002.
8. Ingo Schröder. A case study in part-of-speech tagging using the icopost toolkit. Technical Report FBI-HH-M-314/02, Department of Computer Science, University of Hamburg, 2002.
9. Armando Suárez and Manuel Palomar. A maximum entropy-based word sense disambiguation system. In Hsin-Hsi Chen and Chin-Yew Lin, editors, *Proceedings of the 19th International Conference on Computational Linguistics, COLING 2002*, pages 960–966, August 2002.