Available online at www.sciencedirect.com
ScienceDirect
Data & Knowledge Engineering 61 (2007) 449–466
DATA & KNOWLEDGE ENGINEERING
www.elsevier.com/locate/datak

Combining data-driven systems for improving Named Entity Recognition ☆

Z. Kozareva *, O. Ferrández, A. Montoyo, R. Muñoz, A. Suárez, J. Gómez

Natural Language Processing and Information Systems Group, Department of Software and Computing Systems, University of Alicante, Alicante, Spain

Received 20 June 2006; accepted 20 June 2006
Available online 17 July 2006

Abstract

The increasing flow of digital information requires the extraction, filtering and classification of pertinent information from large volumes of texts. All these tasks greatly benefit from involving a Named Entity Recognizer (NER) in the preprocessing stage. This paper proposes a completely automatic NER system. The NER task involves not only the identification of proper names (Named Entities) in natural language text, but also their classification into a set of predefined categories, such as names of persons, organizations (companies, government organizations, committees, etc.), locations (cities, countries, rivers, etc.) and miscellaneous (movie titles, sport events, etc.). Throughout the paper, we examine the differences between language models learned by different data-driven classifiers confronted with the same NLP task, as well as ways to exploit these differences to yield a higher accuracy than the best individual classifier. Three machine learning classifiers (Hidden Markov Model, Maximum Entropy and Memory Based Learning) are trained on the same corpus in order to resolve the NE task. After comparison, their output is combined using voting strategies. A comprehensive study and experimental work on the evaluation of our system, as well as a comparison with other systems has been carried out within the framework of two specialized scientific competitions for NER, CoNLL-2002 and HAREM-2005. Finally, this paper describes the integration of our NER system in different NLP applications, in concrete Geographic Information Retrieval and Conceptual Modelling.

Keywords: Named Entity Recognition; Machine learning; Feature selection; Classifier combination; Multilinguality; Natural Language Processing

1. Introduction

The vision of the information society as a global digital community is fast becoming a reality. Progress is being driven by innovation in business and technology, and the convergence of computing, telecommunications

☆ This research has been funded by the Spanish Government under project CICyT number TIC2003-07158-C04-01, TIN2004-00779 and PROFIT number FIT-340100-2004-14, CESS-ECE number HUM2004-21127-E and by the Valencia Government under project numbers GV04B-276 and GV04B-268.
* Corresponding author.
E-mail addresses: zkozareva@dlsi.ua.es (Z. Kozareva), ofe@dlsi.ua.es (O. Ferrández), montoyo@dlsi.ua.es (A. Montoyo), rafael@dlsi.ua.es (R. Muñoz), armando@dlsi.ua.es (A. Suárez), jgomez@dlsi.ua.es (J. Gómez).

0169-023X/$ - see front matter
doi:10.1016/j.datak.2006.06.014

and information systems. Access to knowledge resources in the information society is vital to both our professional and personal development. However, access alone is not enough. We need to be able to select, classify, assimilate, retrieve, filter and exploit this information, in order to enrich our collective and individual knowledge and skills. This is a key area of application for language technologies. The approach taken in this area is to develop advanced applications characterized by more intuitive natural language interfaces and content-based information analysis, extraction and filtering. Natural Language Processing (NLP) is crucial in solving these tasks. In concrete, Named Entity Recognition (NER) has emerged as an important preprocessing tool for many NLP applications such as Information Extraction (IE), Information Retrieval (IR) among other text processing applications, mainly because Name Entities (NEs) provide important cues for identifying relevant information in text.

One example of application that benefits greatly from NER is Question Answering, as, in order to answer questions like "Who is the president of the Spanish Government?", it is useful to know that the expected answer is the name of a person, and, obviously, to consider as candidate answers only entities of this type. Accurate NE detection can also help IE systems to gain knowledge about where an event happened, who was involved in it, etc. Developing and populating ontologies is another task to be largely improved by NER. Even in automatic semantic role labelling, the extra information about an item being a person or an artifact could be crucial to determine what entity denominates the agent, depending on the type of the verb. In the area of Conceptual Modelling, a NER module is used to extract functional requirements from user's documents written in natural language.

As can be seen, NER has crucial importance for the performance of various NLP applications. To gain a better insight into this area, we will further present a short history, together with different approaches adopted in the identification and classification of NEs.

1.1. Origins and background of NER

The NER task has been first introduced by the sixth Message Understanding Conference (MUC-6). The MUC Conferences [1] have been established in 1987 with the aim of evaluating Information Extraction Systems. During MUC-6, a definition and description of the term Named Entity (NE) was introduced, together with measures for the accuracy of a system performing NER. NER involves processing a text and identifying certain occurrences of words or expressions as belonging to a particular category of Named Entities (NEs) such as names of persons, organizations, locations and numeric expressions like money or percent expressions. In fact, NER has two main sub-problems. One is NE detection (NED), that is the identification of the portion of text that forms a NE, (for example, "El Presidente Rodriguez Zapatero"[1]) and the second is NE classification (NEC), the process of assigning a category label to the identified span of text (the system would assign the tag person to the NE "El Presidente Rodriguez Zapatero").

Conferences such as MUC, CoNLL[2] or ACE[3] proposed a limited set of categories (for example, person, organization, location, and numeric expressions like money or percent expressions), but there are many more possible labels ([2] proposed an extended Named Entity hierarchy which contains about 150 NE types). So far the research community does not commonly agree on the relationship between the number of classes and the domain of application.

1.2. Different approaches

Normally, an automatic NE tagger follows one of two possible approaches: the approach that employs dictionaries and hand-made rules (knowledge-based systems like the one presented in Arevalo et al. [3]) or the approach based on supervised learning techniques. The difficulty of building accurate systems following the first approach has lead to many researchers showing recently increased interest in supervised learning techniques. Supervised learning has the advantage to acquire automatically linguistic knowledge from a pool of

[1] The president Rodriguez Zapatero.
[2] http://www.cnts.uia.ac.be/conll2002/ner/, http://www.cnts.ua.ac.be/conll2003/ner/.
[3] http://www.nist.gov/speech/tests/ace/.

correctly hand-annotated examples. Therefore, this type of techniques are easy to adapt to another domain or languages, but they depend on appropriate corpus. This is the main limitation of these approaches, because annotated data is not usually available. Supervised learning techniques employed so far in NER include: Decision Tree [4], Hidden Markov Model [5,6], Maximum Entropy Model [7], Support Vector Machine [8], Boosting and voted perceptron [9], AdaBoost [10], Conditional Random Fields [11] and Finite State Automata [12].

In order to detect and classify the NEs, typically these NER systems use the surrounding context or syntactically annotated text. But it is also usual to use external resources such as lists of trigger words[4] or gazetteers.[5] In general, the larger the set of defined features is, the more information a classifier possesses to perform its task. Some problems that may arise are related to domain or language dependency, or over-fitting.

In fact, NER suffers the same limitations as many other NLP tasks, the bottleneck of knowledge acquisition both for corpus-based and knowledge-based systems. Knowledge-based systems rely on manually defined rules and dictionaries. This poses the difficulty of being successfully tuned to new domains or applications. On the other hand, corpus-based approaches are dependent on the availability of properly annotated corpora even when a restricted domain is needed. In both cases the high cost of acquiring such knowledge (explicit or in the form of sets of examples) is a serious handicap. Several attempts have been made to decrease the effort in acquisition and application of knowledge for NER.

The Muse system [13] is a multi-purpose NER system which focuses on this matter. One of its properties is the ability to deal with heterogeneous sources of text. [14] proposed to use active learning techniques in order to minimize the human annotation effort by automatically selecting the most useful examples for training. For the NE classification task, [15] presented a pair of algorithms based on co-training [16] that use seven simple "seed" rules. This approach is based on natural redundancy in data because, they say, both spelling and the context in which the name appears is sufficient to determine its type. Three classes of entities are location, person and organization. Other approaches based on the usage of unlabeled data were presented by [17,18]. They applied self-training techniques to resolve the NE detection task and proposed a new voted co-training method for classify the NEs. Compared to [15] this approach did not need a split of features, but rather different machine learning classification methods that are combined through voting. The NER has been done for Spanish and the identified NEs have been classified into location, organization, person and miscellaneous categories.

Also by means of a bootstrapping technique, [19] studied the role of syntax-rich features such as constituency and dependency in order to iteratively feed a semi-supervised system based on Expectation Maximization(EM) with labeled and unlabeled data. Other approaches combine supervised and unsupervised methods with multilinguality such as [20] did for Catalan and Spanish using both AdaBoost.MH and Greedy Agreement Algorithm [21]. This system takes advantage of the syntactic similarity between the two languages. Their conclusion is that, by using multilingual resources, one can clearly outperform other approaches.

A common approach for NER and other NLP tasks, is related to the combination of several methods and classifiers that improve the performance. For example, [22] solved NER using a combination of four classifiers: robust linear classification, Maximum Entropy, transformation-based learning and Hidden Markov Model. [23] presented a system based on stacking and voting of strong classifiers, such as Support Vector Machine, boosting and memory-based learning.

Our system has been developed following this last approach. We combined several machine learning methods and linguistic knowledge resources. The main goals of the work presented in this paper are:

- to propose a completely automatic NER which involves the identification of proper names in texts, and their classification into a set of predefined categories;
- to examine the differences between language models learned by different data-driven systems performing the same NLP tasks and how they can be exploited to yield a higher accuracy than the best individual system;
- to use voting strategy to combine effectively strong classifiers such as Hidden Markov Models, Maximum Entropy and Memory-based;

[4] Semantically significant word pointing to some of the categories person, location, organization; e.g. city is a trigger word for locations.
[5] Collections of names of people, locations, organizations.

- to evaluate the proposed NER system in two specialized scientific competitions for NER, such as CoNLL-2002 and HAREM-2005;
- to utilize our NER system in a NLP application such as IR;
- to evaluate the applicability of our system in the area of Conceptual Modelling.

The organization of the paper is the following: the system is described in Section 2, the conducted experiments and a discussion of the obtained results follow in Sections 3, Section 4 describes the combination of the classifiers, Section 5 presents a comparison with other NER systems, followed by the integration of the system within an IR application and a Conceptual Modelling application (Section 6). Finally, we conclude in Section 7 with a summary of the most important achievements and plans for future work.

2. NERUA system description

In the following subsections, we describe the Named Entity Recognition system developed at the University of Alicante and called NERUA[6] [24]. It is composed of two main modules, each module corresponds to a NER subtask:

- entity detection – the identification of a sequence of words that makes up the name of an entity;
- entity classification – the assignation of a category (LOCation, ORGanization, PERson or MISCellaneous) to each detected entity.

Both modules can choose between two types of algorithms – based on knowledge or employing machine learning. As our aim was to create a domain independent NER system, the most reasonable choice was the machine learning approach.

2.1. Classification methods

Considering the previous work in the area of NER described in Section 1.2 and studying the pros and cons of several machine learning methods, we have selected the following three machine learning methods for our system:

- Memory-based learning: suitable when there are not enough training examples and characterized by a fast training and a slow testing phase.
- Maximum Entropy: predicts the example's class with a high degree of certainty, but it is slow during training.
- HMM: lacks in using plenty of features, but has quick the training and testing phases.

The following subsections describe briefly each of these methods individually.

2.1.1. Memory-based learning

Memory-based learning is a supervised inductive learning algorithm for solving classification tasks. It treats a set of training instances as points in a multi-dimensional feature space, and stores them as such in an instance base in memory. Test instances are then classified by comparing them with all instances in memory and calculating a distance function between the test instance x and each of the n memory instances $y_1 \ldots y_n$. With each match the distance, given by the classification with memory-based learning is performed by the *k–NN* algorithm that searches for the k nearest neighbours among the memory instances according to the distance function. The class that the majority of the k nearest neighbors belong to determines the class of the test instance x. Our system employed the default learning algorithm of the TiMBL software package [25], that is instance-based learning with information gain weighting (IB1IG).

---

[6] Named Entity Recognition system of the University of Alicante.

2.1.2. Maximum Entropy

The Maximum Entropy (ME) framework [26,27] estimates probabilities based on the idea of making as few assumptions as possible, other than the imposed constraints. The probability distribution that satisfies the above property is the one with the highest entropy [28]. A classifier obtained by means of ME consists of a set of parameters or coefficients estimated using an optimization procedure. Each coefficient is associated with one feature observed in the training data. The main purpose is to obtain the probability distribution that maximizes the entropy. An advantage of using the ME framework is that even knowledge-poor features may be applied accurately; the ME framework thus allows a virtually unrestricted ability to represent problem-specific knowledge in the form of features. We used a very basic implementation of the toolkit [29], with no smoothing or feature selection.

2.1.3. Hidden Markov Models

Hidden Markov Models (HMM) are stochastic finite-state automata with probabilities associated to the transitions between states and to the emission of symbols from states. The Viterbi algorithm is often used to find the most likely sequence of states for a given sequence of output symbols. In our case, let T be defined as the set of all tags, and Σ the set of all NEs. One is given a sequence of NEs $W = w_1 \ldots w_k \in \Sigma^*$, and is looking for a sequence of tags $T = t_1 \ldots t_k \in T^*$ that maximizes the conditional probability $p(T|W)$, hence we are looking for

$$\arg\max_T p(T|W) = \arg\max_T \frac{p(T)p(W|T)}{p(W)} \tag{1}$$

$p(W)$ is independent of the chosen tag sequence, thus it is sufficient to find

$$\arg\max_T p(T)p(W|T) \tag{2}$$

The toolkit we used is called ACOPOST [7] implemented for the task of POS tagging, but adapted to NER [30].

2.2. Feature description

To perform the Named Entity Recognition task, different sets of features can be used. This section focuses on the description of the characteristics we have been using for the NE detection and classification phases. Both the memory-based learning and the ME machine learning algorithms use identical features. However, as HMM lacks to manage numerous attributes, its set of features contains only of the three most informative ones.

2.2.1. Features for NE detection

The entity detection task consists of determining the NE boundaries. This is possible through the well-known BIO model [31] that assigns to each word a label indicating if a word is at the beginning of a NE (B), inside a NE (I) or outside a NE (O). For the sentence "Juan Carlos está esperando.",[8] according to this model, the following tags have been associated, "B I O O O". *Juan* starts the named entity; *Carlos* continues it and neither *está* nor *esperando* nor the full stop are part of the NE.

The features used in the NE detection task are described in Fig. 1. They are denoted as set A. The set embed lexical, contextual and orthographic information, as well as information provided by manually created trigger lists. As previously studied in [32], in order to improve a classifier's performance, different feature combinations from the original set should be designed and explored.

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

[8] Juan Carlos is waiting.

- **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 present 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 present 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.

2.2.2. Features for NE classification

The task of NE classification has the role of assigning to the previously detected entity the category it belongs to: LOC, ORG, PER or MISC. The feature set employed by this task consists of the first seven features from entity detection (e.g. a, c[1–6], p) plus the set described in Fig. 2. The number of gazetteer entries per class reaches 900. The lists have been manually compiled consulting the Yellow Pages.

2.3. System description

We have described so far the machine learning methods together with the features. This section presents the integration of these methods and the features in what resulted to be the NERUA system. The underlying methodology is represented in Fig. 3. In order to identify the existing entities for a text collection, two stages are performed – Named Entity Detection (NED) and Named Entity Classification (NEC).

The first stage starts with the feature extraction for the entity detection. The text enriched with feature values corresponding to each word is passed to HMM and TiMBL classifiers. Due to high processing time, ME was not used in the detection phase. Its absence is not crucial, as entity delimitation is considered to be easier than entity classification. Classifiers' output are combined through a voting scheme which is described in Section 4.

The second stage has as starting point the text with the identified named entities. Only entities that have been previously detected are going to be classified, therefore, only for these entities the classification feature values are extracted. The performance of the second stage is obviously influenced by the results of the first one.

- ***eP***: entity is trigger for PER
- ***eL***: entity is trigger for LOC
- ***eO***: entity is trigger for ORG
- ***eM***: entity is trigger for MISC
- ***tP***: word ±1 is trigger for PER
- ***tL***: word ±1 is trigger for LOC
- ***tO***: word ±1 is trigger for ORG
- ***gP***: part of NE in PER gazetteer
- ***gL***: part of NE in LOC gazetteer
- ***gO***: part of NE in ORG gazetteer
- ***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***: encodes capitalization and presence of non-letter characters within the entity's tokens

Fig. 2. Features for NE classification.

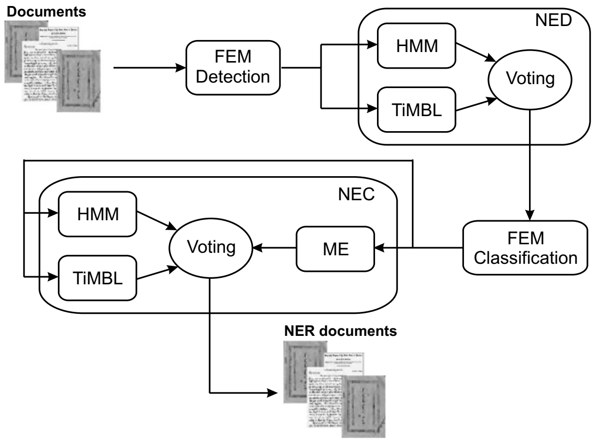

Fig. 3. System description.

The classifiers involved at this stage are: HMM, TiMBL and ME. Each one of them uses labeled training examples, in order to predict the class of the unseen example. Even if the classifiers use the same set of characteristics, they propose for one and the same example various classes. The final outcome is the result of the voting scheme. This second stage yields all the identified NEs together with the class each entity belongs to.

NERUA presents several advantages. One of them is the fact that the feature set it uses at the detection phase does not require language dependent resources.[9] NERUA can be easily adapted to other languages, as demonstrated in [32] when it was adapted to Portuguese. The high performance obtained for Portuguese is also influenced by its close relatedness to Spanish. The system restrains from morphologic and syntactic analysis, which lessens the processing time and eases the adaptation to other languages. Three classifiers process text in parallel and later their output are combined through a voting scheme. This leads to a higher degree of correct label assignment. NERUA has been tested for Spanish, Portuguese and English, and has been involved in real applications, as mentioned in Section 6.2.

3. Experimental setup and discussion of results

The initial experiments were conducted for Spanish using the labelled train and test data of the CoNLL-2002 [31] competition. The train corpus contains 264 715 tokens, out of which 18 794 are entities. The development corpus sums 52 923 tokens out of which 4351 are entities, and the test corpus contains 51 533 tokens and 3558 entities. Scores were computed per NE class with the help of *conlleval*[10] script. The measures are

$$\text{Precision} = \frac{\text{number of correct answers found by the system}}{\text{number of answers given by the system}} \tag{3}$$

$$\text{Recall} = \frac{\text{number of correct answers found by the system}}{\text{number of correct answers in the test corpus}} \tag{4}$$

$$F_{\beta=1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

The next two subsections describe in more detail the experiments we conducted for NER.

3.1. NE detection

Several experiments have been carried out using the provided development data in order to find the best performing combination of features. Following the BIO model described briefly in Section 2.2.1, our experi-

[9] It does not use dictionaries, tools as lemmatizers, stemmers etc.
[10] http://www.cnts.ua.ac.be/conll2002/ner/bin/.

Table 1
NE detection for Spanish – individual classifiers

Tags	B (%)			I (%)			BIO (%)		
Classifier	Prec.	Rec.	$F_{\beta=1}$	Prec.	Rec.	$F_{\beta=1}$	Prec.	Rec.	$F_{\beta=1}$
TMB-*C*24	93.12	93.29	93.20	86.49	84.48	85.48	91.24	90.72	*90.99*
TMB-*C*17	87.76	85.04	86.38	80.92	65.57	72.44	86.03	79.42	82.59
TMB-*C*24*r*	85.93	85.34	85.63	80.25	64.44	71.48	84.53	79.30	81.83
TMB-*E*12	93.43	92.48	92.95	87.81	84.03	85.88	91.81	90.04	90.94
TMB-*E*17	92.73	93.54	93.14	87.70	85.16	86.41	91.32	91.12	*91.22*
HMM-CD	91.10	92.71	91.90	83.16	83.86	83.51	88.82	90.16	89.49
HMM-CW	91.07	92.37	91.72	81.43	82.67	82.05	88.29	89.57	88.92

ments start with a subset of the original set of features A presented in Fig. 1. We will consider only the first 24 features of A and we will call the resulting set $C24 = A/\{aSubStr[1\text{–}5]\}$. The results obtained using this feature set have been satisfactory, as can be seen in Table 1. In order to find the feature set F that maximizes the performance, minimizes the computational cost and is language and resource independent, we have designed more candidate feature sets. We have studied all the features and selected the most informative ones according to the information gain measure. We have developed four candidate sets to validate this study. They are divided into resource dependent and resource independent groups, as some need tools such as lemmatizers and stemmers, and others do not. Each set is denoted with a capital letter, followed by a number which corresponds to the number of features it contains. As set $C24r$ is a reduced representation of the set $C24$, we have placed the index "r" to indicate the reduction.

Language and resource dependent sets

- $C24r = \{a, c[1\text{–}6], C[1\text{–}7], p, aC, wD, wT, aL, aS\}$
- $C17 = C24r/\{c[5\text{–}6], C[6\text{–}7]\}$

Language and resource independent sets

- $E12 = \{a, c[1\text{–}4], C[1\text{–}5], p, aC\}$
- $E17 = E12 \cup \{aSubStr[1\text{–}5]\}$

The results presented in Table 1 show the performance of individual classifiers followed by a simple post-processing stage. The post-processing stage substitutes each appearance of tag I preceded by O with the tag B, only if the analyzed word starts with a capital letter, all the other cases are tagged with O. Sequences such as O B I B I, are transformed into O B I I I.

The labeled corpus we used contains many words that do not form part of an entity (words of tag type O). We have considered studying the efficiency of each feature set for each entity class and also observe the efficiency of the combined BIO tags. Therefore, Table 1 presents the performance measures for B, I and BIO tags.

The language dependent set of features that obtained the best score for the three categories B, I and O, was $C24$ with a F-measure of 90.99%. $E17$ was the best performing language independent set of features with a score of 91.22%. This study of each individual set comes as a consequence of investigation done by various authors, who demonstrated that the combination of different methods improves the results and outperforms the achievement of a single classifier.

The lower accuracy of HMM compared to the accuracy of the best performing TiMBL feature set is compensated by the difference between the number of features the two classifiers use (2 vs. 17). As illustrated by Rössler [33] very few features can be passed to HMM through corpus or tag transformation. We have studied both possibilities and found that tag transformation gives better results. The three most informative attributes have been identified with the help of the information gain measure. These attributes are word capitalization, whole word in capitals and word in the gazetteer list. The tags B and I were modified to also embed the values of these three most informative attributes and this resulted in an performance increase of 2%. In the previous

Table 2
NE classification for Spanish

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-*F*24	69.18	80.00	74.20	63.44	45.62	53.07	73.68	79.88	76.66	90.89	79.21	84.65
TMB-*F*24	64.07	80.20	71.24	54.44	51.01	52.67	75.84	75.35	75.60	89.11	74.96	81.42
ME-*F*24*c*	69.13	79.80	74.08	63.08	46.07	53.25	74.72	79.82	77.19	90.82	80.93	85.59
TMB-*F*24*c*	64.55	79.49	71.25	55.47	50.11	52.66	75.41	75.06	75.24	87.36	75.78	81.16
TMB-*R*24	64.23	80.20	71.33	53.24	44.27	48.34	74.16	76.47	75.30	89.94	74.63	81.57
TMB-*R*24*c*	63.15	80.91	70.94	51.09	42.02	46.12	73.94	74.76	74.35	88.73	74.06	80.73
HMM	64.48	76.85	70.13	48.78	47.90	48.34	73.14	70.82	71.97	58.20	76.12	65.96

table, the experiment using the features word capitalization and word in dictionary is denoted by HMM-CD, and the one using word capitalization and whole word in capitals is denoted by HMM-CW.

3.2. NE classification

NE detection is followed by NE classification. According to the CoNLL-2002 data, the system had to identify the following classes: LOC, MISC, ORG and PER. In the classification task, NERUA makes use of the three classifiers TiMBL, ME and HMM. The first two classifiers were tested using the following sets of characteristics:

- $F24 = \{a, c[1\text{–}6], p, eP, eL, eO, eM, tP, tL, tO, gP, gL, gO, wP, wL, wO, NoE, f, s\}$
- $F24c = F24 \cup \{clx\}$
- $R24 = \{a, c[1], eP, gP, gL, gO, wP, wL, wO, NoE, f\}$
- $R24c = R24 \cup \{clx\}$

We performed different experiments with these feature sets. The obtained results of the development data are shown in Table 2. As one can see, there is no set of characteristics that leads to the best results for all four classes (LOC, PER, ORG and MISC). After validating all results with z' [34] statistics, set $F24c$ for ME was found to reach the best performance for ORG, PER and MISC classes. The sets $R24$ and $R24c$ lowered the performance for MISC in comparison with the sets $F24$ and $F24c$. However, these sets deal with a reduced set of features, thus allowing smaller classifier computational cost. The maximum score of 53.25% for the MISC class is gained by ME with set $F24c$. Among all classifiers, HMM obtains the lowest score per class. HMM fails for example to categorize correctly the entity *Don Simon*, which in a text can designate either a person or an organization. In order to assign the correct category, HMM needs more contextual information.

The elimination or addition of various attributes from different feature sets, has revealed their impact over certain types of classes. On the other hand our system performs well when classifying into PER, ORG and LOC classes, but not when dealing with the MISC class whose nature is quite heterogeneous.

In order to improve the system's performance, this variety of feature sets and classifiers allows us to conduct an adequate voting, as described in Section 4.

4. Classifier combination

In order to develop a system that outperforms the best individual classifiers, we combined the individual classifiers. According to [35], the simplest approach for classifier combination is voting. The output of various ML algorithms is examined and classifiers with weight exceeding a certain threshold are selected. Normally the weight is dependent upon the models that proposed a particular classification.

It is possible to assign various weights to the classifiers, in effect giving more importance to one classifier than to the others. A simple voting scheme gives equal weight to all classifiers and the final decision is taken according to the majority votes. For every examples, the classifiers have assigned a tag. At the end, only the

Table 3
Simple and weighted voting for NE detection

	B (%)			I (%)			BIO (%)		
	Prec.	Rec.	$F_{\beta=1}$	Prec.	Rec.	$F_{\beta=1}$	Prec.	Rec.	$F_{\beta=1}$
Language independent classifier									
TMB-*E*12	94.33	94.91	94.62	87.00	85.29	86.14	92.38	92.30	92.34
TMB-*E*17	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
liSimpleVot	94.32	95.70	95.01	87.94	85.82	86.87	92.64	93.02	92.83
liWeightedVot	94.43	95.73	95.07	88.31	86.05	87.17	92.81	93.10	92.96
Language dependent classifier									
TMB-*C*24	94.42	95.19	94.81	87.25	85.67	86.45	92.51	92.61	92.56
TMB-*C*17	94.47	95.11	94.79	87.28	85.37	86.31	92.56	92.47	92.51
TMB-*C*24*r*	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
ldSimpleVot	94.97	95.53	95.25	88.96	86.27	87.60	93.38	93.02	93.20
ldWeightedVot	95.31	95.36	95.34	88.02	87.56	87.79	93.34	93.24	93.29

Table 4
Simple and weighted voting for NE classification

	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-*F*24	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-*F*24*c*	74.84	75.46	75.15	55.88	50.29	52.94	75.88	79.79	77.79	85.42	85.31	85.36
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
simpleVot	80.29	75.92	78.05	62.41	48.82	54.79	75.57	82.64	78.95	83.27	88.71	85.90
weightedVot	81.16	75.92	78.46	66.80	49.71	57.00	75.06	83.21	78.93	83.72	89.52	86.52

tag with the highest number of votes is selected. For weighted voting, more weight is given to the classifier that reached the best individual performance. Voting schemes were thus used to furtherly improve the base model. In NERUA [24] both simple and weighted voting have been applied.

Tables 3 and 4 present the best performing feature sets described in Section 3 for the CoNLL-2002 test data. The table shows how the classifiers' combination through simple and weighted voting affect the final results.

At the detection stage, several voting techniques have been applied to the language independent and dependent feature sets. Table 3 shows the scores achieved by three language independent and five language dependent sets. When combined through simple and weighted voting, their performances was evaluated with the z' [34] test. According to this measure, among individual and voted classifiers, weighted voting is the one giving better correct class identification. The achieved results are 92.83% for simple voting and 92.96% for weighted voting in the case of language independent sets, and 93.20% for simple and 93.29% for weighted voting in the case of language dependent sets. However, in Table 4 can be seen that simple voting obtains nearly the same results as ME-F24. Though, globally, weighted voting improved the results for all classes except for the MISC.

The results show that different classifiers like TiMBL, HMM and ME can be combined to yield promising results and also to prove that more complicated voting strategies such as weighted voting perform better than simple ones.

5. System comparison

In this section, we describe NERUA's performance compared to other existing systems that participated in the CoNLL-2002 and HAREM-2005 competitions.

Table 5
CoNLL-2002 overall system performance

Classifier	Prec. %	Rec. %	$F_{\beta=1}$ %
Carreras [10]	81.36	81.40	81.39
Florian [36]	78.70	79.40	79.05
NERUA	78.09	79.10	78.59
Cucerzan [37]	78.19	76.14	77.15

5.1. CoNLL-2002

The CoNLL-2002 competition involved language-independent named entity recognition. Participants were supposed to recognize four types of named entities: persons, locations, organizations and miscellaneous. They were offered training and test data and they developed NER systems using machine learning.

The previous sections presented the performance of NERUA considering different machine learning methods. As we used common data sets with the CoNLL-2002 participants, it was reasonable to make a comparative study.

Table 5 presents the first three systems in the CoNLL-2002 competition together with our NER system. Compared to all 12 participant, NERUA scores on third position. This performance is very encouraging since we restrained from morphologic, syntactic or semantic tools and information. Moreover, the gazetteer lists which were consulted, were manually created, limited and with generic nature.

6. Applications

In this section, we want to show the applicability of NERUA in real applications such as Information Retrieval (IR) and Conceptual Modelling (CM). For IR, NERUA has been applied in order to support the IR module in retrieving relevant documents and this application has been proved in the GeoCLEF 2005. In what CM is concerned, NERUA has been used to extract functional requirements. In the followings subsections these applications are presented in detail.

6.1. HAREM

HAREM[11] is a competition that evaluates Named Entity Recognizers for Portuguese. Within the competition, three NE tasks have been proposed:

- detection of NE boundaries;
- semantic classification, that is assigning an appropriate category and subtype to each identified NE;
- morphological classification, realized through an assignation of number and gender to each NE.

All participants were provided with a small golden corpus, previously tagged with the categories described in Table 6.

Our participation in HAREM involved applying our already developed Spanish NER to Portuguese. We have preserved the four NE categories (location, organization, person and miscellaneous)[12] that our system was designed to recognize for Spanish.

In order to apply NERUA to Portuguese which was the target language in the HAREM competition, a Portuguese annotated corpus was needed. Although the HAREM organizing committee has supplied all participants with a small annotated corpus (e.g. 92761 tokens and 5072 NEs), this amount was insufficient to train our system. Therefore, for training purposes, we have merged our already available Spanish corpus with the Portuguese one. The feature sets developed for Spanish were directly ported to detect and classify Portu-

[11] http://poloxldb.linguateca.pt/harem.php.
[12] In HAREM, these classes correspond to local, organizacao, pessoa and abstraccao.

Table 6
HAREM categories and subtypes

Category	Sub-type	English gloss
PESSOA	INDIVIDUAL	Individual person
	CARGO	Title of employment
	MIEMBRO	Members
	GRUPOIND	Group of people
	GRUPOCARGO	Group of titles
	GRUPOMEMBRO	Group of members
ORGANIZACAO	ADMINISTRACAO	Administration
	INSTITUICAO	Institution
	EMPRESA	Company
	SUB	Sub-organization
TEMPO	DATA	Date
	HORA	Time
	PERIODO	Period
	CICLICO	Cyclic
LOCAL	CORREIO	Address
	ADMINISTRATIVO	Administrative
	GEOGRAFICO	Geographic
	VIRTUAL	Virtual
	ALARGADO	Extended
OBRA	PRODUCTO	Product
	REPRODUZIDA	Reproducible work
	ARTE	Unique work
	PUBLICACAO	Publication
ACONTECIMENTO	EFEMERIDE	Unique
	ORGANIZADO	Large event
	EVENTO	Atomic event
ABSTRACCAO	DISCIPLINA	Subject
	MARCA	Brandname
	ESTADO	Condition
	ESCOLA	School
	IDEIA	Ideal
	PLANO	Plan
	OBRA	Complete works
	NOME	Name
COISA	OBJECTO	Object
	SUBSTANCIA	Substance
	CLASSE	Class
VALOR	CLASSIFICACAO	Classification
	QUANTIDADE	Amount
	MOEDA	Money
VARIADO		Other

guese NEs. This was possible due to the proximity and the common characteristics of the Spanish and Portuguese languages.

Among 13 participants, NERUA came fifth in the task of NE detection and classification. It obtained better results in the identification task compared to the classification one. This is due to the lack of annotated resources for Portuguese and the fact that we have focused on the recognition of only four types of entities. In this competition we showed that our NER system, initially designed and developed for Spanish was adapted with little effort to Portuguese and achieved promising results.

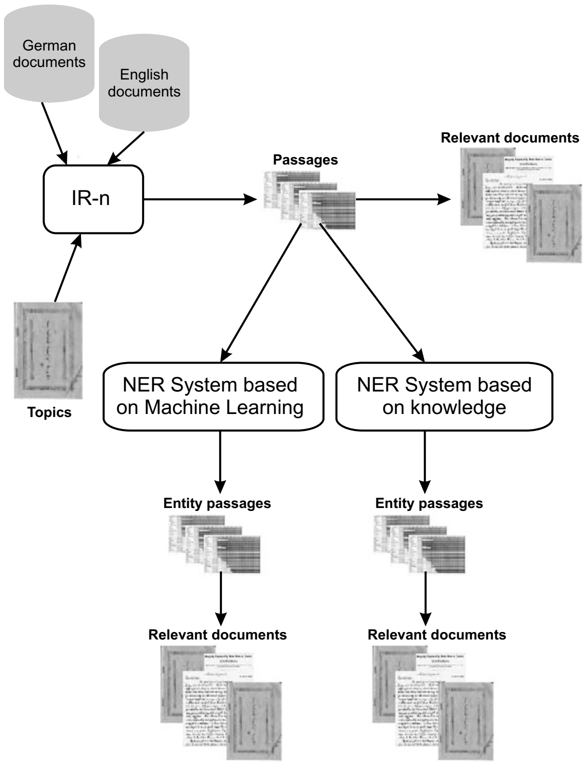

Fig. 4. GeoCLEF system architecture.

6.2. Application of NERUA in IR

The CLEF conference has introduced in 2005 the first Cross-Language Geographic Information Retrieval task, called GeoCLEF [38]. Geographical Information Retrieval (GIR) concerns the retrieval of information involving spatial awareness. Geographical references (georeferences) can play an important role in IR, as many documents contain various types of spatial references. The GeoCLEF monolingual and bilingual tasks require that the participating GIR systems retrieve relevant documents by involving geographic information related to places, events, etc. The aim of GeoCLEF is to provide the framework to evaluate GIR systems in information retrieval tasks involving both spatial and multilingual aspects.

Our participation in GeoCLEF [39] was performed with a system combining three modules: an Information Retrieval (IR) module [40] (IR-n), a NER based on machine learning (NERUA) and a NER based on knowledge (DRAMNERI) [41]. Fig. 4 illustrates the system's architecture. The IR system was combined first with the ML-based and then with the rule-based NER, thus allowing a comparison between the performance of NERUA against DRAMNERI in the context of IR. Although the machine-learning and rule-based NER systems used the same gazetteer lists, NERUA outperformed DRAMNERI in every task.

The Named Entity Recognizers were mainly applied to recognize only the location entities in the text collection. The Named Entity Recognizers process the returned passages of the IR system and then annotate the location entities. In this way, the NEs provide extra information and assign higher relevance score to the retrieved passages.

In order to participate with our system in the GeoCLEF monolingual English and bilingual X2English[13] tasks, we have trained our NER system for English using the CoNLL-2003 resources [42]. NERUA performed better than DRAMNERI when combined with IR-n. Table 7 illustrates the increase in performance for all

[13] The X2English bilingual tasks consist of retrieving English documents belonging to a certain topic specified in either Spanish, Portuguese or German.

tasks when NERUA was added to IR-n. When applying NERUA to the outcome of IR-n, the results improve with more than 2% in the monolingual task and with 1% in the bilingual task. The results could improve more if NERUA's capabilities were combined with a geographical knowledge source embedding relations of the type "X city is in Y country".

Thanks to the participation in GeoCLEF, we were able to investigate and evaluate NERUA's influence over an Information Retrieval system. At the same time we have demonstrated NERUA's easy portability to other languages such as English.

6.3. Application of NERUA in conceptual modelling

Another area of interest for us was the applicability of NERUA for Conceptual Modelling purposes. In this section, we describe how Named Entities Recognition can be used to clearly extract functional requirements from documents and systematically decompose these high-level software requirements into a more detailed specification that constitutes the conceptual schema of the desired system. We name this extension NER4CM (NER for Conceptual Modelling) and we view it as a plug-in for CASE tools. Our goal here is to develop a semiformal model that allows software engineers to systematically produce Object Oriented (OO) models starting from a collection of documents that constitute the specification requirements. We focus on designing the methodology to develop the OO model from a textual problem description.

Fig. 5 illustrates this methodology. The input consists of a set of documents describing the specification of requirements for a software system. The first step is preparing the text finding synonyms and homonyms with the help of a Word Sense Disambiguation tool [43] and producing as a result semantically annotated text. At the second stage the NERUA system extracts the information required for the Conceptual Modelling process, that is the set of named entities from the enhanced textual description. As a result, a set of entities classified as Persons, Locations or Organizations are obtained. The last step is the organization of the information in a structured way by means of an XML specification. Therefore, the resulted specification can be imported into a CASE tool, thus allowing the analyst to use the Named Entity Recognition and annotation process in developing the Object Model (OM).

As a proof of concept to verify the applicability of this process, a running example using Google news service (in Spanish) is presented. Let's suppose we have the following news, "*Santaclara, 26 de Julio de 2005. Fred Amoroso Presidente y CEO de Macrovision Inc. en USA anuncia la adquisicion de Trymedia Systems Inc. En el tercer cuarto de 2005 por un importe de 34M €*".

NER4CM processes this news according to the methodology described above and, as a result, an XML specification that contains the relevant information is generated. Then, this XML specification is imported in a CASE environment in order to develop the OM. The experiments have been performed in the Enterprise Architect CASE environment. Fig. 6 illustrates a snapshot of this tool with the NER4CM plug-in. Once the designer has applied the plug-in and imported the generated XML specification, a new view is added to the system. This is the NER4CM view (that can be observed in the Project View tree), and it shows the different sets of named entities that the designer can use in defining the OM. The first step the designer has to complete is assigning the semantic relations between these named entities. Considering our example, the designer assigns the following relations: Fred Amoroso is the PresidentOf Macrovision, Santa Clara is partOf USA and Santa Clara is the HQ of Macrovision. Once this step has been completed, the designer has all the relevant information needed to define the structural part of the OM.

Table 7
GeoClef 2005 officials results for monolingual and bilingual tasks

Task	Avg. precision (%)	Ranking	Number of participants	Best avg. precision (%)
Monolingual English (only IR-n)	32.53	3rd	11	39.36
Monolingual English (IR-n + NERUA)	34.95	3rd	11	39.36
Bilingual X2English (only IR-n)	30.83	3rd	5	37.15
Bilingual X2English (IR-n + NERUA)	31.78	3rd	5	37.15

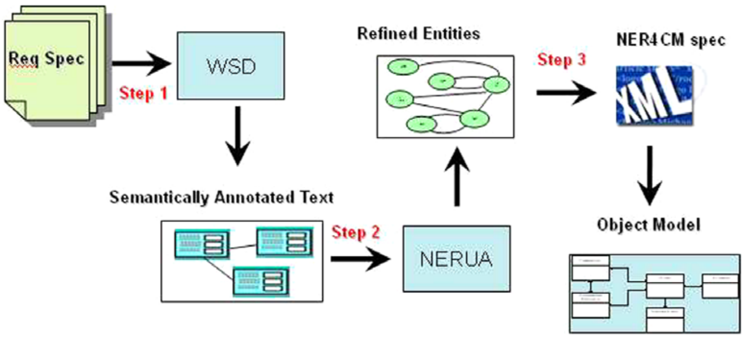

Fig. 5. Steps needed to create an OM.

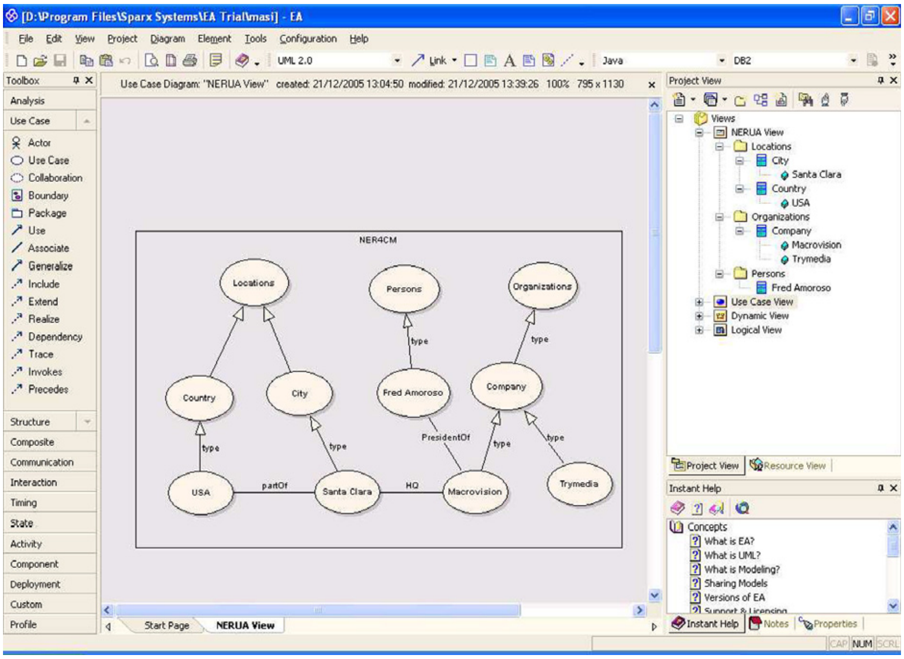

Fig. 6. Snapshot of the EA tool with the NER4CM plug-in.

NER4CM's main drawbacks are related to the deficiencies in the specification requirements. It works on the assumption that the textual description is correct. However, if problems are detected during any of the processing steps, the textual description needs to be refined furtherly.

7. Conclusions and future work

In this paper, we presented a completely automatic Named Entity Recognition approach, based on three different machine learning techniques (Memory-based learning, Hidden Markov Models and Maximum Entropy). We have examined the performances of each method individually for the NER task, as well as we studied the best combination among them using various voting strategies.

The system was initially developed for Spanish, but afterwards it was easily adapted to other languages, such as Portuguese and English. This was possible as we restrain from morphologic, syntactic and semantic analyzers.

Our system was evaluated in two NER scientific competitions, such as CoNLL-2002 and HAREM-2005. This evaluation revealed a good performance of the system for the LOCation, PERson and ORGanization categories. However, the MISCellaneous class did not obtain a good coverage, due to its heterogeneous nature. In the future to obtain better coverage of the MISC names, we will divide them into subcategories.

We have also evaluated the contribution of our trilingual NER system in the geographic Information Retrieval competition (GeoCLEF). The obtained results are very promising. Moreover, we have applied NERUA in a real tool for Conceptual Modelling.

In the future, we will focus on the improvement and tuning of the features needed for the classification phase. We will adapt the system to other languages and we will incorporate more complicated voting schemes to yield a better coverage. Our future intentions involve also extending the system to more categories and resolving the so-called weak named entities.

References

[1] R. Grishman, B. Sundheim, Message understanding conference-6: a brief history, in: Proceedings of the 16th Conference on Computational Linguistics, Copenhagen, Denmark, 1996, pp. 466–471.

[2] S. Sekine, K. Sudo, C. Nobata, Extended named entity hierarchy, in: Proceedings of the Third International Conference on Language Resources and Evaluation, LREC2002, Las Palmas, Canary Islands, Spain, 2002.

[3] M. Arevalo, M. Civit, M.A. Martí, Mice: a module for named entity recognition and classification, International Journal of Corpus Linguistics 9 (1) (2004) 53–68.

[4] S. Sekine, R. Grishman, H. Shinnou, A decision tree method for finding and classifying names in japanese texts, in: Proceedings of the 6th Workshop on Very Large Corpora, WVLC-98, Montreal, Canada, 1998.

[5] D.M. Bikel, S. Miller, R. Schwartz, R. Weischedel, Nymble: a high-performance learning name-finder, in: Proceedings of the Fifth Conference on Applied Natural Language Processing, Morgan Kaufman Publishers Inc., San Francisco, CA, USA, 1997, pp. 194–201.

[6] G. Zhou, J. Su, Named entity recognition using an hmm-based chunk tagger, in: ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, Pennsylvania, 2002, pp. 473–480.

[7] A. Borthwick, J. Sterling, E. Agichtein, R. Grishman, Exploiting diverse knowledge sources via maximum entropy in named entity recognition, in: Proceedings of the 6th Workshop on Very Large Corpora, WVLC-98, Montreal, Canada, 1998.

[8] M. Asahara, Y. Matsumoto, Japanese named entity extraction with redundant morphological analysis, in: NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Edmonton, Canada, 2003, pp. 8–15.

[9] M. Collins, Ranking algorithms for named-entity extraction: boosting and the voted perceptron, in: ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, Pennsylvania, 2002, pp. 489–496.

[10] X. Carreras, L. Màrques, L. Padró, Named entity extraction using adaboost, in: Proceedings of CoNLL-2002, Taipei, Taiwan, 2002, pp. 167–170.

[11] A. McCallum, W. Li, Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons, in: W. Daelemans, M. Osborne (Eds.), Proceedings of CoNLL-2003, Edmonton, Canada, 2003, pp. 188–191.

[12] M. Padró, L. Padró, A named entity recognition system based on a finite automata, Procesamiento del Lenguaje Natural 35 (2005) 319–326.

[13] D. Maynard, V. Tablan, C. Ursu, H. Cunningham, Y. Wilks, Named entity recognition from diverse text types, in: R. Mitkov, N. Nicolov, G. Angelova, K. Bontcheva, N. Nikolov (Eds.), Recent Advances in Natural Language Processing, RANLP 2001, Tzigov Chark, Bulgaria, 2001.

[14] D. Shen, J. Zhang, J. Su, G. Zhou, C.L. Tan, Multi-criteria-based active learning for named entity recognition, in: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004), Barcelona, Spain, 2004, pp. 589–596.

[15] M. Collins, Y. Singer, Unsupervised models for named entity classification, in: Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, 1999.

[16] A. Blum, T. Mitchell, Combining labeled and unlabeled data with co-training, in: COLT' 98: Proceedings of the Eleventh Annual Conference on Computational Learning Theory, Madison, Wisconsin, United States, 1998, pp. 92–100.

[17] Z. Kozareva, B. Bonev, A. Montoyo, Self-training and co-training applied to spanish named entity recognition, in: Proceedings of the Fourth Mexican International Conference on Artificial Intelligence, MICAI 2005, Monterray, Mexico, 2005.

[18] Z. Kozareva, A. Montoyo, Learning spanish named entities using unlabeled data, in: G. Angelova, K. Bontcheva, R. Mitkov, N. Nikolov, N. Nikolov (Eds.), Proceedings of the Recent Advances in Natural Language Processing, RANLP 2005, Botovetz, Bulgaria, 2005, pp. 284–287.

[19] B. Mohit, R. Hwa, Syntax-based semi-supervised named entity tagging, in: Proceedings of the ACL Interactive Poster and Demonstration Sessions, Association for Computational Linguistics, Ann Arbor, Michigan, 2005, pp. 57–60.

[20] L. Márquez, A. de Gispert, X. Carreras, L. Padró, Low-cost named entity classification for catalan: exploiting multilingual resources and unlabeled data, in: Proceedings of the 1st ACL Workshop on Multilingual and Mixed-language Named Entity Recognition: Combining Statistical and Symbolic Models, MulNER'03, 2003.

[21] S. Abney, Bootstrapping, in: ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, Pennsylvania, 2002, pp. 360–367.
[22] R. Florian, A. Ittycheriah, H. Jing, T. Zhang, Named entity recognition through classifier combination, in: W. Daelemans, M. Osborne (Eds.), Proceedings of CoNLL-2003, Edmonton, Canada, 2003, pp. 168–171.
[23] D. Wu, G. Ngai, M. Carpuat, A stacked, voted, stacked model for named entity recognition, in: W. Daelemans, M. Osborne (Eds.), Proceedings of CoNLL-2003, Edmonton, Canada, 2003, pp. 200–203.
[24] O. Ferrández, Z. Kozareva, A. Montoyo, R. Muñoz, Nerua: sistema de detección y clasificación de entidades utilizando aprendizaje automático, Procesamiento del Lenguaje Natural 35 (2005) 37–44.
[25] W. Daelemans, J. Zavrel, K. van der Sloot, A. van den Bosch, Timbl: Tilburg memory-based learner, Tech. Rep. ILK 03-10, Tilburg University, November 2003.
[26] R. Lau, R. Rosenfeld, S. Roukos, Adaptative statistical language modeling using the maximum entropy principle, in: Proceedings of the Human Language Technology Workshop, ARPA, 1993.
[27] A.L. Berger, S.A.D. Pietra, V.J.D. Pietra, A maximum entropy approach to natural language processing, Computational Linguistics 22 (1) (1996) 39–71.
[28] A. Ratnaparkhi, Maximum entropy models for natural language ambiguity resolution, Ph.D. Thesis, Computer and Information Science Department, University of Pennsylvania, 1998.
[29] A. Suárez, M. Palomar, A maximum entropy-based word sense disambiguation system, in: H.-H. Chen, C.-Y. Lin (Eds.), Proceedings of the 19th International Conference on Computational Linguistics, COLING 2002, 2002, pp. 960–966.
[30] I. Schröder, A case study in part-of-speech tagging using the icopost toolkit, Tech. Rep. FBI-HH-M-314/02, Department of Computer Science, University of Hamburg, 2002.
[31] E.F. Tjong Kim Sang, Introduction to the CONLL-2002 shared task: language-independent named entity recognition, in: D. Roth, A. van den Bosch (Eds.), Proceedings of CoNLL-2002, Taipei, Taiwan, 2002, pp. 155–158.
[32] Z. Kozareva, O. Ferrández, A. Montoyo, R. Muñoz, Using language resource independent detection for spanish named entity recognition, in: G. Angelova, K. Bontcheva, R. Mitkov, N. Nikolov, N. Nikolov (Eds.), Proceedings of the Recent Advances in Natural Language Processing, RANLP 2005, Botovetz, Bulgaria, 2005, pp. 279–283.
[33] M. Rössler, Using markov models for named entity recognition in German newspapers, in: Proceedings of the Workshop on Machine Learning Approaches in Computational Linguistics, Trento, Italy, 2002, pp. 29–37.
[34] T.G. Dietterich, Approximate statistical tests for comparing supervised classification learning algorithms, Neural Computation 10 (7) (1998) 1895–1923.
[35] H. van Halteren, J. Zavrel, W. Daelemans, Improving accuracy in word class tagging through the combination of machine learning systems, Computational Linguistics 27 (2) (2001) 199–230.
[36] R. Florian, Named entity recognition as a house of cards: classifier stacking, in: D. Roth, A. van den Bosch (Eds.), Proceedings of CoNLL-2002, Taipei, Taiwan, 2002, pp. 175–178.
[37] S. Cucerzan, D. Yarowsky, Language independent NER using a unified model of internal and contextual evidence, in: D. Roth, A. van den Bosch (Eds.), Proceedings of CoNLL-2002, Taipei, Taiwan, 2002, pp. 171–174.
[38] F. Gey, R. Larson, M. Sanderson, H. Joho, P. Clough, Geoclef: the clef 2005 cross-language geographic information retrieval track, Working Notes in CLEF 2005.
[39] O. Ferrández, Z. Kozareva, A. Toral, E. Noguera, A. Montoyo, R. Muñoz, F. Llopis, The University of Alicante at geoclef 2005, Working Notes in CLEF 2005.
[40] F. Llopis, Ir-n un sistema de recuperación de información basado en pasajes, Procesamiento del Lenguaje Natural 30 (2003) 127–128.
[41] A. Toral, Dramneri: a free knowledge based tool to named entity recognition, Proceedings of the 1st Free Software Technologies Conference, 2005, pp. 27–32.
[42] E.F. Tjong Kim Sang, F. De Meulder, Introduction to the CONLL-2003 shared task: language-independent named entity recognition, in: W. Daelemans, M. Osborne (Eds.), Proceedings of CoNLL-2003, Edmonton, Canada, 2003, pp. 142–147.
[43] A. Montoyo, MTtodo basado en marcas de especificidad para WSD, Procesamiento del Lenguaje Natural 26 (2000) 53–60.

Zornitsa Petrova Kozareva (1982) is a Ph.D. student in Computer Science at the University of Alicante, Spain. Since 2004 she works in the Department of Software and Computing Systems (GPLSI division) as a researcher with a fellowship from the Rector of Investigation at the University of Alicante. Her research interests are focused on Natural Language Processing and Machine Learning, in concrete Named Entity Recognition and Textual Entailment. She has published 15 papers in International Conferences.

Óscar Ferrández Escámez (1980). Ph.D. student in Computer Science by the University of Alicante. He is working since 2004 in the Department of Software and Computing Systems (GPLSI division) at this University as a researcher with a fellowship from the Valencia Government. His research interests are focused on Computational Linguistics and Natural Language Processing and more precisely in Named Entity Recognition and Textual Entailment. He has already published nine papers in International Conferences.

Andrés Montoyo is full Professor and member of the Department of Software and Computing Systems (GPLSI division) (since 1992). His scientific interests include lexical and semantic analysis, word sense disambiguation, Name Entity Recognition, information extraction and information retrieval applications. He has been the main researcher in the development of different Word Sense Disambiguation systems (Specification Marks) and cross-language Geographic Information Retrieval (GIR) systems (GIRUA) that have been evaluated in several CLEF evaluations. He is editor of the special issue on Natural Language and Information Systems held by the journal Data and Knowledge Engineering. He has been a member of programme committees of several international conferences (NLDB, RANLP). He is responsible for several national research projects and has edited several books, and contributed more than 40 papers to several journals and conferences.

Rafael Muñoz (1967) is working since 1996 in Software and Computing Systems at the University of Alicante as a researcher and lecturer. Moreover, he is the head of Technology Transfer Office of University of Alicante. His research areas are: Computer Linguistics and Natural Language Processing and in concrete Information Extraction and Retrieval, Question Answering, Temporal information applied to Question Answering, Named Entity Recognition and Anaphora resolution (definite description). He was conference chairman in several conferences to note: NLDB'05. He is a member of the Research Group on Language Processing and Information systems at the University of Alicante since 1996 and member of Natural Language Processing Spanish Society. He wrote two book, edited three books and published more than 40 papers in journals and conferences to note Computational Linguistics and DKE journal and international conferences like ACL in 2004.

Armando Suárez (1965) is working since 1992 in Software and Computing Systems at the University of Alicante as a researcher and lecturer. His research areas are: Computer Linguistics and Natural Language Processing and in concrete Word Sense Disambiguation and semantic annotation. He is a member of the Research Group on Language Processing and Information systems at the University of Alicante since 1996 and member of Natural Language Processing Spanish Society. He published more than 40 papers in journals and conferences to note Journal of Artificial Intelligence Research and international conferences like COLING in 2002.

Jaime Gómez is a Professor of Software Engineering at the Computer Science School at the University of Alicante. He received a B.S. in Computer Science from Technical University of Valencia in 1992 and Ph.D. in Computer Science from the University of Alicante in 1999. His current research interests include, Application of NLP to Web Engineering, Semantic Web and Model-Driven Engineering. For more than eight years, his research group has develop methods and tools to facilitate software development in Industry. He is author of more than 50 papers in Journals and International Conferences.