A Resource-Based Method for Named Entity Extraction and Classification *

Pablo Gamallo and Marcos Garcia

Centro de Investigação em Tecnologias da Informação (CITIUS),
Universidade de Santiago de Compostela, Galiza, Spain
pablo.gamallo@usc.es marcos.garcia.gonzalez@usc.es

Abstract. We propose a resource-based Named Entity Classification (NEC) system, which combines named entity extraction with simple language-independent heuristics. Large lists (gazetteers) of named entities are automatically extracted making use of semi-structured information from the Wikipedia, namely infoboxes and category trees. Language-independent heuristics are used to disambiguate and classify entities that have been already identified (or recognized) in text. We compare the performance of our resource-based system with that of a supervised NEC module implemented for the FreeLing suite, which was the winner system in CoNLL-2002 competition. Experiments were performed over Portuguese text corpora taking into account several domains and genres.

1 Introduction

Named Entity Recognition and Classification (NERC) is the process of identifying and classifying proper names of people, organizations, locations, and other Named Entities (NEs) within text. NERC is a crucial task in several natural language applications, namely Question Answering and Information Extraction. This paper will be focused on the second step of the task, i.e., on Named Entity Classification (NEC). Note that we use here the term NER (Named Entity Recognition) in a narrow sense: it is defined as the process of just identifying NEs.

Most approaches for NEC adopt machine learning techniques as a way to automatically induce statistic classifiers starting from a collection of training examples. The main drawback of these supervised techniques is the requirement of a large amount of annotated corpora. The unavailability of such corpora and the high cost required to build them lead to search for alternative resource-based methods.

In this paper, we propose a NEC system requiring no human intervention such as manually labeling training data (supervised learning) or manually creating gazetters (i.e., repositories of named entities). This system combines named

* This work has been supported by Ministerio de Educació y Ciencia (Spain), within the project Ontopedia, with reference : FFI2010-14986.

entity extraction with simple language-independent heuristics for named entity disambiguation. In order to extract named entities, we use semi-structured information from the Wikipedia, namely infoboxes and category trees. This technique allows us to create large gazetteers of entities, such as lists of persons, organizations, and locations. The second step uses language-independent rules to classify entities in the context of a given text (i.e., entity disambiguation). Only disambiguation among different homonyms is considered (e.g., "Austin" as town or person). Polysemy and metonymy of proper names are not taken into account. We compare the performance of our resource-based system with that of a supervised NEC module implemented for the FreeLing suite [7, 9], which was the winner system in CoNLL-2002 competition. Experiments were performed over Portuguese text corpora considering several domains and genres.

More precisely, the major contributions of this paper are the following:

- adding a new Portuguese NEC module to the FreeLing package,
- comparing a supervised NEC method with a resource-based strategy,
- analyzing the portability of both strategies to new domains and textual genres.

The article is organized as follows. The following section (2) introduces some related work. Then, Section 3 describes and justifies the classification criteria employed by our NEC systems. Next, Section 4 describes in detail the resource-based method and, in Section 5, we report the experiments performed on several Portuguese corpora. Finally, some conclusions are put forward in 6.

2 Related Work

The current dominant methods to named entity identification and classification are based on supervised learning. This is evidenced by the fact that most of the 28 systems presented at both CoNLL-2002 and CoNLL-2003 rely on supervised strategies. These learning strategies consist in creating disambiguation rules (classifier) based on discriminative features found in an annotated corpus (training corpus). The variants of this general strategy include different learning algorithms: Boosting [9], Support Vector Machines (SVM) [3], or Conditional Random Fields (CRF) [17].

Alternative systems are based on resource-based techniques that can classify named entities without prior training. These techniques require external resources such as WordNet [1] or gazetteers [18]. The latter work is very close to our proposal. As our system, [18] describes a method consisting of two modules. The first one automatically creates gazetteers of entities, and the second one uses language and domain independent heuristics for entity classification. However, there is a significant difference between their system and our proposal: their extractor of gazetteers needs some manual supervision. In particular, their extraction strategy requires some lists of seed named entities to generate queries and retrieve Web pages containing occurrences of the seed entities. By contrast,

we do not need to manually define any prior seeds since our extraction strategy takes advantage of the semi-structured information found in Wikipedia.

In this sense, we must mention there exists recent interesting work using Wikipedia as gold standard corpora to train supervised NEC classifiers [19].

Finally, there also exist several NEC systems for Portuguese language. Most are rule-based, language dependent approaches [5, 2, 23, 10], few supervised systems [11], and even one hybrid (stochastic and rule-based) method [12]. None of them is based on a resource-based strategy.

3 Classification Criteria

Since the MUC-6 competition [13], the main three semantic classes of proper names used for the NEC task are "persons", "locations", and "organizations". These classes were known as "enamex". Then, in CoNLL 2002 [24] the type "miscellaneous" was included to encode proper names falling outside the three "enamex" categories. Temporal expressions and some numerical expressions such as amounts of money and other types of units are also accepted as NEs in many NEC tasks.

The criteria given to the annotators to tag proper names in context, using a predefined set of classes, may change considerably according to the guidelines of the competition. One of the main problems arising in annotation is metonymy/polysemy. For instance, it is common to use names of countries, cities, or other locations to make reference to some kind of organization ("*Germany* signed the treaty", "In the morning *Tokyo* lost 3.7%"), a group of people ("*Spain* is against the Iraq war"), or even abstract entities such as economic systems ("*Portugal* grew 0.2% in the second quarter"), complex structures and cultural entities ("I miss *Portugal*"), etc. Different classification criteria have been used in previous competitions: in MUC-6 metonymy/polysemy was not taken into account, only homonyms were considered. In CoNLL, only some metonymy types were distinguished. For instance, countries referring to their governments are annotated as organizations: in "*Germany* signed the treaty" the proper name "Germany" is classified as an organization and not as a location. However, metonymies dealing with other types of organizations were not considered: in "Tokyo lost 3.7%", "Tokyo" is taken as the name of a city and then a location. Finally, in HAREM[22] were considered many types of metonymy and polysemy. For example, the proper name "Portugal" in "I miss *Portugal*" is annotated as an abstract entity. In this competition, countries are perceived as very ambiguous words and, in consequence, may be annotated not only as locations, but also as organizations, groups of people, or even abstract entities.

In our experiments, we decided to annotate training and test corpora considering only homonymy, and disregarding metonymic interpretations of NEs. Such a decision was motivated by several reasons, which are described in the following subsections.

3.1 Criticism in Lexical Semantics

NEC is perceived as a specific type of Word Sense Disambiguation (WSD). One of the main drawbacks of most WSD tasks is that they are based on the naive enumerative model of word meaning [21]. According to this model, the meaning of an ambiguous word is a set of senses, and one of them is selected in the context of the utterance. Differences between unrelated and related senses, that is, between homonymy and metonymy/polysemy, are not taken into account by the disambiguation procedure. By contrast, other lexical models try to represent the range of possible word senses in a more compact way than by enumeration. In particular, [20] distinguishes two representations:

- underspecified representations for a polysemous word,
- a list of alternative (or disjoint) readings for homonyms.

Polysemous words are specified (or *precisified*) under certain conditions, while homonyms are enumerated and disambiguated as in the well-known WSD models. Similar assumptions can be found in [15]. So, following these approaches, we will use classic WSD techniques, not to identify metonymy interpretations or very specific senses of NEs, but just to disambiguate their homonyms. Metonymy resolution and sense *precisification* are much more complex tasks requiring more complex and sophisticated techniques which are beyond the objective of our work.

3.2 Polysemy and Homonymy in Psycholinguistics

Psycholinguistic experiments seem to prove that polysemy and homonymy are different phenomena. In [4], the experiments performed offer neurophysiological support for modelling homonymy by means of different mental entries, while polysemy is compacted in single entries. Similarly, the experiments described in [16] revealed different cognitive processing strategies depending on the type of lexical ambiguity (homonymy or polysemy). So, the use of the same WSD technique for dealing with both lexical ambiguities seems to be not very appropriate.

3.3 Identity Criteria in Formal Ontology

There are some work on Formal Ontology [6, 14] distinguishing between lexical (metonymy/polysemy) and ontological relations (IS-A, which is used to classify entities). The Formal Ontology framework criticizes the abuse of IS-A roles and multiple inheritance to deal with lexical polysemy. This results in overloading ontologies. To simplify ontologies and classification, they define the IS-A relation by means of the notion of *identity criteria*. Individuals with different identity criteria are in different classes, even if they can be semantically related by different kinds of dependencies, colocalizations, etc. For instance, a university is a social organization, and not a place or building, according to their functional criteria of identity. If its functional identity is destroyed, then the organization

ceases to exist, even if the place or building is not destroyed. From this viewpoint, a university should be always classified as a functional organization, even if its *dependent* entities (location or group of people) can be highlighted in some linguistic contexts.

3.4 Encyclopedic Organization

As in the case of traditional dictionaries, in encyclopedias only homonyms are separated in different entries. For instance, if the same name is used for two different individuals, the encyclopedia defines two separated entries. By contrast, the location, population, and government of a country are not separated in different entries. All this information is organized within a single one. But there exist some borderline cases where the difference between homonymy and polysemy is not very clear, for instance the case of national football teams. Should they be a component of countries? or should be considered as separated entries? In Wikipedia, a national football team is assigned a single entry, so it is not perceived as a part of the country. We will follow the same convention.

3.5 Commercial NEC Systems

Finally, we must point out that many commercial NEC systems make classification without polysemy: Alchemy[1], Extractiv[2], and Daedalus[3].

4 Resource-Based NEC System

The NEC system we propose classifies 4 types of named entities: persons, locations, organizations, and other entities ("miscellaneous"). It can be considered as a resource-based strategy which consists of two tasks. First, three large lists of NEs (persons, locations, and organizations) are automatically generated with the aid of semi-structured information from Wikipedia. Second, some disambiguation rules are applied on previously identified NEs, in order to solve homonyms and unknown NEs. Even if our experiments will be focused on Portuguese, the disambiguation rules we propose can be considered as (almost) totally independent on a specific language and knowledge domain.

4.1 Automatic Generation of Gazetteers and Trigger Words

The main objective here is to generate three lists of NEs, one for each semantic class, by exploiting both the category trees and infoboxes of Wikipedia. In particular, category trees and infoboxes will allow us to identify common nouns referring to different subclasses of persons, locations, and organizations. This way, the extraction task consists of two steps: first, we select common nouns (trigger words) denoting subclasses of the three target classes and, then, by means of these subclasses we extract the lists of NEs (gazetteers).

[1] `http://www.alchemyapi.com/api/entity/`
[2] `http://extractiv.com`
[3] `http://www.daedalus.es/productos/stilus/stilus-ner.html`

Subclasses In the first step, the goal is to search into the category tree of Wikipedia a set of categories which are subclasses of persons, locations, and organizations. The strategy is the following: we identify the categories containing in the *head* position the words "People" "Places", and "Organizations" as well as their synonyms, and then, we extract the *head* of their hyponyms. Let us see how, in Portuguese, we extract `Partidos` (Parties) as a subclass of `Organizações` (Organizations). First, we select the generic category `Organizações políticas` (political organizations), since it contains in the *head* position the target category `Organizações`. Then, we search its hyponyms and identify, among others, the category `Partidos políticos` (political parties). Finally, we extract the head of its expression, namely `Partidos` (parties), and put it in the list of subclasses of organizations. Table 1 shows a sample of the Portuguese subclasses selected for each target class.

Persons	**Locations**	**Organizations**
`Misses, Políticos,`	`Terra, Planeta, Mundos,`	`Instituições, Partidos,`
`Designers, Professores,`	`Locais, Ilhas, Cidades,`	`Federações, Associações,`
`Personagens, Criminosos,`	`Subdivisões, Rios,`	`Sindicatos, Clubes,`
`Chefes, Escritores,`	`Países, Monumentos,`	`Entidades, Empresas,`
`Artistas, Treinadores`	`Hoteis`	`Cooperativas`

Table 1. Portuguese subclasses of Persons, Locations, and Organizations

These lists of nouns will be used, on the one hand, as *trigger* words (after lemmatization) in the disambiguation process (see section 4.2) and, on the other, as seeds to generate the gazetteers.

Gazetteers The second step consists in extracting those NEs considered as instances of the selected subclasses. Two strategies were implemented.

The first one verifies whether the set of categories of each Wikipedia article contains one or more of the selected subclasses. If a subclass is contained in the set of categories, then the title of the article, which is a named entity, is classified as an instance of this subclass and, then, of the corresponding generic class. For instance, let us suppose that the article with the title `Rui Zink` is assigned the category `Escritores de Portugal` (writers of Portugal). As this category contains `Escritores` (writers), which is a subclass of *Persons*, then we add `Rui Zink` to the list of persons.

The second strategy follows the same procedure but, instead of checking the set of categories of an article, we search within the "attribute-value" structure of infoboxes. If one of the subclasses is contained in the value of an infobox, then we add the title of the article to the list of the corresponding generic class. To filter out noise, it is possible to restrict the search by using only those values tagged with some specific infobox attributes, e.g., "type", "occupation", etc.

This process let us generate three gazetteers of NEs: a list of people, a list of locations, and a list of organizations. In case of homonymy, a NE can be in more than one list. Finally, the list of subclasses are lemmatized and used as *trigger words* in the process of disambiguation.

4.2 Disambiguation and Classification

The input of our NEC system is PoS tagged text containing single and composite proper names already identified as NEs. In addition, two external resources are required: both the gazetteers and trigger words automatically generated from Wikipedia.

Given an identified NE, the algorithm we use to select a semantic class can be informally described as follows:

list lookup strategy: if the NE matches an entry appearing in only one gazetteer, then it can be considered as an unambiguous NE and can be assigned the class of the gazetteer.

contextual checking: if the NE appears in various gazetteers (homonymy) or it is an unknown NE (missing in gazetteers), then we search within its linguistic context for relevant trigger words. In particular, we check if the words appearing to the left and to the right of the target NE (in our experiments, the window size is 3) match the lists of trigger words. For instance, the NE "Austin" will be classified as a location in the context "Austin, a town in ..." because the common noun "town" is a trigger word in the list of locations. If there are several trigger words of different classes in the context of the target NE, we give preference to the closest one. If there are two triggers at the same distance, the preference is given to the left position. Finally, if there is a preposition between the trigger and the NE, then the trigger is not considered. For instance, in "the king of Spain", the trigger "king" is a *person* but the NE "Spain" is a *location*. This last heuristic is motivated by the fact that prepositions tend to be used to syntactically relate nouns and NEs of different semantic classes.

class ranking: if the NE is ambiguous and cannot be disambiguated by contextual checking (previous step), then we select a single class by taking into account our ranking of classes: $person > location > organization$. That is, if the NE appears in the gazetteers of persons and locations, we select the *person* reading. If it is in gazetteers of locations and organizations, the preference is given to the *location* class. This ranking was not set *ad hoc*. It was defined by taking into account the distribution of classes within the gazetteers extracted from the Wikipedia.

internal checking: if the NE is unknown and cannot be assigned a class by contextual checking, then we check some of its constituent expressions. In particular, we check whether the first expression of a NE matches the first expression of a NE in a gazetteer or a common noun in a trigger list. For instance, since the first expression of the NE "University of Alberta" is a trigger word ("university") for organizations, the target NE is classified as

an organization. In case of several options, we give preference to gazetteers over trigger words and follow the class ranking defined above.

default rule: if no rule is applied, the NE is classified as "miscellaneous".

Note that these rules are language and domain independent. We do not make use of specific cues such as organizational designators (e.g., Corp.) or personal suffixes (e.g., Jr.).

5 Experiments

The resource-based method described in this paper was compared with a supervised learning system, namely the NEC module of FreeLing [7]. FreeLing is an open source suite of modules for natural language processing: lemmatization, PoS tagging, named entity recognition/identification (NER), named entity classification (NEC), chunking, etc. The NEC module was ranked among the top performing systems in the CoNLL-2002 competition. It was based on a boosting algorithm (AdaBoost) which consists in combining many base classifiers. It may make use of external resources such as gazetteers and trigger words to define specific features. In the experiments reported in this paper, both our resource-based NEC method and the supervised FreeLing system take as input the basic NER module of FreeLing, an heuristic rule based strategy, which takes into account capitalization patterns, functional words and dictionary lookup. This module achieves 90% precision [8].

The overall organization of our experiments is the following: first, we generate the different lists of gazetteers and triggers that will be used in the experiments. Then, we train the NEC module of FreeLing for Portuguese language. Next, five different test corpora, belonging to different domains and genres, are annotated. Finally, the two systems are applied to the five test corpora, and the results obtained are compared.

5.1 Gazetteers and Triggers

We follow the two strategies defined above in section 4, with the aim of generating two versions of gazetteers. First, three lists with 115,650 named entities were built by checking the categories of the Wikipedia articles (first strategy). Second, the attribute-value structure of the infoboxes was used to select three lists with 37,445 NEs (second strategy). In addition, a third version was also considered for our experiments, namely the gazetteers freely available in the Spanish NEC module of FreeLing.

To compare the impact of gazetteers in terms of size in a NEC task, we will make use of three sets of gazetteers (see Table 2): *es* stands for the gazetteers taken from the Spanish version of FreeLing, *es+infobox* corresponds to the union of *es* with the gazetteers extracted from infoboxes, and finally *es+infobox+cat* consists of the previous set and the NEs extracted using the article's categories. Table 2 shows the number of persons (PER), locations (LOC), and organizations (ORG) found in each gazetteer version.

	es	es+infobox	es+infobox+cat
PER	2,598	17,600	64,735
LOC	7,312	23,732	58,305
ORG	2,263	4,586	13,599
TOTAL	12,173	45,918	136,639

Table 2. Size of the three gazetteers used in the experiments

Concerning trigger words, only 419 nouns (57 types of organizations, 82 types of locations, and 320 types of persons) will be used in the experiments.

5.2 Training the NEC module of FreeLing

We trained both the PoS tagged and NEC modules of FreeLing on an manually annotated European Portuguese corpus. In particular, we selected 87,000 tokens from Bosque 8.0[4], containing about 5,000 proper names which we have manually classified according to the classification criteria defined in Section 3, that is, only homonymy was considered. The training corpus consists of news of a Portuguese newspaper (*Público*). The NEC module for Portuguese will be freely available in the next version of FreeLing.

5.3 Test Corpora

In order to perform experiments in different domains and textual genres, 5 different test corpora were elaborated:

bosque: 50,000 test tokens from Bosque 8.0 (part of CETEMPúblico), which is constituted by news of *Público* (journalistic genre, open domain)

wiki: 30,000 tokens from Portuguese Wikipedia (first paragraph per article) (encyclopedic genre, open domain)

europarl: 30,000 tokens from the Portuguese version of the parallel corpus Europarl[5] (formal genre, political domain)

br: 24,000 tokens from the Brazilian Portuguese part of European Corpus Initiative Multilingual Corpus I (ECI/MCI)[6] (technical genre, economical domain)

harem: 70,000 tokens from HAREM competition [22] (open genre, open domain)

The proper names contained in *bosque*, *wiki*, *europarl*, and *br* were manually annotated by us following the simple criterion of homonymy disambiguation. No metonymy was considered. By contrast, *harem*, which is the corpus used

[4] `http://www.linguateca.pt/floresta/corpus.html`
[5] `http://www.statmt.org/europarl/`
[6] `http://www.elsnet.org/eci.html`

as reference in the Portuguese NEC competition, was annotated by other linguists according to more complex criteria, since many types of metonymy were taken into account. We had to adapted the set of categories used in HAREM competition to the four main categories of our experiments.

The heterogeneity of these test corpora will allow us to verify whether the two compared systems may be ported to new domains or textual genres without losing their performance.

5.4 Results

	null	**es**	**es+infobox**	**es+infobox+cat**
baseline (*bosque*)	.33	.33	.33	.33
resource (*bosque*)	.33	.54	.69	.74
superv (*bosque*)	.74	.75	.77	**.78**
baseline (*wiki*)	.57	.57	.57	.57
resource (*wiki*)	.32	.48	.75	**.92**
superv (*wiki*)	.79	.80	.83	.88
baseline (*br*)	.35	.35	.35	.35
resource (*br*)	.61	.73	.74	**.75**
superv (*br*)	.50	.53	.62	.63
baseline (*europarl*)	.45	.45	.45	.45
resource (*europarl*)	.46	.48	.48	.74
superv (*europarl*)	.77	**.78**	.76	.76
baseline (*harem*)	.41	.41	.41	.41
resource (*harem*)	.26	.39	.53	.58
superv (*harem*)	.56	.56	.59	**.60**
baseline (**average**)	.42	.42	.42	.42
resource (**average**)	.40	.52	.64	**.75**
superv (**average**)	.67	.68	.71	.73

Table 3. F-score$_{\beta=1}$ values of the two compared systems (and a baseline), provided with four different gazetteers. Experiments performed on five test corpora

The two NEC systems were applied on the five test corpora provided with the same sets of gazetteers. In terms of computational efficiency, the resource-based system turned out to be about 40% speeder than the supervised one.

The performance (f-score$_{\beta=1}$) of the two NEC systems (and a baseline) are presented in Table 3. The resource-based system is noted *resource* and the supervised is *superv*. As a baseline, we include the results obtained by using the strategy based on the most frequent category. In each column, we show the results obtained with different sets of gazetteers. In column *null* no gazetteer is used. Column *es* shows the results with the gazetteers taken from the Spanish NEC. Column *es+infobox* shows the results by combining *es* with the infobox

extraction technique, and finally in column *es+infobox+cat*, the gazetteers also include the NEs learnt with the article's categories. Precision, recall, and f-score$_{\beta=1}$ were computed by means of the evaluation script (*conlleval*) used in CoNLL competition, considering only for evaluation those NEs correctly identified by the NER.

We can observe in Table 3 that no system is clearly better than other. On the one hand, the supervised strategy performs better on three test corpora: *bosque*, *europarl*, and *harem*. The good results obtained from *bosque* were expected, since this test is constituted by the same type of documents as in the training corpus. On the one hand, the resourceised system performs better on *wiki* and *br*. The high score obtained from *wiki* (92%) is not a surprise because most gazetteers were extracted from that corpus. In average, the resourceised technique achieves a slightly better f-score with the largest set of gazetteers: 75% against 73%. Let us note the low scores of *harem* are due to the fact that this test corpus was annotated according to different criteria as those used to elaborate both the training corpus of *superv* and the disambiguation rules of *resource*.

The main difference between the two approaches concerns the degree of dependence on knowledge-rich gazetteers. While the performance of *superv* remains quite stable regardless the size of gazetteers, *resource* requires very rich gazetteers to reach acceptable performance. Figure 1 shows how the performance of the two systems improves as the size of gazetteers increases, but the improvement curve is clearly more marked in the case of *resource*. It means that, as it was expected, the resource-based strategy is much more dependent on external gazetteers. By contrast, the supervised one relies on many different features, being only those defined from external resources a small part of the decision model. This is in accordance with the experiments reported in [8], where the same supervised NEC system merely improved 2 or 3 points its performance when it was trained with external resources such as gazetteers.

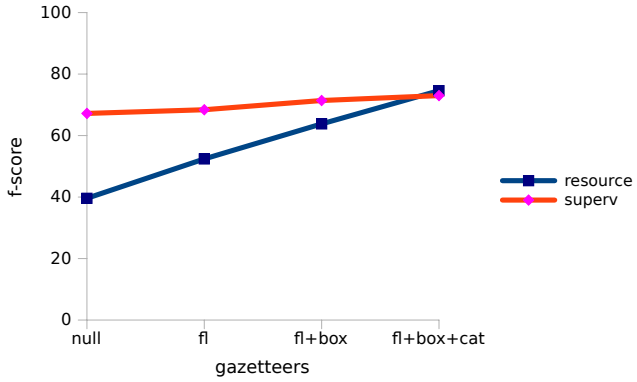

Fig. 1. Improvement curve of the two systems in function of the size of gazetteers (average f-score)

We observe a similar tendency in the case of trigger words (see Table 4). The resource-based strategy is clearly more dependent on the contextual information provided by trigger words than the supervised one. In brackets, we add the difference (in percentage points) between these results and those obtained by the same system but provided with trigger words. In average, *resource* decreases 5 percentage points while *superv* only 2.

	bosque	**wiki**	**br**	**europarl**	**harem**
resource (*no triggers*)	.69 (-5)	.92 (=)	.67 (-8)	.65 (-9)	.56 (-2)
superv (*no triggers*)	.76 (-2)	.88 (=)	.59 (-4)	.74 (-2)	.59 (-1)

Table 4. F-score$_{\beta=1}$ of the two systems without triggers (and with gazetteers *es+infobox+cat*)

.

Finally, in order to analyze in more detail the behaviour of the two systems, Table 5 breaks down the results into four categories: PER, LOC, ORG, and MISC. In particular, this table shows the results obtained with the largest gazetteers (*es+infobox+cat*), and the *bosque* text corpus. We can observe that the best performance is reached with the PER category, while the MISC category turns out to be the most hard to predict. The low values of MISC could be caused by either the difficulty of identifying a so general and heterogeneous category, or by the fact that we have not created any specific gazetteers with *miscellaneous* named entities.

resource (*bosque*)	**precision**	**recall**	**f-score**$_{\beta=1}$
PER	.86	.89	.88
LOC	.76	.54	.63
ORG	.79	.68	.73
MISC	.33	.51	.40
overall	74.48%	74.19%	74.33%
superv (*bosque*)	**precision**	**recall**	**f-score**$_{\beta=1}$
PER	.88	.94	.91
LOC	.85	.57	.68
ORG	.81	.70	.75
MISC	.35	.52	.42
overall	77.77%	77.54%	77.65%

Table 5. Results of the two systems for the four semantic categories. They were obtained from the *bosque* text corpus, with the largest gazetteers (*es+infobox+cat*).

5.5 Comparing with Related Work

It is difficult to establish a fair comparison between our systems and those described in other works, due to the specificities of each evaluation setup. However we should note that the performance of *superv* on *bosque* (about 78% f-score) is similar to that achieved by the same NEC system trained for Dutch [9]. It should also be noted that our training data (87,000) is three times smaller than that available in CoNLL-2002 for Dutch and Spanish. This could be the reason of the slightly lower score obtained by our supervised method in comparison to the Spanish NEC system. On the other hand, a full comparison of our results with those obtained by the participants in the HAREM competition is not possible because the semantic classification criteria do not coincide. Besides, some errors could be produced when converting the original set of categories of HAREM to the basic set used in our experiments. However, we observed that our resource-based method achieves the same f-score (58%) as the best NEC system in that competition.

6 Conclusions

We have presented a named entity recognition system that avoids the need for supervision by making use of some language independent rules on automatically extracted external resources, namely gazetteers and trigger words. When comparing with a supervised system, we made the two following observations: First, the supervised strategy performs better when both the test and training corpora are similar (same genre and same domain). Second, our resource-based strategy is not worse than the supervised system when they are applied on a great variety of texts, especially if the domains and genres of these texts are not found in the training corpus. So, we conclude that if we need to work on domain-specific texts, it is worth manually annotating a corpus to tune a supervised system. Nevertheless, if we require a more generic NEC with acceptable performance on any type of text, our resource-based system could be a reasonable solution.

References

1. E. Alfonseca and S. Manandhar. An unsupervised method for general named entity recognition and automated concept discovery. In *International Conference on General WordNet*, 2002.
2. Cristian Aranha. O cortex e a sua participação no HAREM. In Diana Santos and Nuno Cardoso, editors, *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM*, pages 113–122. 2007.
3. M. Asahara and Y. Matsumoto. Japanese named entity extraction with redundant morphological analysis. In *Human Language Technology Conference - North American chapter of the Association for Computational Linguistics*, 2003.
4. Alan Beretta, Robert Fiorentino, and David Poeppel. The effets of homonymy and polysemy on lexical access: an MEG study. *Cognitive Brain Research*, 24:57–65, 2005.

5. Eckhard Bick. Functional aspects on portuguese ner. In *7th Workshop on Computational Processing of Written and Spoken Language (PROPOR'2006)*, Itatiaia, Brazil, 2006.
6. S. Borgo, N. Guarino, and C. Masolo. Stratified ontologies: the case of physical objects. In *Workshop on Ontological Engineering (ECAI-96)*, 1996.
7. X. Carreras, I. Chao, L. Padró, and M. Padró. An Open-Source Suite of Language Analyzers. In *4th International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, 2004.
8. X. Carreras, L. Marquez, and L. Padró. Wide coverage spanish named entity extraction. In *IBERAMIA 2002 Proceedings of the 8th IberoAmerican Conference on AI: Advances in Artificial Intelligence*, 2002.
9. X. Carreras, L. Marquez, L. Padró, and M. Padró. Named entity extraction using adaboost. In *COLING-02 proceedings of the 6th Conference on Natural Language Learning*, 2002.
10. Cristina Mota e Max Silberztein. Em busca da máxima precição sem almanaques. o stencil/nooj no HAREM. In Diana Santos and Nuno Cardoso, editors, *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM*, pages 191–208. 2007.
11. O. Ferrández, Z. Kozareva, A. Toral, R. Mu noz, and A. Montoyo. Tackling HAREM's portuguese named entity recognition task with spanish resources. In Diana Santos and Nuno Cardoso, editors, *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM*, pages 137–144. 2007.
12. E. Ferreira, J. Balsa, and A. Branco. Combining rule-based and statistical methods for named entity recognition in portuguese. In *V Workshop em Tecnologia da Informação e da Linguagem Humana*, pages 1615–1624, 2007.
13. Ralph Grishman and B. Sundheim. Message understanding conference -6: A brief history. In *International Conference on Computational Linguistics*, 1996.
14. Nicola Guarino. The role of identity conditions in ontology design. In *IJCAI-99 Workshop on Ontologies and Problem-Solving Methods*, Stockholm, Sweden, 1998.
15. George Kleiber. *Problèmes de Sémantique.* Presses Universitaires Septentrion, Lille, 1999.
16. E. Klepousniotou. The processing of lexical ambiguity: Homonymy and polysemy in the mental lexicon. *Brain and Language*, 81:205–223, 2002.
17. A. McCallum and W. Li. Early results for named entity recognition with conditional random fields, features induction and web-enhanced lexicons. In *Conference on Computational Natural Language Learning (CoNLL-2003)*, pages 117–126, 2003.
18. D. Nadeau, P. Turney, and S. Matwin. Unsupervised named entity recognition: Generating gazetteers and resolving ambiguity. In *Canadian Conference on Artificial Intelligence*, 2006.
19. J. Nothman, T. Murphy, and J.R. Curran. Analysing wikipedia and gold-standard corpora for ner training. In *12th Conference of the European Chapter of the ACL*, pages 612–620, Athens, Greece, 2009.
20. Manfred Pinkal. Vagueness, ambiguity, and underspecification. In *Conference on Semantics and Linguistic Theory*, pages 181–201, 1996.
21. James Pustejovsky. *The Generative Lexicon.* MIT Press, Cambridge, 1995.
22. Diana Santos, Nuno Seco, Nuno Cardoso, and Rui Vilel. HAREM: An advanced NER evaluation contest for portuguese. In *5th International Conference on Language Resources and Evaluation - LREC'2006*, pages 1986–1981, Genova, Italy, 2006.

23. Luís Sarmento. O SIEMÊS e a sua participação no HAREM e no mini-HAREM. In Diana Santos and Nuno Cardoso, editors, *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM*, pages 173–189. 2007.
24. Kim Sang Tjong and F. Erik. Introduction ot the CoNLL-2002 shared task: Language independent named entity recognition. In *Conference on Natural Language Learning*, 2002.