Capítulo 11

Tackling HAREM's Portuguese Named Entity Recognition task with Spanish resources

Óscar Ferrández, Zornitsa Kozareva, Antonio Toral, Rafael Muñoz e Andrés Montoyo

Diana Santos e Nuno Cardoso, editores, *Reconhecimento de entidades mencionadas em português: Documentação e actas do* HAREM*, a primeira avaliação conjunta na área*, Capítulo 11, p. 137–144, 2007.

This chapter presents our participation in the HAREM evaluation contest. This is a challenge regarding the identification and classification of named entities in Portuguese. Our NER system, initially designed for Spanish, combines several classifiers in order to resolve the classification of the entities. Besides, a rule-based module has been used to deal with entity types easily recognized by applying knowledge resources such as regular expressions (e.g. `TEMPO:DATA`).

The rest of this chapter is organized as follows. The next section introduces our system and the modules it is made of. The carried out experiments are explained and discussed in Section 11.2. Finally, Section 11.3 outlines our conclusions.

11.1 System Description

For our participation in HAREM (Santos et al., 2006), we have used the architecture of our system NERUA (Ferrández et al., 2005; Kozareva et al., 2007). This is a NER system that was developed combining three classifiers by means of a voting strategy. This system carries out the recognition of entities in two phases: detection[1] of entities and classification of the detected entities. The three classifiers integrated in NERUA use the following algorithms: Hidden Markov Models (HMM) (Schröer, 2002), Maximum Entrophy (ME) (Suárez e Palomar, 2002) and Memory Based Learning (TiMBL) (Daelemans et al., 2003). The outputs of the classifiers are combined using a weighted voting strategy which consists of assigning different weights to the models corresponding to the correct class they determine. An overview of our system is depicted in Figure 11.1.

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

The second stage has as starting point the text with the identified named entities. Therefore, only entities that have been previously detected are going to be classified and features for the classification of these entities will be extracted. The performance of the second stage is obviously influenced by the results of the first one. 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. The final outcome is the result of the voting scheme. This second stage yields all the identified NE together with the class each entity belongs to.

Our voting approach regarding both the identification and the classification phases has been already evaluated in Ferrández et al. (2005) and Kozareva et al. (2007). TiMBL is the classifier that obtains the best results for identification, while ME is the one reaching the

[1] **Editors' note.** As in the previous chapter, the authors use *detection* to mean what we dubbed *identification* in HAREM.

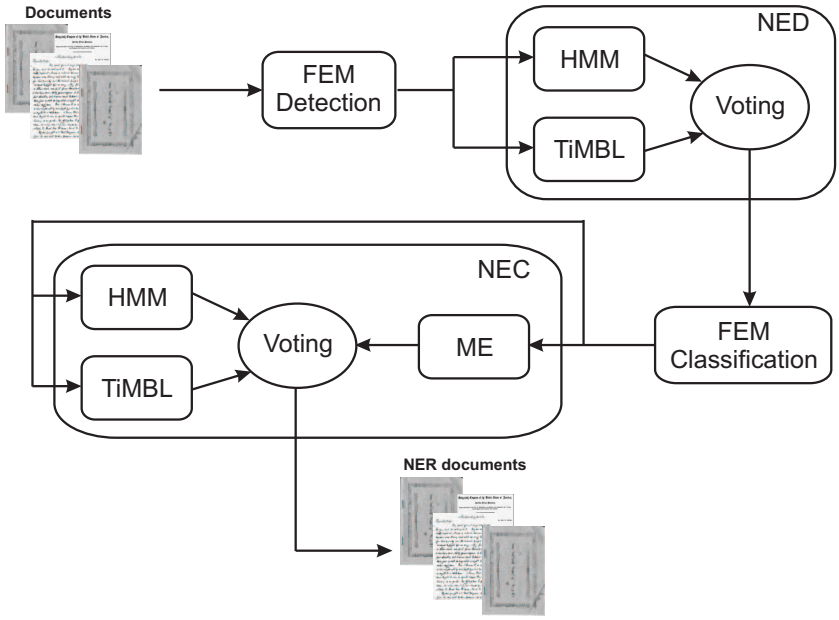

Figure 11.1: The NERUA architecture.

best score for the classification. The voting strategy meaningfully increases the final score above the results provided separately by the algorithms.

Due to the small size of tagged corpora available for Portuguese and the facts that our NER system was initially designed for Spanish and Spanish and Portuguese are close-related languages, we decided to merge the Spanish and Portuguese training corpora in order to train our system. The Spanish training corpus we used was provided for the CoNLL-2002 shared task (Sang, 2002). As in CoNLL-2002 only four kind of entities were considered (`PERSON`, `ORGANIZATION`, `LOCATION` and `MISCELLANEOUS`) we have focused in the following HAREM correspondent types: `PESSOA`, `ORGANIZACAO` and `LOCAL`.

By studying the entity taxonomy of HAREM (Santos et al., 2006), we saw that for some of the NE types, a knowledge-based approach could perform better. Entities such as `TEMPO:DATA` or `VALOR:QUANTIDADE`, have regular and a priori known structure, therefore they can be tackled more efficiently by using regular expressions and dictionaries.

Therefore, apart from the machine-learning system, we used a knowledge-based one which classifies the following entity types: `LOCAL:VIRTUAL`, `TEMPO:DATA`, `TEMPO:CICLICO`, `TEMPO:HORA`, `VALOR:MOEDA` and `VALOR:QUANTIDADE`. The system we used is called DRAMNERI (Toral, 2005). This system is a NER application belonging to the knowledge paradigm and adaptable to different domains and languages. In this research, this system has been adapted to recognize the aforementioned types of entities by hand-coding the correspondent dictionaries and rules.

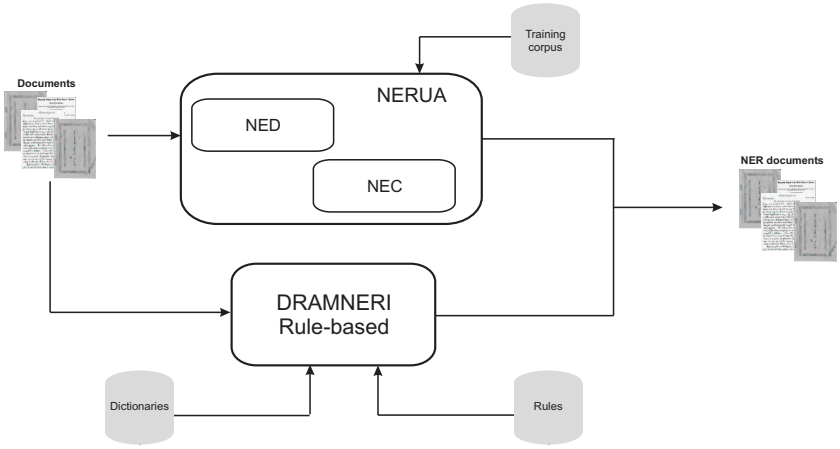

Figure 11.2: System description.

For this purposes, DRAMNERI uses 32 rules (4 for `LOCAL:VIRTUAL`, 21 for `TEMPO:DATA`, 1 for `TEMPO:CICLICO`, 2 for `TEMPO:HORA`, 3 for `VALOR:MOEDA` and 1 for `VALOR:QUANTIDADE`). The applied dictionaries contain 80 tokens. These resources were adapted from the Spanish resources. The adaptation consisted only of translating the language dependent strings in the dictionaries and in the rules (e.g. January (*Enero* to *Janeiro*)). In other words, the rules' structure was not modified.

Figure 11.2 depicts our system using both the machine learning and rule-based NER sub-systems. Both NER sub-systems are applied to the input-text in a parallel way. Afterwards, a postprocessing module receives both tagged texts and composes a final tagged text. If a snippet is tagged as an entity by both modules then the rule-based one is given precedence, i.e., the entity tagged by this latter NER system would be the one preserved[2].

11.1.1 Feature sets

To improve the performance of the classifiers, a large number of features were extracted from the training corpus to get a pool of potentially useful features (this procedure is shown in detail in Ferrández et al. (2006)). Many of these features are acquired from the best performing NER systems such as Carreras et al. (2002) and Florian et al. (2003). We have divided our features into several groups: orthographic (about the orthography of the word), contextual (about the context of the word), morphological (about morphological characteristics), statistic (about statistical characteristics) and handcrafted-list (test whether or not the word is contained in some handcrafted list of general entities obtained from several web pages). Below, we describe the features in detail:

[2] This case rarely happens, since the systems were designed to classify different kind of entities.

- **Orthographic**
  - **a**: anchor word (e.g. the word to be classified)
  - **cap**: capitalization of the word and context
  - **allcap**: whole word and context are in upper case
  - **lower**: whole word and context are in lower case
  - **internal**: word and context have internal upper case letters
  - **digits**: word and context are only made up of digits
  - **contdig**: word and context contain digits
  - **ispunct**: word and context are punctuation marks
  - **contpunct**: word and context contain punctuation marks
  - **hyphen**: word and context are hyphenated
  - **initial**: word and context are initials (e.g. B.O.E. or D.O.G.V.)
  - **url**: word and context represent an URL
  - **prefix**: the first three and four characters of the word and context
  - **suffix**: the last three and four characters of the word and context
  - **middle**: half substring of the word and context
  - **firstword**: first word of the whole entity
  - **secondword**: second word of the whole entity
  - **clx**: words within the entity are upper-cased (c), lower-cased (l) or made up of other symbols (x), e.g. *Charles de Gaulle*: clc
- **Contextual**
  - **cntxt**: word context at position ± 1, ± 2, ± 3
  - **verbword**: the nearest verb that comes with the entity
- **Morphological**
  - **postag**: PoS tag of the word and context
  - **lemma**: lemma of the word and context
  - **stem**: stem of the word and context
- **Metrical**
  - **length**: number of characters of the word and context
  - **firstpos**: word is the first word of the sentence

- **Handcrafted list**
  - **stopword**: word and context are stop-words
  - **dict**: word and context are in handcrafted dictionaries of entities (locations, persons and organizations)
  - **trigg**: word and context are in handcrafted dictionaries of trigger words
  - **connec**: context is contained in a dictionary of connectives
  - **WNword**: the WordNet semantic prime of the word from the Spanish WordNet

Since in HAREM we did not have enough training resources for the target language (Portuguese), we have considered only sets containing features that do not depend on a language-specific tool (called IDL sets) (Ferrández et al., 2006). In order to select the most meaningful features, we have followed a bottom-up strategy. This strategy iteratively adds one feature at a time and checks the effect of this feature in the results according to the information gain of this feature. The feature sets used for HAREM were:

- IDL sets for the detection phase
  - IDL1d: a, cntxt, cap, allcap[3], firstpos, url[3], ispunct[3], contpunct[3], digits[3], contdig[3], internal[3], ishyphen[3], lower[3].
  - IDL2d: IDL1 + prefix[3], suffix[3], middle[3].
- IDL sets for the classification phase
  - IDL1c: a, cntxt, firstpos, firstword, secondword, clx, url[3], ispunct[3], cont-punct[3], digits[3], contdig[3], internal[3], ishyphen[3], lower[3].

11.2 Experiments and discussion

This section presents the experiments carried out for our participation in HAREM. We show the obtained results and briefly discuss them. The aim of our study is to evaluate the recognition of entities with resources for a close-related language.

We have carried out three runs: one for the identification (*r_detection*) and the remaining two for the semantic classification. Regarding the two classification runs, one (*r_clas_total*) deals with all the entity types that we have considered while the other one (*r_clas_partial*) treats the ones that we thought the system could obtain better results (all categories but OBRA and ABSTRACCAO).

Table 11.2 shows the results obtained for the identification phase in HAREM. Table 11.2 presents the results for the semantic classification task according to CSC (combined) measure (Santos et al., 2006).

[3] only the word (not the context)

Category	Run	Total scenario			Selective scenario		
		Precision	Recall	F measure	Precision	Recall	F measure
all	r_detection	56.93%	64.39%	0.6043	-	-	-
	r_clas_partial	59.43%	64.39%	0.6181	52.25%	65.43%	0.5810
	r_clas_total	57.19%	63.51%	0.6019	-	-	-

Table 11.1: Results of the identification task, for the total and selective scenarios.

Category	Run	Absolute scenario			Relative scenario		
		Precision	Recall	F measure	Precision	Recall	F measure
PESSOA	r_clas_partial	26.93%	16.44%	0.2042	84.37%	49.86%	0.6268
	r_clas_total	19.59%	26.67%	0.2259	79.15%	79.62%	0.7938
ORGANIZACAO	r_clas_partial	27.35%	21.44%	0.2404	76.63%	46.36%	0.5777
	r_clas_total	25.57%	27.61%	0.2655	65.56%	68.44%	0.6697
LOCAL	r_clas_partial	40.13%	19.27%	0.2603	89.72%	52.37%	0.6614
	r_clas_total	32.90%	29.78%	0.3126	82.38%	83.50%	0.8294
TEMPO	r_clas_partial	75.26%	65.36%	0.6996	91.58%	91.88%	0.9173
	r_clas_total	53.58%	66.57%	0.5937	91.22%	91.80%	0.9151
VALOR	r_clas_partial	35.23%	71.12%	0.4712	77.42%	79.22%	0.7831
	r_clas_total	34.72%	72.26%	0.4690	77.61%	79.39%	0.7849
ABSTRACCAO	r_clas_total	15.14%	6.72%	0.0931	58.52%	59.66%	0.5908
OBRA	r_clas_total	6.62%	5.36%	0.0592	60.74%	52.98%	0.5660
VARIADO	r_clas_partial	1.28%	21.96%	0.0241	85.64%	85.64%	0.8564

Table 11.2: Results of the semantic classification task according to the CSC (combined) measure, for the selective scenario (runs *r_clas_partial*) and for the total scenario (*r_clas_total*).

Regarding identification (see Table 11.2), even if we have not made an extensive use of Portuguese specific resources, we have reached the 5th best score in F measure. Considering the small effort realised in order to adapt our system to Portuguese, the overall results are promising. It should be noted as well that the result for the selective scenario is worst (see r_class_partial) than that for the total scenario. This is due to the fact that for the selective scenario the categories ABSTRACCAO and OBRA are not considered but they might be detected by our system although afterwards they will not be classified (this is why the results for the selective scenario in the semantic classification (see Table 11.2) are better than for the total scenario).

As to the entity classification (see Table 11.2), our system obtains quite high scores for TEMPO (F measure of 0.9173) and LOCAL (F measure of 0.8294). This is due to the fact that, in the first case, temporal expressions can be appropriately tackled with regular expressions and, in the second case, local entities do not depend that much on the specific language.

11.3 Conclusions

In this paper we have presented our participation in HAREM. In order to recognize named entities in Portuguese, we decided to apply our previously developed NER system for Spanish. We have merged our already available Spanish corpus with the Portuguese one because of the lack of sufficient training data. The feature sets developed for Spanish were directly ported to detect and classify Portuguese NE. This was possible due to the proximity and the common characteristics of the two languages. Apart from this, we treated some entities (`VALOR`, `TEMPO`, `LOCAL:VIRTUAL`) with a knowledge-based approach.

NERUA came on fifth position in the NE identification task in the first HAREM. 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 a subset of entities. In this contest, we showed that our NER system, initially designed and developed for Spanish, was adapted with little effort to Portuguese and achieved promising results.

Acknowledgements

This research has been partially funded by the Spanish Government under project CICyT number TIC2003-07158-C04-01.