Comparative Analysis of Portuguese Named Entities Recognition Tools

Daniela O. F. Amaral, Evandro B. Fonseca, Lucelene Lopes, Renata Vieira
PUCRS University – Porto Alegre – Brazil
daniela.amaral@acad.pucrs.br, evandro.fonseca@acad.pucrs.br,
lucelene.lopes@pucrs.br, renata.vieira@pucrs.br

Abstract

This paper describes an experiment to compare four tools to recognize named entities in Portuguese texts. The experiment was made over the HAREM corpora, a golden standard for named entities recognition in Portuguese. The tools experimented are based on natural language processing techniques and also machine learning. Specifically, one of the tools is based on Conditional random fields, an unsupervised machine learning model that has being used to named entities recognition in several languages, while the other tools follow more traditional natural language approaches. The comparison results indicate advantages for different tools according to the different classes of named entities. Despite of such balance among tools, we conclude pointing out foreseeable advantages to the machine learning based tool.

Keywords: Named entities recognition, Information extraction tools, Conditional random fields

1. Introduction

Named entity recognition (NER) is an important task in natural language processing (NLP). The typical NER task associates a semantic category to each actor of the discourse in a written text. The most common categories found in NER are: proper names identifying people, organizations and places, but also more abstract concepts as time, currency and events can be identified as well.

The possible approaches to NER can be inspired in traditional NLP approaches starting with part-of-speech (POS) tagging and subsequent recognition of patterns. Consequently, many NER efforts are strongly dependent on the text language, or at least dependent on the available tools for the text language. In this context, well resourced languages like English, German and French tend to have at their disposal high quality tools for NLP in general, and of course, NER specifically.

Less resourced language, like Portuguese, which is the subject of this paper, suffer with considerably less efficient tools. Hence, NLP researchers of Portuguese language stimulate NER initiatives creating challenges on the subject. Probably the most important contribution for Portuguese texts NER was the HAREM contest (Santos et al., 2006; Santos et al., 2008). This initiative has established two corpora (first and second HAREM) that have been used as gold standard in most, if not all, recent works for NER in Portuguese. However, it is important to mention that the number of NER studies for Portuguese is relatively small if compared to similar works for English, or even other well-researched languages as French and German (Suakkaphong et al., 2011; Finkel et al., 2005).

Among other information retrieval tasks, NER stands as a vital one, since the identification of actors, and their categories is the basis for many NLP efforts (Jiang, 2012). Therefore, this paper presents a comparative study with four NER systems for Portuguese texts. Specifically, the chosen systems to compare are:

- FreeLing (Padro et al., 2010);
- LanguageTasks (Lan, 2013);
- Palavras (Bick, 2000);
- NERP-CRF (Amaral, 2012).

The performance of each system is observed using basic information retrieval measures (precision, recall and f-measure). The goal of such comparison is to estimate the competitiveness of each system in terms of their effectiveness and efficiency.

This paper is organized as follows: The next section describes basic information about both HAREM corpora initiatives, which is used to all this paper experiments. Section 3, briefly describe the four tools employed in this paper experiments with special emphasis on the NERP-CRF tool that is considerably different from the other tools. Section 4 describes the experiment results. Finally, the conclusion summarizes this paper contributions and draws possible future work.

2. HAREM Corpora

The HAREM corpora are an initiative of the Linguateca consortium. The first HAREM corpus (Santos et al., 2006) was published in 2006 with a human made golden standard reference of named entities. The second HAREM corpus (Santos et al., 2008) was made available in 2008 with basically the same purpose. However, being a more recent contribution, the golden standard for the second HAREM consists in a slightly better test case.

HAREM corpora classify the named entities in ten categories (Tab. 1). Nevertheless, for the purposes of our

comparison only the first three categories ("Person", "Place" and "Organization") were taken into account, since these are the only ones present in all chosen NER systems. Considering only these three categories the number of named entities drops from 7,255 to 4,245.

Corpora	First HAREM 129 texts		Second HAREM 129 texts	
Categories	466,355 words		89,241 words	
Person	1,040	20%	2,035	28%
Place	1,258	25%	1,250	17%
Organization	946	18%	960	13%
Value	484	9%	352	5%
Abstraction	461	9%	278	4%
Time	440	9%	1,189	16%
Work	210	4%	437	6%
Event	128	2%	302	4%
Thing	79	2%	304	4%
Other	86	2%	148	2%
Total	5,132	100%	7,255	100%

Table 1: Number of named entities in each category for the HAREM reference lists.

3. NER Systems for Portuguese

The availability of NER systems for Portuguese leaves very few choices in both academic and commercial communities. In this scarce environment, we pick four systems briefly described as follows.

3.1. NERP-CRF

The first compared system is an academic tool developed in Python language and it is based on a probabilistic mathematical model using Conditional Random Field (CRF) (Lafferty et al., 2001).

The process to recognize named entities in NERP-CRF is made in two phases: test and training. For the comparison carried out, the training set was the first HAREM corpus, while the second HAREM corpus was used, as in the other systems, as test set. The input to the NERP-CRF system was the corpus annotated with POS (Part-Of-Speech) tags, plus the three HAREM defined categories ("Person", "Place" and "Organization"). The named entities of the other HAREM categories were ignored.

3.2. LanguageTasks

LanguageTasks, also known as LTasks, is an open source package that performs the named entities recognition using a simpler approach than NERP-CRF, since it is based on the search for morpho syntatic patterns. The input format for LanguageTasks is a raw corpus, but the possible categories are the same nine possibilities of original HAREM categories, except "Other" category. Therefore, named entities classified in categories other than "Person", "Place" and "Organization" were ignored.

3.3. FreeLing

The third compared system is the NER tool of the FreeLing package, a set of NLP tools capable of a fair number of natural language tasks. It is noticeable that the FreeLing package is capable of handling texts in Portuguese, but also Spanish and English. Another interested aspect of the FreeLing system is the existence of two distinct forms to recognize named entities, one more simple based on morphosyntactic patterns, and another one much more sophisticated based on automatic learning algorithms. In the comparisons carried out in this paper the second module was employed.

3.4. PALAVRAS

The fourth system employed in this paper is the parser PALAVRAS, a traditional NLP software tool for Portuguese texts capable of POS-tagging and even semantic annotation. The input format of PALAVRAS is pure text without any kind of annotation. In contrast, the PALAVRAS output is a very rich annotation where even a syntactic tree structures with all sort of grammatical annotations is available. However, for this paper, only the semantic annotation concerning the three chosen HAREM tags ("Person", "Place" and "Organization") assigned to named entities is considered.

3.5. Example of NER

In order to illustrate the result of the chosen NER tools, Tab. 2 presents fifteen named entities randomly chosen from the Second HAREM corpus and their assigned category by each tool, as well as its reference HAREM annotation.

It is important to recall that any identified category other than "Person", "Place" and "Organization" is ignored. Hence all missing category information in Table 2 indicate either that the tool did not recognized the term as a named entity or the tool did not recognized it as a "Person", "Place" or "Organization".

For instance, the term "Detroit" was correctly recognized as a named entity of "Place" category by all tools. However, the term "Chicago" was correctly recognized as "Place" by all tools, but LanguageTasks, which considered it as a "Person".

The term "Hari Kunzru", on the contrary, was correctly recognized as a "Person" only by NERP-CRF Another interesting example is the term "Model 500" that is a "Person" according to the Second HAREM reference list and was correctly recognized as such by all tools, but FreeLing.

named entity	**HAREM reference**	FreeLing categorization	PALAVRAS categorization	Language Tasks categorization	NERP-CRF categorization
Chicago	**place**	**place**	**place**	person	**place**
Detroit	**place**	**place**	**place**	**place**	**place**
Durban	**place**	**place**	**place**	person	**place**
Ford	**organization**	**organization**	**organization**		**organization**
General Motors	**organization**	**organization**	**organization**	person	person
Hari Kunzru	**person**			organization	**person**
Jeff Mills	**person**	**person**	**person**	**person**	**person**
Kodwo Eshun	**person**	**person**	**person**	**person**	**person**
Matthew Collin	**person**	**person**	**person**	**person**	**person**
Ministério da Justiça	**organization**	**organization**	**organization**	**organization**	**organization**
Model 500	**person**		**person**	**person**	**person**
Motown	**organization**	**organization**	place	place	place
New Orleans	**place**	**place**	person	person	person
Reagan	**person**	place	**person**	**person**	place
Tóquio	**place**	**place**	**place**	**place**	**place**

Table 2: Some randomly chosen named entities in the Second HAREM and their categorization by each tool.

Systems	P	R	F	$\mid ES \mid$	$\mid ES \cap ER \mid$
NERP-CRF	53%	53%	53%	4,239	2,257
LanguageTasks	50%	62%	55%	5,230	2,615
FreeLing	47%	64%	54%	5,798	2,697
PALAVRAS	52%	61%	57%	4,966	2,603

Table 3: Measures for each system considering the 4,245 named entities in "Person", "Place" and "Organization" of the HAREM reference list (ER).

4. Numerical Results

The comparison was made submitting the second HAREM corpus to each NER system. Next, the golden standard was used as reference to compute traditional information retrieval measures: Precision (P), Recall (R) and F-measure (F). The basis for this computation is the set with the terms in the reference (ER) and the set with the terms recognized by each system (ES). Formally, these measures are defined as follows:

$$P = \frac{\mid ES \cap ER \mid}{\mid ES \mid}$$

$$R = \frac{\mid ES \cap ER \mid}{\mid ER \mid}$$

$$F = \frac{2PR}{P + R}$$

Tab. 3 presents the measures to each system in comparison with the golden standard. It is important to stress that we limit our analysis to "Person", "Place" and "Organization".

The results presented in Tab. 3 show a better Precision index achieved by NERP-CRF, since this system is more restrictive than the others, i.e., it identifies less (4,239) named entities. However, the Recall index values show that all systems are equivalent, since the ones with lower precision, delivers higher recall. This fact is noticeable by the similar values of F-measure. It is also worth to mention that NERP-CRF is quite balanced delivering 53% for all measures.

After analyzing the measures for the three categories altogether, Tab. 4 presents the separated analysis for each category individually.

Observing the results in Tab. 4 we see that each system presents performance variations according to different categories. For instance, the precision for LanguageTasks goes from 63% for "Person" to 31% for

"Person" Category $\mid ER \mid$ = 2,035					
Systems	P	R	F	$\mid ES \mid$	$\mid ES \cap ER \mid$
NERP-CRF	57%	51%	54%	1,803	1,028
LanguageTasks	63%	62%	62%	2,017	1,262
FreeLing	55%	61%	58%	2,279	1,243
PALAVRAS	61%	65%	63%	2,158	1,318
"Place" Category $\mid ER \mid$ = 1,250					
Systems	P	R	F	$\mid ES \mid$	$\mid ES \cap ER \mid$
NERP-CRF	52%	57%	55%	1,382	718
LanguageTasks	61%	57%	59%	1,170	714
FreeLing	58%	66%	61%	1,431	823
PALAVRAS	62%	59%	61%	1,193	741
"Organization" Category $\mid ER \mid$ = 960					
Systems	P	R	F	$\mid ES \mid$	$\mid ES \cap ER \mid$
NERP-CRF	48%	53%	51%	1,054	511
LanguageTasks	31%	67%	43%	2,043	639
FreeLing	30%	66%	41%	2,088	631
PALAVRAS	34%	57%	42%	1,615	544

Table 4: Measures for each system considering separately the named entities in "Person", "Place" and "Organization" of the HAREM reference list (ER).

"Organization". However, the precision of NERP-CRF for "Organization" is clearly higher than the other systems.

The terms in "Person" are by far the more numerous (2,035 of 4,245). Therefore, the good performance for this category of both PALAVRAS and LanguageTasks justifies the achievements of these systems in the overall analysis (Tab.3). Such better performance may also indicate a particular suitability of their methods to this category, since unlike NERP-CRF and FreeLing, LTasks and PALAVRAS are mostly based on the search for morphosyntactic patterns, and not on learning algorithms.

The impressive difference between the precision achieved by FreeLing and NERP-CRF for "Organization" is also interesting. Those two systems are the only ones using learning algorithms. For "Person" and "Place" such difference seems to be prejudicial since LTasks and PALAVRAS had clearly higher precision values (around 62%), while NERP-CRF and FreeLing was around 55%. However, for "Organization" we found FreeLing delivering precision values as low as the other systems (around 32%), while NERP-CRF precision is clearly higher (48%). This behavior may indicate a better adaptability of NERP-CRF method, probably due to the use of training and test sets. In fact, the NERP-CRF training set (First HAREM corpus) is particularly rich in named entities categorized as "Organization" (18%) compared to the 13% found in the test set (Second HAREM corpus), as shown in Tab. 1.

5. Conclusion

The performed comparison has raised some interesting points about NER in Portuguese texts. The systems deliver reasonable results, but there is still much room for improvement. The F-measure of the tested systems stayed around 50%, which is far from the values achieved for more resourced languages.

For instance, a previous NER work (Finkel et al., 2005) describe a NER experience with traditional English corpora delivering F-measures values above 80%. In contrast, a previous work in Portuguese (Bick, 2006) describes the use of PALAVRAS for the first HAREM corpus, claiming F-measure values around 65% for "Person" and "Place", and around 57% for "Organization".

It is also noticeable that the quality of named entity recognition is often limited by the quality of parsing and general language tasks available for each language. It is not surprising to find inferior numerical

results for Portuguese, since in parsing of Portuguese language (Bick, 2000; Silva et al., 2010) is usually less accurate than English counterparts (Thede and Harper, 1999; Toutanova et al., 2003). Concept identification also shows such disadvantage towards Portuguese (Lopes et al., 2010; Lopes, 2012), even considering very efficient techniques (Drumond and Girardi, 2010; Lopes et al., 2012).

The produced resources related to this paper (list of terms and categories extracted by each system and reference lists) are electronic available at:

`http://www.inf.pucrs.br/~linatural`

In terms of future work, the conducted comparison led us to believe that the method used by NERP-CRF, due to the use of training and test sets, presents a better potential for improvement than the others. While the other systems seem to be delivering their full potential, NERP-CRF results can be improved by the use of a better training set. This belief is justified by the high precision achieved by NERP-CRF for "Organization".

6. References

Amaral, D. O. F. (2012). Reconhecimento de entidades nomeadas por meio de conditional random fields para a lingua portuguesa. M.sc. dissertation, PUCRS, Porto Alegre, Brazil.

Bick, E. (2000). *The parsing system PALAVRAS: automatic grammatical analysis of portuguese in constraint grammar framework*. Ph.D. thesis, Arhus University, Arhus, Danemark.

Bick, E. (2006). Functional aspects in portuguese ner. In Vieira, R., Quaresma, P., das Graças Volpe Nunes, M., Mamede, N. J., Oliveira, C., and Dias, M. C., editors, *PROPOR*, volume 3960 of *Lecture Notes in Computer Science*, pages 80–89. Springer.

Drumond, L. and Girardi, R. (2010). Extracting ontology concept hierarchies from text using markov logic. In *Proceedings of the 2010 ACM Symposium on Applied Computing*, SAC '10, pages 1354–1358, New York, USA. ACM.

Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jiang, J. (2012). Information extraction from text. In Aggarwal, C. C. and Zhai, C., editors, *Mining Text Data*, pages 11–41. Springer.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

(2013). Language tasks. `http://ltasks.com/`, October. (Last access October 12, 2013).

Lopes, L., Oliveira, L. H., and Vieira, R. (2010). Portuguese term extraction methods: Comparing linguistic and statistical approaches. In *PROPOR 2010 – International Conference on Computational Processing of Portuguese Language*.

Lopes, L., Fernandes, P., and Vieira, R. (2012). Domain term relevance through tf-dcf. In *Proceedings of the 2012 International Conference on Artificial Intelligence (ICAI 2012)*, pages 1001–1007, Las Vegas, USA. CSREA Press.

Lopes, L. (2012). *Extração automática de conceitos a partir de textos em língua portuguesa*. Ph.D. thesis, PUCRS University - Computer Science Department, Porto Alegre, Brazil.

Padro, L., Collado, M., Reese, S., Lloberes, M., and Castellon, I. (2010). Freeling 2.1: Five years of open-source language processing tools. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *LREC*. European Language Resources Association.

Santos, D., Seco, N., Cardoso, N., and Vilela, R. (2006). HAREM: an Advanced NER Evaluation Contest for Portuguese. In *Proceedings of LREC'2006*, pages 1986–1991, Genoa, Italy, 22-28 May.

Santos, D., Freitas, C., Oliveira, H. G., and Carvalho, P. (2008). Second harem: New challenges and old wisdom. In Teixeira, A. J. S., de Lima, V. L. S., de Oliveira, L. C., and Quaresma, P., editors, *PROPOR*, volume 5190 of *Lecture Notes in Computer Science*, pages 212–215. Springer.

Silva, J. a., Branco, A., Castro, S., and Reis, R. (2010). Out-of-the-box robust parsing of portuguese. In *PROPOR 2010 – International Conference on Computational Processing of Portuguese Language*, pages 75–85.

Suakkaphong, N., Zhang, Z., and Chen, H. (2011). Disease named entity recognition using semisupervised learning and conditional random fields. *JASIST*, 62(4):727–737.

Thede, S. M. and Harper, M. P. (1999). A second-order hidden markov model for part-of-speech tagging. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 175–182, Stroudsburg, PA, USA. Association for Computational Linguistics.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Stroudsburg, PA, USA. Association for Computational Linguistics.