Comparing NERP-CRF with Publicly Available Portuguese Named Entities Recognition Tools

Daniela O. F. do Amaral[1], Evandro Fonseca[1], Lucelene Lopes[1], Renata Vieira[1]

[1] Pontifical Catholic University of Rio Grande do Sul, Computer Science,
Av. Ipiranga, 6681, Partenon, Porto Alegre, Brazil
{daniela.amaral, evandro.fonseca}@acad.pucrs.br
{lucelene.lopes, renata.vieira}@pucrs.br

Abstract. This paper presents the evaluation of NERP-CRF, a Conditional Random Fields (CRF) based tool for Portuguese Named Entities Recognition (NER) against other publicly available NER tools. The presented evaluation is based on the comparison with three other NER tools for Portuguese. The comparison is made observing Recall and Precision measures obtained by each tool over the HAREM corpus, a golden standard for NER for Portuguese texts. The experiments were initially conducted considering ten categories and then, considering a reduced number of categories. The results show that NERP CRF outperforms the others tools when sufficiently trained for four entity categories.

Keywords: Named Entity Recognition, Conditional Random Fields, Portuguese Language.

1 Introduction

Named Entity Recognition (NER) comprises extraction and classification of named entities according to several semantic categories [1]. The entities fall under categories such as person, organization and place. This task may consider also temporal entities such as date and time. NER is an important task in many research areas, including both general and specialized domains. For instance, well known NER applications are the recognition of disease and gene names in biomedical texts [2, 3].

The number of studies on NER for the Portuguese Language [4] is quite restricted when compared to other languages such as English. HAREM is the first and only initiative for Portuguese NER [5], which had so far two editions. HAREM set out two Golden Collections: the first and the second HAREM. The corpus has annotations for NE in ten categories: Person, Place, Organization, Value, Abstraction, Time, Work, Event, Thing and Other.

This paper presents a comparative study based on the Second HAREM corpus. First, we compare our tool NERP-CRF [6] trained over the ten HAREM categories with three other tools: FreeLing [7], LTasks [8], and PALAVRAS [9]. Then we modify the training sets, considering a reduced number of categories. This paper is organized as follows: Section 2 presents the HAREM corpora; Section 3 describes the tools under evaluation; Section 4 presents the evaluation process and results; and Section 5 presents our conclusions.

2 Corpora

HAREM is an event for the joint assessment of NER for Portuguese, established by Linguateca [10,11]. HAREM Golden Corpus (GC), was annotated by humans and has been used as a reference for NER systems evaluation. In [5, 12] evaluations of NER systems on the basis of HAREM corpora are presented. HAREM has two editions, Table 1 shows the distribution of NE in the 10 different categories for the corresponding golden corpora.

Table 1. Number of NE in each category according to both HAREM golden corpora.

Corpora	GC First HAREM 129 texts 466,355 words		GC Second HAREM 129 texts 89,241 words	
Categories				
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%	79	2%
Total	5,132	100%	7,255	100%

3 NER systems

We developed a system for Portuguese, called NERP-CRF. We compare it with three other tools. In general, there are few options for systems that perform NER for Portuguese. These three tools under evaluation were all that we could access and execute by ourselves. Two of them were publicly available, a third one is commonly used in Portuguese NLP groups, although it is not a freely available tool. In the following we present a brief description of the NER tools under analysis.

NERP-CRF: is a system based on the probabilistic mathematical model called Conditional Random Fields (CRF) [13]. The system was trained with First HAREM GC [14] using two input vectors. The first vector contains the POS tagging, and the Harem NE categories using BILOU notation [15]. The second is a vector of features, as described in [6]. The output is a vector with categories in BILOU notation.

Freeling: This system comprises a package of NLP tools, such as coreference, POS tagging and NER [7]. The Freeling works with texts in English, Spanish and Portuguese. It has two NER functions: the first one, simpler, is based on morphosyntactic patterns, and the second one, more elaborated, is based on machine learning algorithms. The latter form was used for comparisons in this work. This tool considers only the following NE categories: Person, Place, Organization and Other.

LTasks: LTasks is a set of web tools [8]. These tools are available but unfortunately they do not specify which techniques are used for NER. The categories are the same of HAREM with the exception of the category other.

PALAVRAS: The PALAVRAS parser is a software tool for Portuguese [9]. The output of PALAVRAS is a very rich annotation, where even syntax tree structures with all kinds of grammatical and semantic annotations are available. The system is rule-based.

4 Evaluation

The four tools described above were run over HAREM 2 by ourselves. First we present a comparison of the output for ten categories of each of the four systems (Table 2).

Table 2. P/R/F for all categories: Person, Place, Organization, Event, Work, Abstraction, Thing, Time, Value, and Other.

Person | RL | = 2,035						Place | RL | = 1,250				
Systems	P	R	F	| OE |	|OE ∩ RL |	P	R	F	| OE |	| OE ∩ RL |
FreeLing	54%	60%	57%	2,279	1,230	52%	60%	56%	1,431	751
LTasks	62%	61%	62%	2,017	1,249	56%	53%	54%	1,170	658
PALAVRAS	60%	64%	62%	2,174	1,297	54%	55%	54%	1,264	685
NERP-CRF	56%	50%	53%	1,803	1,012	48%	53%	51%	1,382	667
Organization | RL | = 960						Event | RL | = 302				
Systems	P	R	F	| OE |	OE ∩ RL |	P	R	F	OE |	| OE ∩ RL |
FreeLing	28%	60%	38%	2,088	575	-	-	-	-	-
LTasks	28%	60%	38%	2,043	576	12%	28%	17%	736	86
PALAVRAS	30%	51%	38%	1,630	491	53%	26%	35%	150	80
NERP-CRF	44%	48%	**46%**	1,054	460	42%	4%	7%	26	11
Work | RL | = 437						Abstraction | RL | = 278				
Systems	P	R	F	| OE |	| OE ∩ RL |	P	R	F	| OE |	| OE ∩ RL |
FreeLing	-	-	-	-	-	-	-	-	-	-
LTasks	26%	19%	22%	321	84	19%	14%	16%	201	39
PALAVRAS	36%	30%	33%	367	132	14%	6%	8%	117	16
NERP-CRF	44%	9%	15%	93	41	14%	8%	10%	155	22
Thing | RL | = 304						Time | RL | = 1,189				
Systems	P	R	F	| OE |	| OE ∩ RL |	P	R	F	| OE |	| OE ∩ RL |
FreeLing	-	-	-	-	-	-	-	-	-	-
LTasks	11%	5%	6%	129	14	5%	3%	4%	633	32
PALAVRAS	0%	0%	0%	22	0	-	-	-	-	-
NERP-CRF	6%	1%	1%	32	2	7%	3%	5%	624	41
Value | RL | = 352						Other | RL | = 79				
Systems	P	R	F	| OE |	| OE ∩ RL |	P	R	F	| OE |	| OE ∩ RL |
FreeLing	-	-	-	-	-	2%	15%	3%	638	12
LTasks	46%	46%	46%	351	163	-	-	-	-	-
PALAVRAS	-	-	-	-	-	-	-	-	-	-
NERP-CRF	42%	38%	40%	321	134	100%	3%	5%	2	2

After that we retrained NERP-CRF, with a reduced number of categories (Tables 4 and 5). For all these experiments the output of NE (OE) was compared with the second Harem GC annotation which was used as the reference list (RL) for the calculation of t Precision (P = | OE ∩ RL | / | OE |), Recall (R = |OE ∩ RL | / | RL |), and F-measure (F) as the harmonic average between P and R.

Table 2 shows that there was no system that outperformed in all categories. In fact, for categories Person, Place and Organization there is a balance of precision and recall among the results.

NERP-CRF had a better performance for the Organization category with 46% of F-measure. FreeLing had the best F-measure (56%) for Place category, while LTasks and PALAVRAS got the best F-measures for Person category.

FreeLing, LTasks and PALAVRAS were not able to identify all categories. Alas, besides the Person, Place and Organization categories the performance of all systems was either quite low (below 40%) or not representative (Value category was only detectable by LTasks and NERP-CRF).

Next we considered different distributions of categories for the training of NERP-CRF, focusing only on Person, Place and Organization categories. Thus, we performed four new trainings, all using First Harem. Again we used Second Harem for testing.

Initially, we grouped all other Harem categories (Event, Work, Abstraction, Thing, Time, Value and Other) into a single one called Everything Else (EE). The results for this new situation is presented in Table 3.

Table 3: P/R/F of NERP-CRF for categories: Person, Place, Organization and EE.

NERP-CRF – Four Categories					
	P	R	F	| OE |	| OE ∩ RL |
Person	84%	60%	**70%**	1,462	1,230
Place	49%	54%	51%	1,378	671
Organization	48%	46%	47%	918	442
EE	42%	11%	18%	793	332

We then we did three new trainings and these results are presented in Table 4. These results represent three different runs with two categories. The first with Person and all other categories grouped as EE. The second with Place and EE, and the third with Organization and EE.

The results of Tables 2, 3 and 4 show an interesting evolution in both Recall and Precision for the Person category when trained over four classes. The Precision values for Organization category that went from 44% for ten categories, to 48% for four categories, until impressive 60% for two categories. A similar evolution was observed for Place category with Precision evolving 48%, 49% and 62%. The accuracy increased as expected, due to the reduced number of categories to be learned.

Table 4: P/R/F of NERP-CRF for two categories (isolating Person, Place and Organization).

NERP-CRF – Two Categories – Person					
	P	R	F	| OE |	| OE ∩ RL |
Person	70%	36%	48%	1,043	732
EE	51%	47%	49%	4,721	2,422
NERP-CRF – Two Categories – Place					
	P	R	F	| OE |	| OE ∩ RL |
Place	62%	47%	53%	947	587
EE	66%	45%	54%	3,999	2,659
NERP-CRF – Two Categories – Organization					
	P	R	F	| OE |	| OE ∩ RL |
Organization	60%	38%	47%	620	369
EE	61%	50%	55%	5,153	3,143

5 Final Considerations and Future Work

This paper presented an evaluation of NERP-CRF against other NER systems. The first experiment has shown balanced results for the categories Person, Place and Organization, among the systems. The second experiment show possible evolutions for NERP-CRF system by reducing the number of categories, and this improvement is more relevant in terms of Precision. The overall comparison performed led us to believe that the method used by NERP-CRF, due to the use of sets of training and testing, has a better potential for improvement than the other systems.

Additionally to the results obtained with the second experiment, NERP-CRF system also can be improved by the development of a more elaborate set of features which will be applied to the training corpus. This belief is justified by the high accuracy achieved by NERP-CRF in experiment involving Place category.

Another future work of our interest is to specialize the Place category into sub-categories considering specific domains such as Geology to perform the classification task of NE.

References

1. Jiang, J.: Information extraction from text. In: Mining Text Data. Chap. 2, pp. 11--41. Springer, New York (2012)
2. Settles, B.: Biomedical named entity recognition using conditional random fields and rich feature sets. In: Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications, pp. 104--107 (2004)

3. Suakkaphong, N., Zhang, Z., Chen, H.: Disease Named Entity Recognition Using Semisupervised Learning and Conditional Random Fields. Journal of the American Society for Information Science and Technology, pp. 727--737 (2011)

4. Batista, S., Silva, J., Couto, F., Behera, B.: Geographic Signatures for Semantic Retrieval. In: 6th Workshop on Geographic Information Retrieval, pp.18--19. ACM (2010)

5. Freitas, C., Mota, C., Santos, D., Oliveira, H. G., Carvalho, P.: Second HAREM: Advancing the State of the Art of Named Entity Recognition in Portuguese. In: 7th International Conference on Language Resources and Evaluation, pp.363--3637. LREC. European Language Resources Association. ELRA, Valletta. (2010)

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

7. Padró, L., Collado, M., Reese, S., Lloberes, M., Castellón, I.: FreeLing 2.1: Five Years of Open-Source Language Processing Tools. In: 7th international conference on Language Resources and Evaluation, LREC pp.3485--3490 (2010)

8. LTasks – Language Tasks, http://ltasks.com

9. Bick, E. 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, Springer, pp. 80–89. (2006).

10. Santos, D., Cardoso, N.: Reconhecimento de entidades mencionadas em português: Documentação e atas do HAREM, a primeira avaliação conjunta na área. In: Santos, D., Cardoso, N. (eds.). Chap.1, pp. 1--16 (2008)

11. Santos, D.: Caminhos percorridos no mapa da portuguesificação: A linguateca em perspectiva. Linguateca. Vol.1, pp.25--59 (2009)

12. Carvalho, P., Oliveira, H.G., Mota, C., Santos, D., Freitas, C.: Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. In: Mota, C., Santos, D. (eds.) Linguateca. Chap.1, pp.11--31 (2008)

13. Lafferty, J., McCallum, A., Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In:18th International Conference on Machine Learning ICML, pp.282--289 (2001)

14. Santos, D., Cardoso, N.: Reconhecimento de entidades mencionadas em português: Documentação e atas do HAREM, a primeira avaliação conjunta na área, Chap. 20, p. 307–326 (2007)

15. Ratinov, L., Roth, D.: Design Challenges and Misconceptions in Named Entity Recognition. In: 13th Conference on Computational Natural Language Learning, CONLL, pp. 147--155 (2009)