Proceedings of Symposium in Information and Human Language Technology. Natal, RN, Brazil, November 4–7, 2015. ©2015 Sociedade Brasileira de Computação.

Comparative Analysis between Notations to Classify Named Entities using Conditional Random Fields

Daniela Oliviera F. do Amaral, Maiki Buffet, Renata Vieira

[1]Faculdade de Informática – Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS)
Caixa Postal 1429 – 90619-900 – Porto Alegre – RS – Brazil

{daniela.amaral, maiki.buffet}@acad.pucrs.br, renata.vieira@pucrs.br

Abstract. *Conditional Random Fields (CRF) is a probabilistic Machine Learning (ML) method based on structured prediction. It has been applied in several areas, such as Natural Language Processing (NLP), image processing, computer vision, and bioinformatics. In this paper we analyse two different notations for identifying the words that compose a Named Entity (NE): BILOU and IO. We found out that IO notation presents better results in F-measure than BILOU notation in all categories of HAREM corpus.*

1. Introduction

NER is the task of identifying Named Entities (NEs), mostly proper nouns, from free texts and to classify them within a set of pre-defined categories that includes Person, such as "Carlos Ribeiro"; and Place, such as "Porto Alegre" [Freitas et al. 2010]. NER has been largely applied in texts through methods such as supervised learning to classify addition to the above categories, also, diseases and genes in the abstracts of the medical field [Ray et al. 2014]. Labeled data and a set of automatically extracted features are used to train models, such as Maximum Entropy Markov Models (MEMMs) [McCallum et al. 2000] or CRF [Pinto et al. 2003]. The key difference between CRF and MMEMs is that MMEMs use exponential models by states for conditional probabilities of upcoming states, considering the current state. Within this context, the method chosen for this study was CRF, that was evaluated in previous studies for this task [Amaral and Vieira 2014b].

Different notations are used to annotate data for the NER task. In previous studies, we used BILOU [Ratinov and Roth 2009]. This notation demarcates the NEs as follows: B (Begin), I (Inside), L (Last), O (Outside) and U (Unit), indicating the beginning, continuation and end of a compound NE, or whether the word does not refer to a NE or refers to an unit NE. The IO notation [Tjong Kim Sang and De Meulder 2003] is a simpler alternative. It defines whether a word is a NE or not I (Inside) or O (Outside), respectively. Therefore, this paper presents a comparative study, which consists in two different notations for identifying the words that compose a Named Entity (NE): BILOU and IO.

This article is structured as follows: Section 2 presents Related Work. Section 3 describes the development of the NERP-CRF system. Section 4 presents the evaluation process and the results we obtained. Section 5 points to the conclusions and further work.

2. Related Work

Therefore, as one analyses the results generated from CRF, it is found that it is possible to improve them by modifying the identification markings of the NEs through different

types of notations. In [Ratinov and Roth 2009], for example, were applied two popular notations in the literature, BILOU and BIO, in their experiments for NER with the use of CRF.

Another interesting notation was applied in [Weber and Vieira 2014] using Stanford NER model [Sobhana et al. 2010]. Words that were not recognized as NEs were labeled as O (Outside). Words identified as NEs, in turn, received the classification Person, Place or Organization.

Similarly, the study by [Finkel et al. 2005] implemented a model based on the algorithm Gibbs sampling, in which specific labels were applied to the domain used, such as Person, Place and Organization, as well as consistent features extracted to generate the CRF model.

For this work, the employee corpus was the The Golden Collection (GC) HAREM [Santos and Cardoso 2007]. The NEs identified and classified by NERP-CRF received one of the ten categories established by HAREM: Abstraction, Event, Thing, Place, Work, Organization, Person, Time, Value and Other. Thus, our study differs from others due to the focus we give to our system, once the literature presents few studies that identify with different kinds of notations, and classify NEs, using the ten categories of HAREM in a corpus in Portuguese through CRF.

3. NERP-CRF System

This section describes the development of the NERP-CRF system [Amaral and Vieira 2014a] since the preprocessing of texts, as well as the model generated by CRF for NER. The elaboration of the model consists of two steps: training and testing. Thus, we adopted the HAREM's (GC) corpus that is divided into a set of texts for training and a set of texts for testing. The texts used as input for NERP-CRF are in the XML format with the categorization of the ENs and POS tagging. The system creates a preprocessing vector with this data. After this, the NEs are labeled with two alternative notations: BILOU and IO. These labels are also put in the previous vector. The goal of comparing BILOU and IO is to examine if a simplified notation such as this can increase learning performance. After the labeling, the feature vector is generated [Amaral and Vieira 2014a]. The features aim to characterize all the words in the corpus chosen for this process, directing the CRF in the identification and classification of the NEs. The input used for the CRF in the training step are the preprocessing vector and the features vector.

In the testing step, a set of texts is sent to NERP-CRF. This system: (a) creates the POS vector; (b) sends these vector and the same features vector to the CRF model generated in the training step, which, in turn (c) classifies the NEs of the corpus under study. Finally, the extracted NEs and the metrics precision, recall and F-measure are presented to the users of the system. The system process is completed with the output vector, which classifies the text with the notation applied and with the categories of the Second HAREM. The Table 1 illustrates the system output given the sentence: "Maria Antonia sonha em visitar Roma" (Maria Antonia dreams about visiting Rome).

Table 1. Two outputs of NERP-CRF: BILOU notation and IO notation.

	Maria	Antonia	sonha	em	visitar	Roma
BILOU	B	L	O	O	O	U
IO	I	I	O	O	O	I
CATEGORIES	PERSON	PERSON	-	-	-	PLACE

4. Evaluation of NERP-CRF

The results from the experiments were obtained according to the metrics: Precision, Recall and F-Measure [Mota and Santos 2008]. Therefore, this evaluation aims to find the most appropriate annotation to the NER task in the HAREM corpus. Our model has been demonstrating good results in comparison with other methods that use machine learning for the NER task [Amaral and Vieira 2014b].

Four experiments were carried out using the NERP-CRF system. For training, they all operated with the GC of the First HAREM, which encompasses 129 texts, and, for testing, with the GC of the Second HAREM, formed by over 129 texts. The two sets total 258 texts and approximately 237.232 words. The experiments differ from one another because of the following characteristics: Experiment 1: uses the BILOU notation and classifies the NEs according to the ten categories of HAREM; Experiment 2: uses the IO notation and classifies the NEs according to the ten categories of HAREM; Experiment 3: uses the BILOU notation and classifies the NEs in the categories Person, Organization, Place and Other. These categories were chosen due to the fact that they have been more widely studied within the field of IE [Weber and Vieira 2014] Experiment 4: uses the IO notation and classifies the NEs according to the same categories of experiment 3.

4.1. Results

Table 2 summarizes the performance of the ten categories with the BILOU and IO notations, respectively. When comparing Experiments 1 and 2, it is found that NERP-CRF presented better results for the ten categories with the IO notation. The highlight is for the category Event, which went from 14.347% to 19.745% in the F-measure. The IO notation contributed for that class to become more comprehensive and precise. Experiments 3 and 4 were carried out to see the learning behavior when the number of categories was reduced. Table 3 shows the performance of the BILOU and IO notations in the classification of NEs with the categories: Person, Place, Organization and Other. Again, there was a percentage increase of the F-measure when NERP-CRF identified them with the IO notation. Only the category Place kept a very similar value.

It is interesting to highlight that the category Organization had an increase of precision in experiment 4 in relation to experiment 3 (from 41.893% to 45.123%). This means that, when the system identified the NEs in the simplest way, it reached a larger number of correctly classified NEs in relation to the NEs that it managed to classify. In this scenario and generally speaking, the IO notation allowed an increase in the results compared to BILOU, both for the ten and for the four categories of HAREM.

Error analysis showed that NERP-CRF needs to improve the identification and classification of the NEs. The most frequent errors were: classification between the categories Place and Person, classification of acronyms and foreign words.

Table 2. Results of NERP-CRF for Experiments 1 and 2.

Categories	Recall		Precision		F-Measure	
	BILOU	IO	BILOU	IO	BILOU	IO
PERSON	58.98%	**61.04%**	65.85%	65.63%	62.23%	63.25%
PLACE	53.91%	55.58%	49.71%	49.81%	51.73%	52.54%
ORGANIZATION	54.18%	52.03%	38.70%	41.34%	45.15%	46.08%
EVENT	08.2%	**11.56%**	56.89%	**67.39%**	14.34%	19.74%
WORK	13.99%	14.48%	52.55%	48.14%	22.10%	22.27%
TIME	30.12%	30.78%	88.91%	87.98%	44.99%	45.60%
THING	01.43%	01.43%	22.85%	**33.33%**	02.70%	02.75%
ABSTRACTION	06.13%	06.42%	15.16%	17.67%	08.73%	09.42%
VALUE	66.11%	67.91%	67.02%	67.41%	66.56%	67.66%
OTHER	02.29%	02.29%	57.14%	80.00%	04.41%	04.46%

Table 3. Results of NERP-CRF for Experiments 3 and 4.

Categories	Recall		Precision		F-Measure	
	BILOU	IO	BILOU	IO	BILOU	IO
PERSON	56.28%	58.07%	67.60%	68.75%	61.42%	62.96%
PLACE	52.08%	51.88%	52.34%	53.87%	52.21%	52.86%
ORGANIZATION	51.70%	49.22%	41.89%	45.12%	46.28%	47.08%
OTHER	35.00%	37.93%	76.33%	72.67%	47.99%	49.84%

5. Conclusion and Future Work

NERP-CRF was the system developed to perform two functions: the identification of NEs and the classification of these NEs based on the ten categories of HAREM. For the four experiments that were conducted, it was possible to observe that all results of the IO notation, both for ten and for four categories, were higher than those of the BILOU notation. Therefore, we perceived that less granularity makes it easier for the system to learn NER. Consequently, the importance of changing notations in sentences enables a better classification of the NEs, so that the CRF can obtain even more accurate and comprehensive results under a specific domain corpus.

The error analysis suggests a future work with experiments using meta-learning algorithms, such as the combination of classifiers, to increase the effectiveness of NERP-CRF as the use of the algorithm AdaBoosting [Carreras et al. 2003] and Coreference Resolution [Fonseca et al. 2014].

References

Amaral, D. O. F. d. and Vieira, R. (2014a). Nerp-crf: uma ferramenta para o reconhecimento de entidades nomeadas por meio de conditional random fields. *Linguamática*, 6(1):41–49.

Amaral, Daniela; Fonseca, E. L. L. and Vieira, R. (2014b). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, pages 2554–2558. International Conference on Language Resources and Evaluation, Reykjavik.

Carreras, X., Màrquez, L., and Padró, L. (2003). A simple named entity extractor using adaboost. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 152–155. Association for Computational Linguistics.

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*, pages 363–370. Association for Computational Linguistics.

Fonseca, E. B., Vieira, R., Vanin, A., and do Sul, G. (2014). Coreference resolution for portuguese: Person, location and organization. *International Conference on Computational Processing of Portuguese (PROPOR)*, pages 1–8.

Freitas, C., Mota, C., Santos, D., Oliveira, H. G., and Carvalho, P. (2010). Second harem: Advancing the state of the art of named entity recognition in portuguese. In *LREC*. Citeseer.

McCallum, A., Freitag, D., and Pereira, F. C. N. (2000). Maximum entropy markov models for information extraction and segmentation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 591–598, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Mota, C. and Santos, D. (2008). Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O segundo harem.

Pinto, D., McCallum, A., Wei, X., and Croft, W. B. (2003). Table extraction using conditional random fields. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 235–242. ACM.

Ratinov, L. and Roth, D. (2009). Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CONLL)*, pages 147–155.

Ray, W. C., Wolock, S. L., Callahan, N. W., Dong, M., Li, Q. Q., Liang, C., Magliery, T. J., and Bartlett, C. W. (2014). Addressing the unmet need for visualizing conditional random fields in biological data. *BMC bioinformatics*, 15(1):202.

Santos, D. and Cardoso, N. (2007). Reconhecimento de entidades mencionadas em português: Documentação e actas do harem, a primeira avaliação conjunta na área.

Sobhana, N., Mitra, P., and Ghosh, S. (2010). Conditional random field based named entity recognition in geological text. *International Journal of Computer Applications*, 1(3):143–147.

Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147. Association for Computational Linguistics.

Weber, C. and Vieira, R. (2014). Building a corpus for named entity recognition using portuguese wikipedia and dbpedia. *International Conference on Computational Processing of Portuguese (PROPOR)*.