A Sequence Model Approach to Relation Extraction in Portuguese

Sandra Collovini, Gabriel Machado, Renata Vieira
PUCRS University – Porto Alegre – Brazil
sandra.abreu@acad.pucrs.br, gabriel.machado.002@acad.pucrs.br, renata.vieira@pucrs.br

Abstract

The task of Relation Extraction from texts is one of the main challenges in the area of Information Extraction, considering the required linguistic knowledge and the sophistication of the language processing techniques employed. This task aims at identifying and classifying semantic relations that occur between entities recognized in a given text. In this paper, we evaluated a Conditional Random Fields classifier for the extraction of any relation descriptor occurring between named entities (Organisation, Person and Place categories), as well as pre-defined relation types between these entities in Portuguese texts.

Keywords: Information Extraction, Relation Extraction, Conditional Random Fields

1. Introduction

Information Extraction (IE) is a process of getting structured data from unstructured information in the text (Bach and Badaskar, 2007; Jurafsky and Martin, 2009). After this structured information can be used by a wide range of NLP applications.

Usually, IE can be regarded as a pipeline process, in which some type of information is extracted at each step. Relation Extraction (RE) is one of the stages of IE, which aims to identify and classify semantic relations that occur between entities recognized in a given text (Bach and Badaskar, 2007; Jurafsky and Martin, 2009). The two major types of RE are closed domain and open domain (Banko and Etzioni, 2008): closed-domain RE systems consider only a closed set of relations between two arguments, while open-domain RE systems do not need a pre-specified definition of the relation.

Currently, there are plenty of systems for RE from unstructured data and there are different methods for dealing with this task. Among supervised methods stands Conditional Random Fields (CRF), which are very powerful for segmenting and labeling sequential data (Lafferty et al., 2001). CRF have now become almost a standard for the task of Named Entity Recognition (NER) (McCallum and Li, 2003), and have more recently been applied to the task of RE from text (Culotta et al., 2006; Banko and Etzioni, 2008; Wu and Weld, 2010; Li et al., 2011; Collovini et al., 2014).

In this paper, we evaluated a CRF classifier in two scenarios: the extraction of any relation descriptor occurring between named entities (Organisation, Person and Place categories), and the extraction of relation descriptors expressing pre-defined types of relations between these entities (*"affiliation"* and *"placement"* relations). We define a relation descriptor as the text chunks that describe the explicit relation, occurring between a pair of named entities in the sentence. For example, we have the relation descriptor *"professor at"* between the Person named entity *"Hugo Doméch"* and the Organisation named entity *"Universidade Jaume de Castellón"* in the sentence:

*"Hugo Doméch, **professor da** Universidade Jaume de Castellón."* (*Hugo Doméch, **professor at** Universidade Jaume de Castellón.*)

This work is organized as follows. In Section 2, we review the related work. The RE model is described in Section 3. In Section 4, we describe the experiments. The results are discussed in Section 5. We conclude in Section 6.

2. Related Work

CRF have been applied efficiently in many tasks of sequential text processing (Culotta et al., 2006; Banko and Etzioni, 2008; Wu and Weld, 2010; Li et al., 2011). (Culotta et al., 2006) presents the integration of a supervised method that learns relational and contextual patterns for the extraction of familiar relations (*"mother"*, *"cousin"*, *"friend"* etc.). In (Li et al., 2011) relation descriptors are extracted, considering pre-defined types of relations (*"employment"* and *"personal/social"*).

These works are systems that extract specific relations (closed-domain RE). There are open-domain RE systems that use CRF, among them, the O-CRF system (Banko and Etzioni, 2008) uses a compact set of lexicon-syntactic patterns, and WOEpos (Wikipedia-based Open Extractor) (Wu and Weld, 2010) uses Wikipedia and features based on POS annotation.

There are very few proposals for RE for Portuguese. Among the RE systems for Portuguese, three systems took part in the ReRelEM[1] track of Second HAREM[2]. REMBRANDT (Recognition of Named Entities Based on Relations and Detailed Text Analysis) (Cardoso, 2008) recognized four relation types: *"identity"*, *"inclusion"*, *"placement"*, and *"other"*, using Portuguese Wikipedia and some grammar rules. SeRELeP (System for Recognition of Relations for the Portuguese language) (Brucksen et al., 2008) aimed at recognizing three relation types: *"identity"*, *"inclusion"* and *"placement"*, using the informations provided by

[1]Recognition of Relation between Named Entities
[2]http://www.linguateca.pt/LivroSegundoHAREM/

PALAVRAS parser (Bick, 2000). SEI-Geo (Chaves, 2008) is an extraction system that deals with NER concerning only the Place category and its relations, using Geo-ontologies. Also work mentioning the work of Batista et al. (Batista et al., 2013), which proposes an approach of distantly supervised relation extraction between two entities. The authors selected 10 relation types from Portuguese Wikipedia, such as *"located-in"*, *"successor-of"*, and others.

For all Portuguese works above, the set of relations was previously defined (closed-domain RE). There are few RE systems which apply Open IE for Portuguese language. A multilingual dependency-based Open IE system (DepOE) has been proposed in (Gamallo et al., 2012), it was used to extract triples from the Wikipedia in four languages: Portuguese, Spanish, Galician and English. In (Santos and Pinheiro, 2015), the RePort system is presented, it is a method of Open IE for Portuguese based on the ReVerb system (Fader et al., 2011) for English.

In previous research (Collovini et al., 2014), we extract relations between named entities (NEs) in the Organization domain, using CRF for Portuguese. We evaluated different feature configurations for CRF based of lexical, syntactic and semantic information.

In this work, we test the same CRF on open and closed RE tasks for Portuguese. We extract any relation descriptors that express any type of relation between the NEs, and two pre-defined types of relations (*"employment"* and *"placement"*) between these entities.

3. The CRF Model

In this work, we applied the linear-chain CRF, which occurs when output nodes of the graphical model are linked by edges in a linear chain.

According to the definition of linear-chain CRF, let $\mathbf{o} = (o_1, o_2, ..., o_T)$ be the sequence of observed input data (values on T input nodes); let S be a set of states, in which each state is associated with a label L; and $\mathbf{s} = (s_1, s_2, ..., s_T)$ is the sequence of states corresponding to the T output nodes.

In this paper, we consider each word of a sentence as an observation $\boldsymbol{o}$, which receives a L label according to an IO notation defined in previous work (Collovini et al., 2015). Two labels are considered: {*I-REL, O*}, where a word labelled with I-REL is Inside of a relation descriptor, while a word labelled with O is Outside of the relation descriptor. An illustration is given in Table 1.

Linear-chain CRFs define the conditional probability of state sequence given an input sequence as $p(\mathbf{s}|\mathbf{o})$, described in (1):

$$p(\mathbf{s}|\mathbf{o}) = \frac{1}{Z_o}\exp(\sum_{t=1}^{T}\sum_{k=1}^{K}\lambda_k f_k(s_{t-1}, s_t, \mathbf{o}, t)), \quad (1)$$

where Z_o is the normalization factor over all state sequences; $f_k(s_{t-1}, s_t, \mathbf{o}, t)$ is an arbitrary feature function over its arguments; $\lambda_k \in (-\infty; +\infty)$ is a learned weight for each feature function. The factor Z_o corresponds to the sum of the scores of all possible state sequences, and the number of state sequences is exponential in the input sequence length T.

Words	IO scheme
Hugo Doméch	O
,	O
professor	**I-REL**
de	**I-REL**
o	O
Universidade Jaume de Castellón	O
.	O

Table 1: Relation Extraction as Sequence Labeling using IO scheme.

Generally, the features functions f_k can ask arbitrary questions about the input sequence, including queries about previous words, next words, and combinations of all these.

In this paper, we use relation-specific features for Portuguese described in previous work (Collovini et al., 2014). The sets of features are: *Part-Of-Speech* (e. g. POS tag); *lexical* (e. g. canonic form), *syntactic* (e. g. syntactic tag); *patterns* (e. g. verb followed by a preposition); *phrasal sequence* (e. g. POS tags of the word sequence between two NEs); *semantic* (e. g. NE category). Feature vectors were generated for the pair of the NEs and also for the words between them, resulting in a vector with 57 elements for each word. An illustration of some features is given in Table 2.

Words	POS	Lexical	Syntactic
Hugo Doméch	PROP	Hugo Doméch	@VOK
,		,	
professor	N	professor	@N>PRED
de	PRP	de	@N>
o	DET	o	@>N
Universidade Jaume de Castellón	PROP	Universidade Jaume de Castellón	@P<
.		.	

Table 2: Example of the features.

4. Experiments

4.1. Experiment Setup

In this paper, we performed two experiments in order to evaluate the results of the CRF classifier for relations between NEs of Organisation, Person and Place categories. In these experiments, we considered the extraction of any type of relations, as well as pre-defined types of relations, from Portuguese texts. The experiments and the respective evaluation are described below:

Experiment 1: CRF classifier extracting any relation descriptor for pair of NEs of type Organisation-Person and Organisation-Place (see Table 3).

Experiments	NEs	Relations	#Total	Positive	Negative
Experiment 1	ORG-PERS	*open*	171	95	76
	ORG-PLACE	*open*	170	97	73
Experiment 2	ORG-PERS	*affiliation*	132	61	71
	ORG-PLACE	*placement*	80	40	40

Table 3: Number of instances.

Experiment 2: CRF classifier extracting specific relation descriptors for pair of NEs of type Organisation-Person and Organisation-Place (see Table 3).

Evaluation: application of 5-folds[3] cross validation for all experiments, and evaluation from to manual annotation of relation descriptors using two criteria (Collovini et al., 2015): *exact matching* (having all words in common) and *partial matching* (having at least one word in common). An illustration is given in Table 4.

Words	Exact matching	Partial matching
Hugo Doméch	O	O
,	O	O
professor	**I-REL**	**I-REL**
de	**I-REL**	O
o	O	O
Universidade Jaume de Castellón	O	O
.	O	O

Table 4: Example of the evaluated criteria for relation descriptors.

4.2. Data

We used subsets of the HAREM's Golden Collections[4] (GCs) for NER. All texts already have the annotations of the NEs, and we opted for the categories Person, Organisation and Place.

First, we selected texts that deal with the Organization domain (e. g. opinion, journalistic etc.) from the First and Second HAREM and added to these texts the manual annotation of any relation descriptor occurring between pairs of NEs (ORG-PERS or ORG-PLACE) in the same sentence of the text (Experiment 1).

After, we selected texts from the ReRelEM track of the Second HAREM that contained two specific relations between pairs of NEs (ORG-PERS and ORG-PLACE) in the same sentence of the text: *affiliation* relations that occur between pairs of Organisation and Person; and *placement* relations that occur between Organization and Place (Experiment 2).

Table 3 illustrates the total number of relation instances (#Total), the number of positive, and the number of negative instances used in each experiment. Positive instances are those that have an explicit relation descriptor between two NEs, negative instances are those that do not meet this condition. Example of positive and negative relation instances from ORG-PERS are presented in Table 5, respectively.

Relation instance	Relation descriptor
"*Steve Jobs*, o **director-geral** da empresa, foi o ponto alto para os fãs da *Apple*." (*Steve Jobs*, the **CEO of** the company, was the highest point for *Apple* fans.)	**director-geral de** (**CEO of**)
"*Saraiva Dias*, vereador da autarquia, referiu ao *Público*." (*Saraiva Dias*, deputy of the municipality, referred to the *Público*.)	–

Table 5: Examples of the positive and negative relation instances from ORG-PERS.

5. Results and Discussion

In this section, we present the results of the application of the CRF classifier for each experiment using the following measures: Precision (P), Recall (R) and F-measure (F).

Overall, we achieved high rates of Precision for the experiments, it occurred due to fact that the CRF classifier is very precise in the process of tagging the relation descriptors. The best rates of F-measure were obtained for *placement* relation considering *exact matching*, and for *affiliation* relation with *partial matching*.

In the open RE task there is a great diversity of relations, which makes their classification more difficult. However, for the extraction of pre-defined relations (closed RE), the CRF classifier only needs to learn specific types of relations. The choice of the extracted relation type depends mainly on the type of information being analyzed and on the objective of the extraction task.

Table 6 presents the results for *exact matching*, in which the best values of F-measure were 47% for *open* relation (Experiment 1), and 57% for *placement* relation (Experiment 2), both results for relations between ORG-PLACE NEs.

For *partial matching*, shown in Table 7, we achieved the best rates for relations between ORG-PERS NEs: F-measure of 61% for *open* relation (Experiment 1) and 63% for *affiliation* relation (Experiment 2).

[3]We apply 5-folds due to reduced data size.

[4]http://www.linguateca.pt/harem/

Experiments	NEs	Relations	P	R	F
Experiment 1	ORG-PERS	*open*	0.45	0.38	0.41
	ORG-PLACE	*open*	0.56	0.41	**0.47**
Experiment 2	ORG-PERS	*affiliation*	0.65	0.49	0.56
	ORG-PLACE	*placement*	0.73	0.47	**0.57**

Table 6: Results of the Experiments with *exact matching*.

Experiments	NEs	Relations	P	R	F
Experiment 1	ORG-PERS	*open*	0.65	0.56	**0.61**
	ORG-PLACE	*open*	0.71	0.52	0.60
Experiment 2	ORG-PERS	*affiliation*	0.73	0.55	**0.63**
	ORG-PLACE	*placement*	0.73	0.47	0.57

Table 7: Results of the Experiments with *partial matching*.

It is difficult to make a comparison with other works for English, since the languages and data sets are different (Abreu et al., 2013). In Table 8 we show the results achieved in other open RE systems for English using Conditional Random Fields: O-CRF (Banko and Etzioni, 2008) and WOEpos (Wu and Weld, 2010). We can see that our results considering any relation descriptors between NEs (open RE) are not distant from these works.

Works	Data/Language	Performance
CRF	HAREM's GC	ORG-PERS: F=0.61 ORG-PLACE: F=0.60
O-CRF	Sent500 corpus (Bunescu, 2007)	F=0.59
WOEpos	Wikipedia and Web pages	Wikipedia: F=0.57 Web: F=0.65

Table 8: Results reported by open RE systems.

One of the obstacles for an adequate evaluation of relation extraction in Portuguese is the lack of common data. However the *placement* relation was also treated by SeRELeP and REMBRANDT in the Second HAREM's ReRelEM track. The comparison is shown in Table 9, we can see that our results are better than other works.

Works	P	R	F
CRF Classifier	**0.73**	**0.47**	**0.57**
SeRELeP	0.36	0.27	0.31
REMBRANDT	0.40	0.12	0.19

Table 9: Results reported to the *"placement"* relation.

6. Conclusion

We evaluated a CRF classifier for the extraction of any relation descriptor occurring between NEs (open RE), as well as pre-defined relation types between these entities (closed RE).

We achieved the best results for an extraction of the pre-defined relations considering *exact* and *partial matching*, but the CRF classifier is capable of extracting any type relation between NEs. Overall, the best results were for *partial matching*, it occurred due to the difficulty of classifying every element included in a descriptor. However, the instances evaluated as *partial matching* were enough to represent the existing relations.

Since there are very few proposals for open relation extraction for Portuguese (Abreu et al., 2013), contrary to the situation for other languages, the difficulty of the task is enhanced. This work contributed for the progress in this area for Portuguese, that has a demand for the development of news methods, tools and specific resources such as annotated data.

The produced resources related to this paper (the subset of texts and corresponding positive relation instances manually annotated in tuples *(NE1, relation descriptor, NE2)* used in each experiment) are electronically available at:

http://www.inf.pucrs.br/linatural/data_set_RE.html

In future work, we plan to use open source tools; as well as performing an extension of the proposed process for other languages.

7. Acknowledgments

The authors acknowledge the financial support of CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) and FAPERGS (Fundação de Amparo à Pesquisa do Rio Grande do Sul).

8. References

Abreu, S. C., Bonamigo, T. L., and Vieira, R. (2013). A review on relation extraction with an eye on portuguese. *Journal of the Brazilian Computer Society*, 19(4):553–571.

Bach, N. and Badaskar, S. (2007). A survey on relation extraction. Technical report, Literature review for Language and Statistics II, Carnegie Mellon University.

Banko, M. and Etzioni, O. (2008). The tradeoffs between open and traditional relation extraction. In

McKeown, K., Moore, J. D., Teufel, S., Allan, J., and Furui, S., editors, *Association for Computer Linguistics - ACL*, pages 28–36. The Association for Computer Linguistics.

Batista, D. S., Forte, D., Silva, R., Martins, B., and Silva, M. (2013). Extracção de relações semânticas de textos em português explorando a DBpédia e a Wikipédia. *linguamatica*, 5(1):41–57.

Bick, E. (2000). *The Parsing System Palavras. Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. University of Arhus.

Brucksen, M., Souza, J. G. C., Vieira, R., and Rigo, S. (2008). Sistema serelep para o reconhecimento de relações entre entidades mencionadas. In Mota, C. and Santos, D., editors, *Segundo HAREM*, chapter 14, pages 247–260. Linguateca.

Bunescu, R. C. (2007). Learning to extract relations from the web using minimal supervision. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 576–583, Prague, Czech Republic. Association for Computational Linguistics.

Cardoso, N. (2008). Rembrandt – reconhecimento de entidades mencionadas baseado em relações e análise detalhada do texto. In Mota, C. and Santos, D., editors, *Segundo HAREM*, chapter 11, pages 195–211. Linguateca.

Chaves, M. S. (2008). Geo-ontologias e padrões para reconhecimento de locais e de suas relações em textos: o sei-geo no segundo harem. In Mota, C. and Santos, D., editors, *Segundo HAREM*, chapter 13, pages 231–245. Linguateca.

Collovini, S., Pugens, L., Vanin, A. A., and Vieira, R. (2014). Extraction of relation descriptors for portuguese using conditional random fields. In *In Proceedings of Advances in Artificial Intelligence - IBERAMIA 2014 - 14th Ibero-American Conference on Artificial Intelligence*, pages 108–119, Santiago de Chile, Chile.

Collovini, S., de Bairros Filho, M., and Vieira, R. (2015). Analysing the role of representantion choices in portuguese relation extraction. In *In Proceedings of Conference and Labs of the Evaluation Forum - CLEF 2015*, pages 91–102, Toulouse, France. Springer.

Culotta, A., McCallum, A., and Betz, J. (2006). Integrating probabilistic extraction models and data mining to discover relations and patterns in text. In *Proceedings of the main conference on HLT-NAACL*, HLT-NAACL '06, pages 296–303, Stroudsburg, PA, USA. Association for Computational Linguistics.

Fader, A., Soderland, S., and Etzioni, O. (2011). Identifying relations for open information extraction. In *Proceedings of Empirical Methods in Natural Language Processing - EMNLP*, pages 1535–1545.

Gamallo, P., Garcia, M., and Fernández-Lanza, S. (2012). Dependency-based open information extraction. In *Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP*, pages 10–18, Avignon, France. Association for Computational Linguistics.

Jurafsky, D. and Martin, J. H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Prentice Hall series in artificial intelligence. Pearson Education Ltd., London, 2 edition.

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.

Li, Y., Jiang, J., Chieu, H. L., and Chai, K. M. A. (2011). Extracting relation descriptors with conditional random fields. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 392–400, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

McCallum, A. and Li, W. (2003). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 188–191, Stroudsburg, PA, USA. Association for Computational Linguistics.

Santos, V. and Pinheiro, V. (2015). Report ? um sistema de extração de informações aberta para língua portuguesa. In *Proceedings of the X Brazilian Symposium in Information and Human Language Technology (STIL)*, Natal, RN, Brazil. SBC.

Wu, F. and Weld, D. S. (2010). Open information extraction using wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 118–127, Stroudsburg, PA, USA. Association for Computational Linguistics.