RelP: Portuguese Open Relation Extraction

Sandra Collovini* de Abreu and Renata Vieira**

Pontifícia Universidade Católica do Rio Grande do Sul—PUCRS,
Av. Ipiranga, 6681, Partenon, Porto Alegre, Rio Grande do Sul, Brasil,
*<sandra.abreu@acad.pucrs.br>, **<renata.vieira@pucrs.br>

Sandra Collovini de Abreu develops research in the area of natural language processing; her main interests are in the following themes: information extraction, relation extraction, machine learning, named entity recognition, computational logic and corpora studies. She has a PhD in computer science from Pontifícia Universidade Católica do Rio Grande do Sul (2014). Currently, she is a post-doctoral researcher at the Natural Language Processing Research Laboratory at Pontifícia Universidade Católica do Rio Grande do Sul.

Renata Vieira develops research in the fields of artificial intelligence and natural language processing. She is mainly interested in the semantic and discursive aspects of language. In the field of knowledge representation, she is interested in the study of ontologies. She received her PhD in 1998 from the Cognitive Science Department at the University of Edinburgh, and has been leading the Natural Language Processing Research Laboratory at Pontifícia Universidade Católica do Rio Grande do Sul since 2008.

Collovini de Abreu, Sandra and Renata Vieira. 2017. "RelP: Portuguese Open Relation Extraction." *Knowledge Organization* 44(3): 163-177. 47 references.

Abstract: Natural language texts are valuable data sources in many human activities. NLP techniques are being widely used in order to help find the right information to specific needs. In this paper, we present one such technique: relation extraction from texts. This task aims at identifying and classifying semantic relations that occur between entities in a text. For example, the sentence "Roberto Marinho is the founder of Rede Globo" expresses a relation occurring between "Roberto Marinho" and "Rede Globo." This work presents a system for Portuguese Open Relation Extraction, named RelP, which extracts any relation descriptor that describes an explicit relation between named entities in the organisation domain by applying the Conditional Random Fields. For implementing RelP, we define the representation scheme, features based on previous work, and a reference corpus. RelP achieved state of the art results for open relation extraction; the F-measure rate was around 60% between the named entities person, organisation and place. For better understanding of the output, we present a way for organizing the output from the mining of the extracted relation descriptors. This organization can be useful to classify relation types, to cluster the entities involved in a common relation and to populate datasets.

Received: 29 November 2016; Revised: 9 March 2017; Accepted: 10 March 2017

Keywords: entities, relations, Portuguese, open relation extraction

1.0 Introduction

Relation extraction (RE) aims at identifying and classifying semantic relations that occur between (pairs of) entities recognized in a given text (Jurafsky and Martin 2009). The problem of relation extraction from natural language texts has been studied extensively, including news articles, science publications, blogs, e-mails and from sources like Wikipedia, Twitter and the Web (Sarawagi 2008). There is an increasing interest in relation extraction, mostly motivated by the exponential growth of information made available through the Web, which makes the tasks of researching and using this massive amount of data impossible through manual means. This context makes relation extraction an even more complex and relevant research area.

Given the importance of exploring semantic relations for a more accurate understanding of language, the need of further advances in techniques for identifying and extracting them emerged; thus, the establishment of the relation extraction task was necessary. Relation extraction can be useful for various knowledge organization activities such as information indexing and retrieval, question answering, text summarization, ontology construction, knowledge acquisition, knowledge representation of the

documents and thesaurus generation (Bertrand-Gastaldy 2001, Green 2001, Santos and Kobashi 2013; da Silva and Milidiú 1991; Guarino 1995).

Information extraction is usually concerned with finding as many semantic relations as possible, where they are relevant to a particular domain or application. On the other hand, the task of automatic ontology construction focuses on inferring knowledge structures from texts where it is important to find those relations which are relevant for the main concepts of the ontology. In other knowledge structures, such as thesauri, the relations between terms help both indexers and searchers to navigate thesauri in order to identify various kinds of related terms. The usefulness of semantic relations in information science is discussed in Khoo and Na (2006) in detail.

Several approaches have been proposed to relation extraction from unstructured data, such as supervised or unsupervised machine learning, corpus-based techniques, linguistic strategies, rule-based heuristics and hybrid systems. For some languages, such as English, there is extensive research and literature regarding relation extraction (Culotta et al. 2006; Banko et al. 2007; Yates et al. 2007; Banko and Etzioni 2008; Sarawagi 2008; Zhu et al. 2009; Wu and Weld 2010; Fader et al. 2011; Li et al. 2011; Zhang et al. 2015), while for Portuguese, there are fewer references to existing work dealing with relation extraction (Brucksen et al. 2008; Chaves 2008; Cardoso 2008; Batista et al. 2013; Ferreira et al. 2009; Santos et al. 2010). Unfortunately, for works in Portuguese, it is not possible to reuse resources and databases developed for English, so rule based approaches are usually applied.

Among supervised methods stand sequential models, such as Hidden Markov Model (HMM), Maximum Entropy Markov Model (MEMM) or Conditional Random Fields (CRF), which are very powerful for segmenting and labeling sequential data (Lafferty et al. 2001). CRF, the most sophisticated of the three families of models, have now become almost a standard for the task of Named Entity Recognition (NER) (McCallum and Li 2003; Dânger et al. 2014) and have more recently (Culotta et al. 2006; Banko and Etzioni 2008; Wu and Weld 2010; Li et al. 2011; Collovini et al. 2016a) been applied to the task of relation extraction from text.

In this paper, we present a system for Portuguese Open Relation Extraction, named RelP, to extract any relation descriptor occurring between Named Entities (NEs) in the organisation domain. This domain was chosen because of its potential applicability to different areas including competitive intelligence, risk management and sales and marketing. We apply the Conditional Random Fields model, and 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 "diretor de" (director of) that occurs between the named entities "Ronaldo Lemos" and "Creative Commons" (parameters of relation) in (1):

1) No próximo Sábado, Ronaldo Lemos, diretor da Creative Commons, irá participar de um debate [...]

Next Saturday, Ronaldo Lemos, director of Creative Commons, will participate in a debate [...]

In (1), the resulting triple extracted by the RelP system is (Ronaldo Lemos, diretor de, Creative Commons). These triples can be used to feed a knowledge base or a semantic network if the triples are connected together.

This paper is organized as follows: in Section 2.0, we present a literature review related to open RE. The RelP system is detailed in Section 3.0. Section 4.0 describes RelP's evaluation and discusses the obtained results. In Section 5.0, we structure the extracted relation descriptors. Finally, Section 6.0 presents the conclusion and future works.

2.0 Relation extraction

Relation extraction (RE) is the task of identifying and classifying the semantic relations between entities from natural language texts. It focuses on extracting structured relations from unstructured sources using different approaches (Zhang et al. 2015). The two major types of Relation Extraction are: 1) Relation Extraction systems in which the set of relations of interest can be previously defined (closed RE); and, 2) when there is no pre-defined relation type in the input (open RE).

Depending on the application and on the resource available, the RE task can be studied for different settings. For Portuguese, there are very few proposals for the task of RE compared to other languages like English. One of the main obstacles for the progress of this task is the lack of resources, such as annotated data, lexical bases and domain ontologies. A survey on RE task is found in (Collovini de Abreu et al. 2013), addressing the progress and difficulties of the area and considering the case of Portuguese in this scenario.

In this section, we present a literature review related to Portuguese RE. After, we present a brief overview of open RE for English.

2.1 Portuguese named entities

In the literature, there are many RE systems which start the process by applying Named Entity Recognition (NER) to identify the named entities in the texts. Next, the systems extract the relations and they also may classify the relation types.

There is a contest exclusively dedicated to the study of named entities of Portuguese language, namely HAREM (Evaluation of Systems for Named Entity Mention). The first event for HAREM evaluation happened in 2005 and it has followed MUC evaluation criteria (MUC-7 1997). Since then the process has gone through alterations (Santos and Cardoso 2007). The Second HAREM took place in 2008 and it allowed systems to choose categories, types/subtypes, and it also included the task of identifying semantic relations between NEs (Carvalho et al. 2008).

Among the resources from the First HAREM, there is the Golden Collection which comprises of 129 Brazilian and European Portuguese texts with 5,132 named entities manually annotated and distributed in ten categories (person, place, organisation, time, title, value, event, abstraction, thing and varied). The Golden Collection from the Second HAREM comprises of 129 Brazilian and European Portuguese texts with 7,255 named entities and 4,803 relations manually annotated.

To illustrate the annotation of the NEs in HAREM's Golden Collection, let us take the sentence described in (1) from a text of the Second HAREM. The corresponding XML annotation format is illustrated in (2). We can identify the annotation of two named entities with tag NE: "Ronaldo Lemos" and "Creative Commons," classified as person and organisation, respectively (tags PESSOA and ORGANIZACAO CATEG).

2)<NE ID=ric-163 CATEG=*PESSOA*> *Ronaldo Lemos* </NE>, *diretor da* <NE ID=ric-170 CATEG=*ORGANIZACAO*> *Creative Commons* </NE>

Systems that take part in joint evaluation conferences for Portuguese follow the conference directives. In the next subsection, the systems that participated in the Second HAREM are presented.

2.2 Portuguese relation extraction

HAREM is a milestone in joint evaluation efforts focused on Portuguese, and its second version included the task of identifying the semantic relations between named entities, called ReRelEM track (Recognition of Relation between Named Entities) (Freitas et al. 2008).

The relations defined in ReRelEM are identity (or coreference: entities with the same referent), inclusion (included/includes), placement (occurs in/place of) and other (relations that do not correspond to any other previously listed category, a set of 22 new relations).

In this section, we present the approaches used by systems that took part in the ReRelEM track (Brucksen et al. 2008; Chaves 2008; Cardoso 2008) and also works approaching relation extraction in Portuguese that are available in the reviewed literature (Cardoso 2012; Batista et al. 2013; Ferreira et al. 2009; Santos et al. 2010).

The REMBRANDT system (Recognition of Named Entities Based on Relations and Detailed Text Analysis) (Cardoso 2008) was developed to recognize all categories of named entities and relations between them (identity, inclusion, placement and other). This system makes use of Portuguese Wikipedia as an external resource, as well as some grammar rules that describe internal and external evidence about named entities. According to Cardoso (2012), REMBRANDT is now a mature tool and it can therefore be used by the NLP community on several information extraction tasks. The SeRELeP system (System for Recognition of Relations for the Portuguese Language) (Brucksen et al. 2008) aimed at recognizing three relations: identity, inclusion and placement. The steps for identification/classification of NEs were carried out by 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.

These systems that participated in HAREM follow the conference directives. For example, the REMBRANDT, SEI-Geo and SeRELeP systems used ReRelEM's Golden Collection during the evaluation of the ReRelEM track. In general, the relations annotated by these systems were compared with those in the Golden Collection (Freitas et al. 2010), and each triple (NE1, Relation, NE2) was scored as correct, missing or incorrect. As results of the ReRelEM track, REMBRANDT system achieved the best results in the global scenario considering all relations, SEI-Geo got best scores for the inclusion relation and SeRELeP reached best results for the placement relation.

Besides the HAREM systems, other systems have been proposed. Batista et al. (2013) propose an approach of distantly supervised relation extraction between two entities. The authors selected 10 relation types in articles from Wikipedia written in Portuguese such as located-in, influenced-by, successor-of and others. In Ferreira et al. (2009), a system for information extraction from medical reports is presented. The authors reported a Golden Collection in the scope of the MedAlert project in which clinical documents relative to hospitalization episodes are annotated with its multiple entities and relations. For automatic extraction of entities and relations, the REMMA system (System for Recognition of Named Entities of MedAlert) was used. In Santos et al. (2010), a system that identifies family relations using rule-based approach is presented. Historical and biographic documents are texts that are rich in that kind of relation.

It is worth highlighting that for all Portuguese works presented here, the set of relations were previously defined (closed-domain RE). There are few RE systems

(Gamallo et al. 2012; Santos et al. 2012; Santos and Pinheiro 2015; Gamallo and García 2015) which apply the Open Information Extraction approach (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 Wikipedia in four languages: Portuguese, Spanish, Galician and English. Santos et al. (2012) present the News2Relations system for extracting relations from titles of news written in Portuguese. News2Relations deals with relations of the type (subject, verb, object), which are not specified in advance. 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. The RePort system used syntactic and lexical rules adapted for Portuguese with linguistic knowledge and lexicon of verbal relations extracted from a Portuguese corpus. Santos and Pinheiro also apply data mining to extract other relations already present but previously unknown. In Gamallo and García (2015), a multilingual rule-based Open IE system (ArgOE) is proposed. It is configured for English, Spanish, French, Galician and Portuguese.

Most of the relation extraction systems for Portuguese are based on rules and few external resources such as Wikipedia and domain ontology, and the set of relations are previously defined (closed RE). They usually do not use machine learning techniques contrary to the situation for English (Collovini de Abreu et al. 2013). Following, we present the relevant English RE systems for this work, which are the ones that apply Open IE approach using machine learning.

2.3 Open relation extraction for English

Open Information Extraction (Open IE), proposed by Banko et al. (2007), does not need a pre-specified definition of relation. In general, Open IE systems aim at extracting a large set of related triples (E1, Rel, E2) from a certain corpus without requiring human supervision, whereas E1 and E2 are strings meant to denote entities or noun phrases, and Rel is a string meant to denote a relation between E1 and E2.

The first Open IE Sytem was the TextRunner (Banko et al. 2007; Yates et al. 2007), which used a Naive Bayes classifier with POS and NP-chunk features. Banko and Etzioni (2008) present the O-CRF system based in a CRF. The authors show that many relations can be categorized using a compact set of lexicon-syntactic patterns. An approach to Open IE which uses Wikipedia as a source of training data is proposed in Wu and Weld (2010). The authors present the WOE system (Wikipedia-based Open Extractor), which generates relation specific training examples by matching between Wikipedia Infobox content with corresponding patterns. WOE can learn two types of extraction: WOEparse learned from dependency path patterns; and WOEpos trained using CRF model with shallow features like POS tags. Fader et al. (2011) presented ReVerb Open IE system, based on syntactic and lexical heuristics, which identifies verbs expressing relations in English.

3.0 RelP system

In this section, we present the RelP system, focusing in open relation extraction for Portuguese texts. RelP extracts any relation descriptors expressing an explicit relation occurring between pairs of named entities (Organisation, Person or Place). For this, initial steps of preprocessing of the texts are necessary: automatic tagging of the texts and NER. The method used to classify the relation descriptor is the probabilistic model CRF, considering the representation scheme (Collovini et al. 2015) and features (Collovini et al. 2014). The data sets used and the steps of the RelP system are presented following.

3.1 Data sets

In this section, we present the process of manual annotation of the data sets used in this work.

3.1.1 Manually annotated subset of HAREM

In this work, we used a subset of the Golden Collections from the two HAREM conferences. The annotation of the data was performed in two steps: we selected texts from these Golden Collections, and then we added the annotation of relations expressed between particular named entities contained in the selected texts.

We only analyzed texts dealing with the organisation domain, such as opinion, journalistic and political texts, among others. We added to the selected texts the annotation of the relation descriptors for each sentence. The manual annotation of the relation descriptors was performed by two linguists in the following way: given two named entities occurring in the same sentence, the text chunk (descriptor) that best describes an explicit relation between these entities is annotated.

We highlight the difficulty to determine which elements between the named entities are in fact part of the descriptor. Thus, the annotation of the relation descriptors was based on guidelines, which are briefly described below.

– The descriptor should be as concise as possible, being composed of the smallest number of elements required to describe it. In general, the elements of the descriptor were considered up to the preposition when

it occurred. In Table 1, example (3) shows where the descriptor "abre perspectivas em" ("opens perspectives in") is sufficient to express the relation between the named entities of the organisation and place categories.
– Nouns can express relations between named entities, such as nouns expressing titles/jobs (see example (4) in Table 1).
– Verbs are the predicates of sentences and generally describe the relations between their arguments (see example (5) in Table 1). In most cases, auxiliary verbs are not part of the descriptor between pairs of named entities, since the main verb is sufficient to express the relation between those entities. In Table 1 example (6) illustrates where the verb plus preposition "fundada em" ("founded in") is sufficient to express the relation between the named entities of the organisation and place categories.
– Prepositions can express relations between the named entities, such as affiliation relations between a person and an organisation (see example (7) in Table 1).
– The negation should be part of the relation descriptor (see example (8) in Table 1).
– There are elements which were not included in the descriptor, such as adjectives and pronouns. In Table 1, example (9) illustrates where the preposition *de* (of) is sufficient to express the relation between the named entities of the person and organisation categories.
– The adjective "em exercício" ("current") only qualifies the named entity "Presidente." In example (10), the possessive pronoun *sua* (its) is not a necessary element to compose the descriptor between the named entities of the organisation and place categories.

In the annotation process, not all sentences of the texts were considered, only those that contained a pair of named entities of interest. After the annotation of the relations between pairs of named entities by each annotator was done, there was a discussion of the annotations to check the consensus.

A sum of 516 relation instances was selected to compound the reference data set divided in four subsets (ORG-ORG, ORG-PERS, ORG-PLACE and ORG-PERS-PLACE), which were defined according to the categories of the pairs of named entities (see Table 2). The small number of instances is due to the difficulty in the manual annotation of the data.

The total number of relation instances and the number of positive and negative instances in each data set are

Positive relation instance	Relation descriptor
(3) A Marfinite abre perspectivas de negócios e geração de empregos, através de novos distribuidores em o Brasil. Marfinite opens business perspectives and job creation through new distributors in Brasil.	abre perspectivas em open perspectives in
(4) Steve Jobs, o director-geral de a 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
(5) Amílcar Cabral criou o Partido Africano. Amílcar Cabral created the Partido Africano.	criou created
(6) A Legião da Boa Vontade, instituição educacional, cultural e beneficiente, foi fundada em o Brasil. Legião da Boa Vontade, educational, cultural and beneficent institution, was founded in Brasil.	fundar em found in
(7) António Fontes de a AIPAN. António Fontes of the AIPAN.	de of
(8) Granada não enfrentou os Reis Católicos por razões de fé ou religião. Granada is not facing the Reis Católicos for reasons of faith or religion.	não enfrentou is not facing
(9) Também o Presidente em exercício de o Conselho de a UE. Also the current Presidente of the Conselho of the UE.	de of
(10) A Legião da Boa Vontade comemora amanhã o 10^{0}. aniversário da sua implantação em Portugal. Legião da Boa Vontade celebrates today the 10th birthday of its establishment in Portugal.	implantação em establishment in

Table 1. Examples of positive relation instances.

Data Sets	Total	Positive	Negative
ORG-ORG	175	90	85
ORG-PERS	171	95	76
ORG-PLACE	170	97	73
ORG-PERS-PLACE	516	282	234

Table 2. Number of instances of data sets.

summarized in Table 2. Positive instances are those that have an explicit relation descriptor between two named entities (see examples of the positive relation instances in Table 1) and negative instances are those that do not meet this condition. Examples of negative relation instances are presented in Table 3, where in example (11) there is not a relation between the entities named person ("Michael Bloomberg") and organization ("Partido Conservador"). The relation (someone deliver something) occurs between "Michael Bloomberg" and "o cartão de militante do Partido Conservador" ("the militant card of the Partido Conservador"). In example (12), there is no relation between the named entities place ("Turquia") and organization ("Pentágono").

Negative relation instance
(11) Michael Bloomberg decidiu entregar o cartão de militante do Partido Conservador. Michael Bloomberg decided to hand the militant card of the Partido Conservador.
(12) Os aparelhos regressaram à base na Turquia, acrescenta o comunicado do Pentágono. The equipment returned to the base in Turquia, added the statement to The Pentagon.

Table 3. Examples of the negative relation instances.

3.2 Pre-processing

The first step of pre-processing is automatic tagging of the data. For Portuguese, there are NLP tools for tagging texts, for example, the PALAVRAS parser (Bick 2000), the POS tagger available in OpenNLP library (http://opennlp.apache.org/), the annotation of POS based on Freeling Library (http://nlp.lsi.upc.edu/freeling/), among others.

In this work, the texts have been annotated with PALAVRAS, which provides Part-Of-Speech (POS) syntactic and semantic information. For example, this parser assigns semantic tags for most nouns like job/title, proper names (for example, semantic tags for place, company, event etc.), verbs (semantic tags indicating verb with human subject/inanimate subject) and some adjectives like semantic tags for color and nationality adjectives.

The example of output of the PALAVRAS parser in Constrain Grammar format is presented in Table 4, following the order: the word, the canonic form, semantic tag, POS tag and syntactic information of the sentence fragment, which is described in (1). For named entities that contain multiple words, the PALAVRAS parser concatenated them into a single word (for example, "Ronado _Lemos"), and preposition-article contraction is split ("da=de+a," "do=de+o"), for example, "diretor de a."

1)No próximo Sábado, Ronaldo Lemos, diretor da Creative Commons, irá participar de um debate [...]

Next Saturday, Ronaldo Lemos, director of Creative Commons, will participate in a debate [...]

The pre-processing step also involves the NER task for identifying the NE categories person, organisation and place; they were found (Collovini et al. 2011) to be the most relevant to the organisation domain.

In the literature, there are some NER systems that treat Portuguese, among which we can mention PALAVRAS-NER (Bick 2003; Bick 2006), REMBRANDT (Cardoso 2008), NERP-CRF (do Amaral and Vieira 2013), Freeling (http://nlp.lsi.upc.edu/freeling/) and Language Tasks (http://ltasks.com/), which can be used in this stage. However, in this work, the texts that we used already had the annotations of the NEs (HAREM's Golden Collection); thus, there was no need for applying an automatic NER system (see Section 3.1).

Word	Canonic form	Semantic	POS	Syntactic
[…]				
Ronaldo_Lemos	Ronaldo_Lemos	<hum>	PROP	@SUBJ>
,				
diretor	diretor	<Hprof>	N	@N<PRED
de	de		PRP	@N<
a	o		DET	@>N
Creative_Commons	Creative_Commons	<org>	PROP	@P<
[…]				

Table 4. Example of output of the PALAVRAS parser.

In this context, the next step is to identify in each sentence the pair of NEs in focus (organisation and person or organisation and placement, not necessarily in that order) that is located nearest to each other, based on the annotation of the HAREM's Golden Collection. However, we consider only one occurrence of named entity pair per sentence. Thus, when there were more than one pair of named entities, these sentences were duplicated considering the different pairs of named entities.

Returning to the example of output of the PALAVRAS parser described in Table 4, we have added a column with the person category (PERS) for "Ronaldo_Lemos" and Organisation category (ORG) for "Creative_Commons," as illustrated in Table 5. As a result of this step, each pair of named entities occurring in the same sentence is a candidate relation instance for the learning stage.

At the last stage, we generate a Pre-processing Vector for each annotated relation instance (positive or negative) with the following information of each word:

- word: return the canonic form of the word;
- syntacticTag: return syntactic tag of the word, or otherwise "null;"
- POSTag: return POS tag of the word, or otherwise "null;"
- semanticTag: return semantic tag of the word, or otherwise "null;"
- dictionary: return "true" if the word is contained in the external dictionary (list of person titles/jobs or places), or otherwise "false;"
- generateFeature: returns "true" if the features should be generated for the word in focus (parameters of the relation and for the words occurring between them), or otherwise "false;"
- NEcategory: return the NE category (PERS or ORG or PLACE) if the word is an NE, or otherwise "null."

This information is extracted from the relation instances in the format presented in Table 5. An illustration of the Pre-processing Vector corresponding to the fragment of the sentence described (1) is given in (13), which indicates in generateFeature: "true" that the vector of the features should be generated for each word.

1)[…] Ronaldo_Lemos, diretor de a Creative_Commons [...]

[…] Ronaldo_Lemos, director of Creative_Commons [...]

13)Pre-processing Vector:

[word: 'Ronaldo_Lemos', syntacticTag: '@SUBJ>', POSTag: 'PROP', semanticTag: 'hum', dicionary: 'false', generateFeature: 'true', category: 'PERS'],

[word: ',', syntacticTag: 'null', POSTag: 'null', semanticTag: 'null', dicionary: 'false', generateFeature: 'true', category: 'null'],

[word: 'diretor', syntacticTag: '@N<PRED', POSTag: 'N', semanticTag: 'Hprof', dicionary: 'true', generateFeature: 'true', category: 'null'],

[word: 'de', syntacticTag: '@N<', POSTag: 'PRP', semanticTag: 'null', dicionary: 'false', generateFeature: 'true', category: 'null'],

[word: 'o', syntacticTag: '@>N', POSTag: 'DET', semanticTag: 'null', dicionary: 'false', generateFeature: 'true', category: 'null'],

[word: 'Creative_Commons', syntacticTag: '@P<', POSTag: 'PROP', semanticTag: 'org', dicionary: 'false', generateFeature: 'true', category: 'ORG']

From the Pre-processing Vector, we can generate the features vectors for the CRF model, described in Section 3.3.1.

Word	Canonic form	Semantic	POS	Syntactic	NE
[…]					
Ronaldo_Lemos	Ronaldo_Lemos	<hum>	PROP	@SUBJ>	PERS
,					
diretor	diretor	<Hprof>	N	@N<PRED	
de	de		PRP	@N<	
a	o		DET	@>N	
Creative_Commons	Creative_Commons	<org>	PROP	@P<	ORG
[…]					

Table 5. Example added: the named entity categories.

3.3 Learning method

We apply a probabilistic method to extract relation descriptors between named entities in Portuguese texts. In this work, we use CRF models, which are undirected graphical models used to calculate the conditional probability of values on designated output nodes given values assigned to other designated input nodes (Lafferty et al. 2001).

Following the definition of linear-chain CRF, let 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 s = (s_1, s_2, ..., s_T) is the sequence of states corresponding to the T output nodes. We consider each word of a sentence as an observation o, which receives a L label according to an IO notation defined in previous work (Collovini et al. 2015). Thus, each word of a sentence is tagged as: {I-REL, O}, where a word labeled with I-REL is inside of the relation descriptor, while a word labeled with O is outside of the relation descriptor.

To exemplify the proposed IO notation, we present the sentence fragment in Table 6, in which we added the last column to illustrate the output labels. These labels represent the output nodes of the CRF model. The words "diretor de" ("director of") received the label I-REL. The PERS and ORG named entities ("Ronaldo_Lemos" and "Creative_Commons," respectively) receive the label O, because they are not part of the relation descriptor.

The IO vectors corresponding to every relation instance are part of the input to the CRF model. The other input are the features that describe the relation instances, which are presented in the next section.

3.3.1 Feature extraction

The next step is the generation of features, because the CRF needs vectors of attributes describing the features of the input data. We generate the features vectors for the NEs involved in a relation and for the words occurring between these entities in the sentence.

We use specific features for Portuguese described in previous work (Collovini et al. 2014). The sets of features are described in Table 7 and examples of features vectors containing some important features are presented in Table 8.

3.4 RelP output

The RelP system uses the IO vector as input along with feature vectors that represent each relation instance. The CRF model is generated from the features vectors; for every feature, it is attributed a certain weight, resulting in a weight matrix. Starting from the generated matrix, the CRF is capable of classifying correctly the words that indicate a relation in new texts not tagged yet. Thus, the RelP output consists in the tagged relation descriptors. An example is given in (14), corresponding to the sentence described in (1).

(14) RelP Output:
(Ronaldo_Lemos<O>, diretor<I-REL> de<I-REL>, Creative_Commons<O>)

4.0 Evaluation

We evaluated the classification using the measures: number of all positive relation instances correctly identified (#C), Recall (R), Precision (P) and F-measure (F). We applied 10-fold cross validation in all of the data sets, and we evaluated the performance using two criteria (Collovini et al. 2014): 1) exact matching, when the extracted relation descriptor is exactly the same as the one manually annotated; and, 2) partial matching, when the extracted relation descriptor has at least one word in common with the manual annotation but it does not match the entire descriptor.

An example of the evaluation criteria (15) is presented in Table 9, where the relation descriptor "comandar por" ("command by") is considered exact matching when both

Word	Canonic form	Semantic	POS	Syntactic	NE	IO
[…]						
Ronaldo_Lemos	Ronaldo_Lemos	<hum>	PROP	@SUBJ>	PERS	O
,						O
diretor	diretor	<Hprof>	N	@N<PRED		I-REL
de	de		PRP	@N<		I-REL
a	o		DET	@>N		O
Creative_ Commons	Creative_ Commons	<org>	PROP	@P<	ORG	O
[…]						

Table 6. Example added: the IO notation.

Feature Set	Explanation
Part-of-Speech	POS tags in a window of ±2 words 2 consecutive POS tags in a window of ±2 words
Lexical Item	canonic form in a window of ±2 words 2 consecutive canonic form in a window of ±2 words number of words in the segment
Syntactic	syntactic tags in a window of ±2 words 2 consecutive syntactic tags in a window of ±2 words head of the segment appositive in a window of ±2 words head of appositive direct object function
Patterns	a verb in a window of ±2 words a verb followed by a preposition or an article a noun followed by a preposition an adverb followed by a preposition or an article
Phrasal Sequence	POS tags of the word sequence between two NEs
Semantic	Semantic tag in a window of ±2 words NE category
Dictionary	list of Person titles/jobs, and Place words

Table 7. Features set.

Features	Ronaldo_Lemos	,	diretor	de [...]
POS (tag)	PROP	null	N	PRP
Lexical Item	Ronaldo_Lemos	,	diretor	De
Syntactic (tag)	@SUBJ>	null	N<PRED	@N<
Patterns (noun+preposition)	false	false	true	False
Phrasal Sequence (POS tag)	N PRP N	null	N PRP N	N PRP N
Semantic (NE category)	PERS	null	null	Null
Dictionary	false	false	true	False

Table 8. Vector with some features corresponding "Ronaldo_Lemos," "diretor de …."

Relation instance	Exact matching	Partial matching
(15) As forças da Escola de Cavalaria eram comandadas por o Salgueiro Maia. The forces of Escola de Cavalaria were commanded by Salgueiro Maia.	comandar <I-REL> por <I-REL> to command <I-REL> by <I-REL>	comandar <I-REL> por <O> to command <I-REL> by <O>

Table 9. Examples of the evaluated criteria for relation descriptors.

words that form it are annotated with I-REL label, and it is considered partial matching when at least one word receives the I-REL label, in this case the verb "comandar" ("to command").

4.1. Results and discussion

In this section, we present the results of the automatic extraction of relation instances, considering the following combinations of entity categories: ORG-ORG, ORG-PERS, ORG-PLACE, ORG-PERS-PLACE.

In Table 10, we illustrate the results of the RelP system for each data set, considering exact and partial matching. In general, ORG-PLACE archived the best rates for all measures compared to others data sets considering both exact and partial matching. This is due to the fact that ORG-PLACE has a greater number of correct labels and fewer number of false-positives. For partial matching, ORG-

PERS presented the best rates of recall, because this set has a greater number of correct instances (54 cases) compared to other sets. Finally, ORG-PERS-PLACE achieved a largest number of the correct instances (108 cases) compared to each data set considering exact matching. We can see in Table 10 that the best results were for partial matching; it occurred due to the difficulty of classifying all elements that compose a descriptor.

However, most instances evaluated as partial matching were able to represent the existing relations. For the 35 cases of partial matching in ORG-PERS-PLACE, 20 cases were able to represent the relation descriptors (57%). Examples of descriptors classified as partial matching are presented in Table 11 where both partial matching and reference are presented.

It is difficult to make a comparison with other works since the resources and data sets are different (Collovini de Abreu et al. 2013). In Table 12, we show the results achieved in other open RE systems for Portuguese: News2Relations (Santos et al. 2012), DepOE (Gamallo et al. 2012), ArgOE (Gamallo and García 2015) and RePort (Santos and Pinheiro 2015; Pires 2015). We can see that

Data Sets	Exact matching					Partial matching			
	#C	R	P	F		#C	R	P	F
ORG-ORG	19	0.21	0.28	0.24		42	0.46	0.63	0.53
ORG-PERS	36	0.37	0.44	0.40		54	0.56	0.66	0.61
ORG-PLACE	42	0.43	0.61	0.50		52	0.53	0.76	0.63
ORG-PERS-PLACE	108	0.38	0.51	0.43		143	0.50	0.67	0.58

Table 10. Results of the RelP system.

Relation instance	Partial matching	Reference
(16) Durão Barroso discursava na sessão plenária do Parlamento Europeu. Durão Barroso discoursed in plenary session of Parlamento Europeu.	discursar discourse	discursar em discourse in
(17) A Legião da Boa Vontade, instituição educacional, cultural e beneficiente, foi fundada no Brasil. Legião da Boa Vontade, educational, cultural and beneficent institution, was founded in Brasil.	em in	fundar em found in
(18) Este mesmo Governo preparou as eleições para uma Assembléia Nacional Popular. This same Government prepared the elections for an Assembléia Nacional Popular.	para for	preparar as eleições para prepare the elections for

Table 11. Examples of RelP output considering partial matching.

Systems	Performance
RelP (Exact matching)	ORG-PERS-LOCAL: P=0.51; F=0.43
RelP (Partial matching)	ORG-PERS-LOCAL: P=0.67; F=0.58
News2Relations (Santos et al. 2012)	Correct=0.80
DepOE (Gamallo et al. 2012)	--
ArgOE (Gamallo and García 2015)	P=0.53
RePort (Santos and Pinheiro 2015; Pires 2015)	P=0.52; F=0.46

Table 12. Results reported by open RE systems.

our results considering any relation descriptors between named entities (open RE) are not distant from these works.

5.0 Organizing the extracted relations

In this section, we present a way to organize the triples resulting from the extraction of open relations between named entities. In the open RE task there is a great diversity of relations, making it difficult to organize the relations. In the total 282 positive instances, there are 162 distinct relation descriptors. This diversity of relations is illustrated in Table 13 where we present triples extracted from data sets.

Our approach is based on mining the triples resulting from the open relation extraction, which considers the following concepts (Collovini et al. 2016b):

- Target: each triple (NE1, Relation Descriptor, NE2);
- Context: any part of the triple used for grouping of targets;
- Configuration: a set of triples associated with one context. Table 14 shows an overview of five possible configurations.

In Table 15, we show examples of configurations applied in the output of the model. Considering the NE "Brasil," in Config. (1), we can see seven mined triples, and in Config. (2) only six were found; that happens, because "Brasil" can be classified as place or as organisation, depending on the situation.

In Config. (1) to (4) presented in Table 15, the mined triples express every relation involving the NE in general or in a determined category. In the case of Config. (1), we cluster all relations involving the "Brasil."

Narrowing the context to the NE's categories, we were able to organize the triples by co-occurrence of these named entities. This makes it easier to classify the relations, identify similar relations as well as make and association between the entities involved (parameters of the relations). In the Config. (4), we can identify two relations expressing foundation ("fundar em/fundar por"): one related to a place ("Brasil") and another related to a person ("Alziro Zarur").

Finally, Config. (5) clusters all the named entities involved in a common relation, since named entities with common descriptors can be classified as similar. The descriptor "presidente de" relates persons with this job, and the affiliation relation between these person and different organisations.

We show here that the extracted relation descriptors can be better structured using different context. This organization can be useful to analysis and discovery of the patterns of the relations expressed between named entities from Portuguese texts.

6.0 Conclusions

This paper presented the RelP system for open relation extraction between named entities previously defined (organisation, person and place). We extract the relation

NE1	Relation Descriptor	NE2
Santos_Ferreira	ter sucesso em	BCP
Nauman_Barakat	vice-presidente de	Macquarie_Futures
Antonio_Ribeiro	em declaração a	Público
Escola_de_Pilotagem	dirigido por	Cmdt_João_Filho
Hospital_São_João	em	Porto_Alegre
Legião_da_Boa_Vontade	fundar em	Brasil
Ministério_da_Indústria_e_Energia	publicar em	Díario_da_República
Câmara_de_Matosinhos	investir em	Guifões
Serrambi_Viagens_e_Turismo	ser agência de viagem em	Pernambuco

Table 13. Diversity of relations.

Configuration	Context
Config. 1	NE
Config. 2	NE of Place category
Config. 3	NE of Person category
Config. 4	NE of Organisation category
Config. 5	Relation descriptor

Table 14. Configurations for different context (the target being the whole triple).

Configuration	Triples
(Config. 1) Context: NE Brasil	(Biblioteca_da_Real_Academia, seguir para, Brasil) (Serrambi, locação de automóvel em, Brasil) (Legião_da_Boa_Vontade, fundar em, Brasil) (Marfinite, abrir perspectiva em, Brasil) (FCI, em Brasil) (Creative_Commons, em, Brasil) (Brasil, manter sobre, Inglaterra)
(Config. 2) Context: NE Place Brasil	(Biblioteca_da_Real_Academia, seguir para, Brasil) (Serrambi, locação de automóvel em, Brasil) (Legião_da_Boa_Vontade, fundar em, Brasil) (Marfinite, abrir perspectiva em, Brasil) (FCI, em Brasil) (Creative_Commons, em, Brasil)
(Config. 3) Context: NE Person Santos_Ferreira	(Santos_Ferreira, saber de, Caixa) (Santos_Ferreira, ter sucesso em, BCP)
(Config. 4) Context: NE Organisation Legião_da_Boa_Vontade	(Legião_da_Boa_Vontade, implantação em, Portugal) (Legião_da_Boa_Vontade, fundar em, Brasil) (Legião_da_Boa_Vontade, em, Hora_da_Boa_Vontade) (Legião_da_Boa_Vontade, em, Rádio_Globo) (Legião_da_Boa_Vontade, fundar por, Alziro_Zarur)
(Config. 5) Context: Relation descriptor presidente de	(Rudy_Giuliani, presidente de, Câmera) (Almeida_Henriques, presidente de, Associação_do_Viseu) (Antônio_Nunes, presidente de, Autoridade_de_Segurança) (Fernando_Gomes, presidente de, Câmara_do_Porto) (Biblioteca_Nacional, presidente de, Pedro_Corrêa_do_Lago)

Table 15. Configurations examples.

descriptor that expresses an explicit relation between these named entities. In general, previously presented Portuguese systems do not use machine learning approaches, and the relations are specified in advance. We evaluated the RelP system considering exact and partial matching, regarding a reference corpus. We achieved best results with exact matching for relations between organisation and place entities. It occurred due to the fact that the relation descriptors contained fewer elements. Sometimes it is only the preposition, or else the verb plus preposition; for example, the descriptor "em" ("in") related the NEs "Fortis" and "Bruxelas," and the descriptor "publicar em" ("publish in") related the NEs "Ministério da Indústria" and "Diário da República."

Overall, the best results were achieved for partial matching, it occurred due to the difficulty of classifying all elements that compose a descriptor. Usually, descriptors that express relations between person and organisation entities are compound by many elements, such as the descriptor "a decisão de criar" ("the decision to create") that occurs between NEs "Sócrates" and "Instituto Europeu de Tecnologia." There are cases of relation descriptors between organizations formed by many elements like the descriptor "vai apoiar uma ação da" ("going to support an action of") that occurs between NEs "Rússia" and "Otan."

We also organized the triples resulting from the extraction of open relations between named entities, mining different configurations. The organization of relation descriptors can be useful to classify relation types, to cluster the entities involved in a common relation and to populate relational datasets, among other uses.

Since there are very few proposals for open relation extraction for Portuguese (Collovini de Abreu et al. 2013), contrary to the situation for other languages, the difficulty of the task is enhanced. This work contributed for the progress of Portuguese processing, that has a demand for the development of news methods, tools and specific resources such as annotated data. The produced resources related to this work will be made available to the community. This is one further step in the construction of technologies that help people finding relevant information for their specific needs. The studies that we are able to extract may serve to further purposes such as text summarization, question answering and ontology learning from texts. In future work, we plan to explore larger datasets: as well as to use open source resources, such as Cogroo (Silva 2013), an open source Brazilian Portuguese

grammar. We also look forward to performing an extension of the proposed process for other languages.

References

do Amaral, Daniela and Renata Vieira. 2013. "O Reconhecimento de Entidades Nomeadas por meio de Conditional Random Fields para a Língua Portuguesa." In *IXth Brazilian Symposium in Information and Human Language Technology (STIL 2013), Fortaleza, Ceara, Brazil, 21-23 October 2013*. Stroudsburg, Pa.: Assn. for Computational Linguistics, 59-68.

Banko, Michele, Michael J. Cafarella, Matt Broadhead and Oren Etzioni. 2007. "Open Information Extraction from the Web." In *International Joint Conference on Artificial Intelligence, IJCAI-07, January 06-12, 2007, Hyderabad, India*, ed. Rajeev Sangal, Harish Mehta and R. K. Bagga. San Francisco, CA: Morgan Kaufmann Publishers Inc., 2670-6.

Banko, Michele and Oren Etzioni. 2008. "The Tradeoffs between Open and Traditional Relation Extraction." In *The Association for Computer Linguistics, ACL* 2008, *June 15-18, 2008, Columbus, Ohio,* ed. K. McKeown, J. D. Moore, S. Teufel, J. Allan and S. Fururi. Stroudsburg, Pa.: Association for Computational Linguistics, 28-36.

Batista, David Soares, David Forte, Rui Silva, Bruno Martins and Mário Silva. 2013. "Extracção de Relações Semânticas de Textos em Português Explorando a DBpédia e a Wikipédia". *Linguamatica* 5: 41-57.

Bertrand-Gastaldy, S. 2001. Review of *Relationships in the Organization of Knowledge*, ed. Carol A. Bean and Rebecca Green. *Knowledge Organization* 28: 208-10.

Bick, Eckhard. 2006. "Functional Aspects in Portuguese NER." In *Computational Processing of the Portuguese Language: Proceeding of 10th International Workshop, PROPOR 2006, Itatiaia, Brazil, May 13-17, 2006,* ed. R. Vieira et al. Lecture Notes in Computer Science. Berlin: Springer, 80-9.

Bick, Eckhard. 2003. "Multi-level NER for Portuguese in a CG Framework." In *Computational Processing of the Portuguese Language: Proceeding of 6th International Workshop, PROPOR 2003, Faro, Portugal, June 26-27, 2003*, ed. Nuno J Mamede, Maria Das Gracas Volpe Nunes, Jorge Baptista and Isabel Trancoso. Lecture Notes in Computer Science 2721. Berlin: Springer, 118-25.

Bick, Eckhard. 2000. *The Parsing System Palavras. Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus: Aarhus University Press.

Brucksen, Mírian, José Guilherme Camargo Souza, Renata Vieira and Sandro Rigo. 2008. "Sistema SeRELeP para o reconhecimento de relações entre entidades mencionadas." In *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM,* ed. C. Mota and D. Santos. N.p.: Linguateca, 247-60.

Cardoso, Nuno. 2008. "REMBRANDT -- Reconhecimento de Entidades Mencionadas Baseado em Relações e ANálise Detalhada do Texto." In *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM,* ed. C. Mota and D. Santos. N.p.: Linguateca, 195-211.

Cardoso, Nuno. 2012. "Rembrandt - a named-entity recognition framework." In *LCREC 2012: Eighth International Conference on Language Resources and Evaluation*. Paris: European Language Resources Association (ELRA), 1240-3. http://www.lrec-conf.org/proceedings/lrec2012/index.html

Carvalho, Paula, Hugo Oliveira Gonçalo, Cristina Mota, Diana Santos and Cláudia Freitas. 2008. "Segundo HAREM: Modelo geral, novidades e avaliação." In *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM,* ed. C. Mota and D. Santos. N.p.: Linguateca, 11-31.

Chaves, Marcirio S. 2008. "Geo-ontologias e padrões para reconhecimento de locais e de suas relações em textos: o SEI-Geo no Segundo HAREM." In *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM,* ed. C. Mota and D. Santos. N.p.: Linguateca, 231-45.

Collovini, Sandra, Marcelo de Bairros P. Filho and Renata Vieira. 2015. "Analysing the Role of Representation Choices in Portuguese Relation Extraction." In *Conference and Labs of the Evaluation Forum, CLEF 2015, Toulouse, France*. Cham: Springer, 91-102.

Collovini, Sandra, Gabriel Machado and Renata Vieira. 2016a. "A Sequence Model Approach to Relation Extraction in Portuguese." In *LREC 2016: Tenth International Conference on Language Resources and Evaluation*, ed. Nicoletta Calzolari et al. Paris: European Language Resources Association.

Collovini, Sandra, Gabriel Machado and Renata Vieira. 2016b. "Extracting and Structuring Open Relations from Portuguese Text." In *Computational Processing of the Portuguese Language: 12th International Conference, PROPOR 2016, Tomar, Portugal, July 13-15, 2016,* ed. João Silva, Ricardo Ribeiro, Paulo Quaresma, André Adami and António Branco. Lecture Notes in Computer Science 9727. Cham: Springer International Publishing, 153-64.

Collovini, Sandra, Lucas Pugens, Aline A. Vanin and Renata Vieira. 2014. "Extraction of Relation Descriptors for Portuguese Using Conditional Random Fields." In *Proceedings of Advances in Artificial Intelligence - IBERAMIA 2014 - 14th Ibero-American Conference on AI, Santiago de Chile, Chile, November 24-27, 2014,* ed. Ana L. C. Bazzan and Karim Pichara. Lecture Notes in Com-

puter Science 8864. Cham: Springer International Publishing, 108-19.

Collovini, Sandra, Fernando Grando, Marlo Souza, Larissa Freitas and Renata Vieira. 2011. "Semantic Relations Extraction in the Organization Domain." In *Proceedings of the IADIS International Conference on Applied Computing 2011, Rio de Janeiro, Brazil 6-8 November 2011,* ed. Hans Weghorn, Leonardo Azevedo and Pedro Isaías. N.p.: AIDIS, 99-106.

Collovini de Abreu, Sandra, Tiago Luis Bonamigo and Renata Vieira. 2013. "A Review on Relation Extraction with an Eye on Portuguese." *Journal of the Brazilian Computer Society* 19: 116. doi:10.1007/s13173-013-0116-8

Dânger, Roxana, Ferran Pla, Antonio Molina and Paolo Rosso. 2014. "Towards a Protein-Protein Interaction Information Extraction System: Recognizing Named Entities." *Knowledge-Based Systems* 57: 104-18.

Culotta, Aron, Andrew McCallum and Jonathan Betz. 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, June 4-9 New York City*. Stroudsburg, PA: Association for Computational Linguistics, 296-303.

Fader, Anthony, Stephen Soderland and Oren Etzioni. 2011. "Identifying Relations for Open Information Extraction." In *Conference on Empirical Methods on Natural Language Processing, Edinburgh, Scotland, UK, July 27–31, 2011*. Stroudsburg, Pa.: Association for Computational Linguistics, 1535-45.

Ferreira, Liliana, César Oliveira, Antônio Teixeira and João Cunha. 2009. "Extração de informação de relatórios medicos." *Linguamatica*, 1: 89-101.

Freitas, Cláudia, Cristina Mota, Diana Santos, Hugo Gonçalo Oliveira and Paula Carvalho. 2010. "Second HAREM: Advancing the State of the Art of Named Entity Recognition in Portuguese." In *LREC 2010: Seventh International Conference on Language Resources and Evaluation,* ed. Nicoletta Calzolari et al. Paris: European Language Resources Association.

Freitas, Cláudia, Diana Santos, Hugo Gonçalo Oliveira, Paula Carvalho and Cristina Mota. 2008. "Relações semânticas do ReRelEM: além das entidades no Segundo HAREM." In *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM,* ed. C. Mota and D. Santos. N.p.: Linguateca, 75-94.

Gamallo, Pablo, Marcos Garcia and Santiago Fernández-Lanza. 2012. "Dependency-Based Open Information Extraction." In *Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in Natural Language Processing.* Avignon, France: Association for Computational Linguistics, 10-8.

Gamallo, Pablo and Marcos García. 2015. "Multilingual Open Information Extraction." In *Progress in Artificial Intelligence: 17th Portuguese Conference on Artificial Intelligence, EPIA 2015, Coimbra, Portugal, September 8-11, 2015,* ed. Francisco Pereira, Penousal Machado, Ernesto Costa and Amílcar Cardoso. Lecture Notes in Computer Science 9273. Cham: Springer International Publishing, 711-22.

Green, Rebecca. 2001. "Relationships in the Organization of Knowledge: An Overview." In *Relationships in the Organization of Knowledge*, ed. C. A. Bean and R. Green. Dordrecht: Kluwer Academic Publishers, 3-18.

Guarino, Nicola. 1995. "Formal Ontology, Conceptual Analysis and Knowledge Representation." *International Journal of Human-Computer Studies* 43: 625-40.

Jurafsky, Daniel and James H. Martin. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition,* 2nd ed. London: Pearson Education Ltd.

Khoo, Christopher S. G. and Jin-Cheon Na. 2006. "Semantic Relations in Information Science." In *Annual Review of Information Science and Technology* 40: 157-228.

Lafferty, John D., Andrew McCallum and Fernando C. N. Pereira. 2001. "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data." In *Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01*. San Francisco, CA: Morgan Kaufmann Publishers Inc., 282-9.

Li, Yaliang, Jing Jiang, Hai Leong Chieu and Kian Ming A. Chai. 2011. "Extracting Relation Descriptors with Conditional Random Fields." In *Proceedings of 5th International Joint Conference on Natural Language Processing*. Chiang Mai, Thailand: Asian Federation of Natural Language Processing, 392-400.

McCallum, Andrew and Wei Li. 2003. "Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons." In *CONLL '03 Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003,* ed. Walter Daelemans and Miles Osborne. Morristown, N.J.: Association for Computational Linguistics, 4: 188-91.

MUC-7. 1997. "Coreference Task Definition." In *Proceedings of the Seventh Message Understanding Conference - MUC-7, Fairfax, Virginia, April 29 - May 1, 1998*. http://www-nlpir.nist.gov/related_projects/muc/proceedings/co_task.html

Pires, Julio Cesar Batista. 2015. "Extração e mineração de informação independente de domínios da web na língua portuguesa." Masters thesis. Universidade Federal de Goiás, Goiânia.

Santos, Diana and Nuno Cardoso. 2007. "Breve introdução ao HAREM." In *Reconhecimento de entidades men-*

cionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área, ed. D. Santos and N. Cardoso. N.p.: Linguateca, 1-16.

Santos, Daniel, Nuno Mamade and Jorge Batista. 2010. "Extraction of Family Relations between Entities." In *Proceedings of the INForum 2010 - II Simpósio de Informática, 9-10 Setembro, 2010,* ed. Luís S. Barbosa and Miguel P. Correia. Braga, Portugal, 549-60.

Santos, Victor and Vladia Pinheiro. 2015. "RePort? Um Sistema de Extração de Informações Aberta para Língua Portuguesa." In *2015 Brazilian Conference on Intelligent Systems BRACIS 2015, 4-7 November 2015, Natal, RN, Brazil.* N.p.: Sociedade Brasileira de Computaçao, 191-200.

Santos, António Paulo, Carlos Ramos and Nuno C. Marques. 2012. "Extração de Relações em Títulos de Notícias Desportivas." In *INFORUM 2012, Simpósio de Informática*, *Lisbon, Portugal.* [Almada, Portugal: Universidade Nova de Lisboa]. http://inforum.org.pt/INForum 2012/programa

Santos, Viviane Neves dos and Kobashi, Nair Yumiko. 2013. "Reflexões sobre processamento e representação automática de conhecimento." In *Complexidade e organização do conhecimento: desafios de nosso século.* Rio de Janeiro: ISKO-Brasil; Marília: FUNDEPE, 183-8.

Sarawagi, Sunita. 2008. "Information Extraction." *Foundations and Trends in Databases* 1: 261-377.

Silva, William Daniel Colen M. 2013. "Aprimorando o corretor gramatical CoGrOO." Masters thesis. Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo.

da Silva, Wagner Teixeira and Ruy Luiz Milidiú. 1991. "Information Indexing and Retrieval with a Belief Function Model." In *Ciência da Informação* 20: 155-64.

Wu, Fei and Daniel S. Weld. 2010. "Open Information Extraction Using Wikipedia." In *ACL '10 Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics Uppsala, Sweden, July 11-16, 2010,* ed. Jan Hajič. Stroudsburg, PA: Association for Computational Linguistics, 118-27.

Yates, Alexander, Michele Banko, Matthew Broadhead, Michael J. Cafarella, Oren Etzioni and Stephene Soderland. 2007. "TextRunner: Open Information Extraction on the Web." In *NAACL-Demonstrations '07 Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, *Rochester, New York, April 23-25, 2007.* Stroudsburg, PA: Association for Computational Linguistics, 25-6.

Zhang, Chunyun, Weiran Xu, Zhanyu Ma, Sheng Gao, Oun Li and Jun Guo. 2015. "Construction of Semantic Bootstrapping Models for Relation Extraction." *Knowledge-Based Systems* 83: 128-37.