Portuguese Relation Extraction in the Organization Domain

Sandra Collovini, Lucas Pugens, Marcelo de Bairros Pereira Filho, Renata Vieira

LogOnto 2014

PUCRS

FACIN PUCRS

Overview

- Introduction
- Previous Work
- Reference Corpus
- Proposed Process
- Evaluation
- Conclusion
- Reference

Introduction

- Relation Extraction (RE)
  - One of the main challenges in Information Extraction (IE)
  - Aims at identifying (and classifying) semantic relations that occur between entities in text
    - Mário Vaz, **diretor da** Central Globo de Qualidade

diretor de a

Person

relation descriptor

Organization

Introduction

- We propose a process for the extraction of **relation descriptors** between Named Entities (NEs) for Portuguese using Conditional Random Fields (CRF)

Introduction

- Organization Domain
  - Potential applicability to different areas
    - Competitive Intelligence, Risk Management, Sales and Marketing
  - Ontology-based NLP tasks
    - Question Answering, Geo-Reference Systems, Information Retrieval and Extraction

Previous Work

- RE with CRF (Abreu et al., 2013)

Works	Corpora	Relation Type
Banko and Etzione (2010)	500 sentences from an IE training corpus	open relation and specifics: acquisition, birthplace, inventorOf, wonAward
Chen et al. (2010)	713 documents - 4 courses of computer science from the Web	preorder, illustration, analogy, no-relation
Li et al. (2011)	150 business articles from NYT and Wikipedia	employment, personal/social
Ling and Weld (2012)	1.8 million news articles from NYT (1987 to 2007)	36 relations

Previous Work

- RE works for **Portuguese** (Abreu et al., 2013)

Systems/Works	Corpora	Method	Relation Type
SeRELeP (Brucksen et al., 2008)	HAREM/ReRelEM	morphosyntactic and semantic rules	inclusion, identity, location
REMBRANDT (Cardoso, 2008)	HAREM/ReRelEM	Wikipedia and grammar rules	inclusion, identity, location, other
SEI-Geo (Chaves, 2008)	HAREM/ReRelEM	rules (patterns) and geo-ontologies	inclusion
Xavier and Lima (2010)	Wikipedia (tourism domain)	Wikipedia and syntactic rules	located-in, is-a
Batista et al. (2013)	DBPedia	distant supervision and k-Nearest-Neighbors	place-funeral, partner, influenced-by, origin-of, part-of, ancestor-of, successor-of, located-in, person-key-in, other

Reference Corpus

- HAREM's Golden Collections[1] for Named Entity Recognition (NER)
  - Manual annotation of the relation descriptors
  - 516 relation instances

Data set	Total	Positive	Negative
ORG-ORG	175	90	85
ORG-PERS	171	105	66
ORG-LOCAL	170	109	61
ORG-PERS-LOCAL	**516**	304	212

[1] http://www.linguateca.pt/

Examples

- Examples of positive relation instances

Data set	Relation Instances	Relation Descriptor	Relation Type
ORG-ORG	Confederação Brasileira de Cinofilia, **órgão filiado ao** FCI	**ógão filiado ao**	affiliation
ORG-LOCAL	Ronaldo Lemos, diretor do Creative Commons **no** Brasil	**em o**	location
ORG-PERS	Mário Vaz, **diretor da** Central Globo de Qualidade	**diretor da**	director-of

Examples

- Examples of positive relation instances

Data set	Relation Instances	Relation Descriptor
ORG-ORG	… da Biblioteca Houghton que **guarda as obras raras de** Harvard	**guarda as obras raras de**
ORG-ORG	A Resistência Islâmica, **ala armada do** Hizbollard	**ala armada do**
ORG-PERS	… Rudy Giuliani, o republicano que já **foi presidente da** Câmara	**foi presidente da**
ORG-PERS	Amílcar Cabral **criou o** Partido Africano …	**criou o**
ORG-LOCAL	… Biblioteca da Real Academia dos Guardas-Marinhas, que **seguiu com a côrte para o** Brasil	**seguiu com a côrte para**
ORG-LOCAL	Goa Tourism Development Corporation Office **organiza excursões a** Goa …	**organiza excursões a**

Examples

- Examples of negative relation instances

Data set	Relation Instances
ORG-ORG	…em consequência da reestruturação orgânica operada na Marinha passou a integrar o Arquivo Central da Marinha
ORG-LOCAL	… embaixador de Portugal em Espanha

Pre-processing

- Parser Palavras (Bick, 2000)

- Mário Vaz, **diretor da** Central Globo de Qualidade

```
Mario=Vaz [Mario=Vaz] <hum> PROP @SUBJ>
,
diretor [diretor] <Hprof> N @N<PRED
de [de] PRP @N<
a [o] DET @>N
Central=Globo=de=Qualidade [Central=Globo=de=Qualidade] <org>
PROP @P<
```

Named Entities

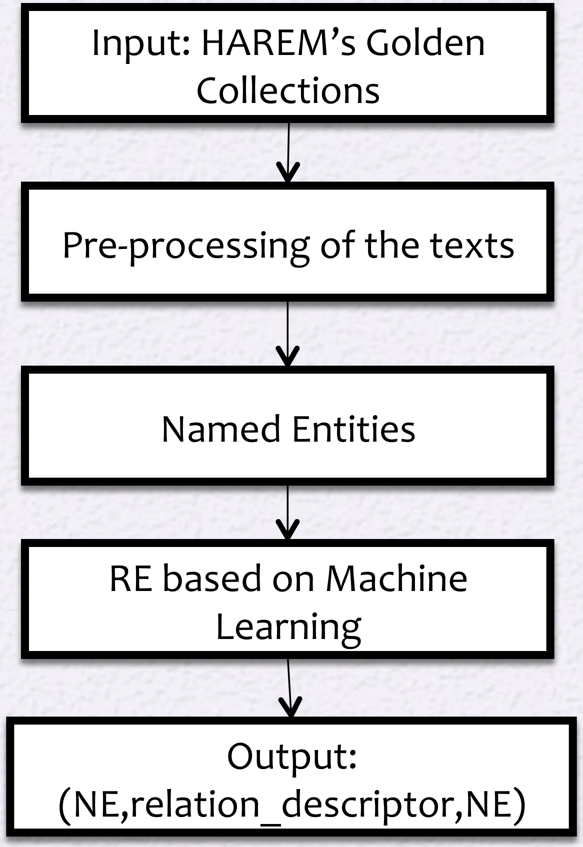
- HAREM’s Golden Collections for NER

- Mário Vaz, **diretor da** Central Globo de Qualidade

```
<EM ID=“ric-13” CATEG="PESSOA” >Mário Vaz<EM>,  diretor da
<EM ID=“ric-14” CATEG="ORGANIZACAO”>Central Globo de Qualidade<EM>
```

```
Mario=Vaz  [Mario=Vaz]  <hum>  PROP  @SUBJ> PERS
,
diretor  [diretor]  <Hprof>  N  @N<PRED
de  [de]  PRP  @N<
a  [o]  DET  @>N
Central=Globo=de=Qualidade [Central=Globo=de=Qualidade]  <org>
PROP  @P<  ORG
```

Proposed Process


Input: HAREM’s Golden Collections
Pre-processing of the texts
Named Entities
RE based on Machine Learning
Output: (NE,relation_descriptor,NE)

Mário Vaz, diretor da Central Globo de Qualidade

Mario=Vaz [Mario=Vaz] <hum> PROP @SUBJ>
diretor [diretor] <Hprof> N @N<PRED
de [de] PRP @N<
a [o] DET @>N
Central=Globo=de=Qualidade [Central=Globo=de=Qualidade] <org>
PROP @P<

Mario=Vaz **<PROP, PERS>**
Central=Globo=de=Qualidade **<PROP, ORG>**

Features: from the annotations above

(Mario=Vaz, **diretor de o**, Central=Globo=de=Qualidade)

Relation Descriptor

- Manual annotation of the relation descriptors between NEs

Mário Vaz, diretor da Central Globo de Qualidade

```
     Mario_Vaz  ,   diretor   da     Central_Globo_de_Qualidade
  [     O       O   REL       REL               O                ]
```

Features

1. **POS**: POS tags in a window of +-2
2. **Lexical :** canonic form in a window of +-2
3. **Syntactic features:** syntactic tags (appositive; direct object; etc.)
4. **Patterns features:** patterns such as a verb followed by a preposition or an article
5. **Phrasal Sequence features**: POS tags of the word sequence between two NEs
6. **Semantic features**: semantic tags provided by parser Palavras and NE category

Features

- Mário Vaz, diretor de o Central Globo de Qualidade

- **Mario=Vaz**
  - **POS**: ‘null’, ‘null’, **‘PROP’**, ‘,’ , ‘N’
  - **Lexical:** ‘null’, ‘null’, **‘Mario=Vaz’**, ‘,’ , ‘diretor’
  - **Syntactic features:** tag= ‘SUBJ’, head: 'sim', directObj: 'nao’, ...
  - **Patterns features:** adv: 'nao', verb: ‘nao’, verbDet: 'nao', ...
  - **Phrasal Sequence features**: ‘N PRP DET’
  - **Semantic features**: semantic: ‘hum’, category: ‘PERS’

- **Feature vector:**

[‘null’, ‘null’, ‘PROP’, ‘,’, ‘N’, ‘null’, ‘null’, ‘Mario=Vaz’, ‘,’, ‘diretor’, tag= ‘SUBJ’, head: 'sim', directObj: 'nao', adv: 'nao', verb: ‘nao’, verbDet: 'nao', 'PROP , N PRP DET PROP’ , ‘N PRP DET’, semantic: ‘hum’, category: ‘PERS’, ...]

Conditional Random Fields

- Conditional Random Fields (CRFs) are used to calculate the conditional probability of the outputs given the inputs (Lafferty et al., 2001)

  - Input: REL-O and features vector
  - Output: classification of relation descriptors

Conditional Random Fields

- Mário Vaz, diretor da Central Globo de Qualidade
  - Input:
    - O-REL: [O, O, REL, REL, REL, O]
    - Features vector: [‘null’, ‘null’, ‘PROP’, ‘,’, ‘N’, ‘null’, ‘null’, ‘Mario=Vaz’, ‘,’, ‘diretor’, tag= ‘SUBJ’, head: 'sim', directObj: 'nao', adv: 'nao', verb: ‘nao’, verbDet: 'nao', 'PROP , N PRP DET PROP’ , ‘N PRP DET’, semantic: ‘hum’, category: ‘PERS’, ….]
  - Output:
    - Mario=Vaz ,<**O**> diretor<**REL**> de<**REL**> o<**REL**> Central=Globo=de=Qualidade

Evaluation

- Method: Cross-validation
- Reference: Manual annotation of the relations
- Two criteria:
  - Complete descriptor matching
  - Partial descriptor matching

Relation Instances	Complete descriptor matching	Partial descriptor matching
PSD passa entre as sombras, ou, em muitos casos, **concordando com o** Governo	**concordar**<REL> **com**<REL> **o**<REL>	**concordar**<REL> **com**<O> **o**<O>

Results

ORG-PERS (10-folds)	Complete descriptor matching				Partial descriptor matching			
	#C	**A**	**P**	**F**	**# C**	**A**	**P**	**F**
F1=POS	31	0.29	0.41	0.34	50	0.47	0.67	0.56
F2=POS+LEX	37	0.35	0.58	0.44	47	0.44	0.74	0.55
F3=POS+LEX+SYN	43	0.40	0.62	0.49	53	0.50	0.76	0.60
F4=POS+LEX+SYN+PAT	42	0.40	0.61	0.48	52	0.49	0.76	0.60
F5=POS+LEX+SYN+PAT+PS	40	0.38	0.63	0.47	49	0.46	0.77	0.57
F6=POS+LEX+SYN+PAT+PS+SEM	**44**	**0.41**	**0.65**	**0.51**	**53**	**0.50**	**0.79**	**0.61**

- Different sets of features for CRF were evaluated
- **Semantic feature** based on **NE category** improved the relation extraction

Results

- Results for each data set

Data set (10-folds)	**Complete descriptor matching**				**Partial descriptor matching**			
	#C	**A**	**P**	**F**	**#C**	**A**	**P**	**F**
ORG-ORG	21	0.23	0.44	0.30	34	0.37	0.71	0.48
ORG-PERS	44	0.41	0.65	**0.51**	53	0.50	0.79	**0.61**
ORG-LOCAL	40	0.38	0.68	0.49	45	0.43	0.77	0.55
ORG-PERS-LOCAL	113	0.37	0.63	**0.46**	133	0.44	0.74	**0.55**

Error Analysis

- Few cases of false-positives
  - Most of the errors were the identification of verbal relation descriptors that do not express an explicit relation between pairs of Organizations

Relation Instances	Output	Reference
Almeida Henriques, presidente da Associação Industrial do Viseu, **é** o novo rosto do Conselho. Almeida Henriques, presidente da ~~Associação Industrial do Viseu~~, **é** o novo rosto do Conselho.	**ser**<B-REL>	**ser**<O>

Error Analysis

- Most false negative examples occur in cases where there are elements interposed between the NE

Relation Instances	Output	Reference
A Legião da Boa Vontade, instituição educacional, cultural e beneficiente, **foi fundada no** Brasil	**ser**<O> **fundar**<O> **em**<O> **o**<O>	**ser**<REL> **fundar**<REL> **em**<REL> **o**<REL>

Conclusion

- Open Relation Extraction (ORG-PER-LOC) for Portuguese
- Seven sets of features for CRF were evaluated
- The **semantic feature** based on the NE category provided us relevant information for the extraction of relation descriptors

Conclusion

- Results of the RE works for Portuguese

System/Works	Corpora	Results, %
SeRELeP (Brucksen et al., 2008)	HAREM/ReRelEM	3 Harem relations: F = 36%
REMBRANDT (Cardoso, 2008)	HAREM/ReRelEM	4 Harem relations: F= 45%
SEI-Geo (Chaves, 2008)	HAREM/ReRelEM	1 Harem relation: F= 44%
Batista et al. (2013)	DBPedia: 97.988 sentences	10 relations: F= 55.6%
Proposed Process (Abreu, 2014)	subset from HAREM	Open for ORG-PERS-LOCAL: complete matching: F= 46% partial matching: F= 55%

Future Work

- Specific relations (HAREM/ReRelEM)
- Consider descriptors not only between NEs, but also before and after
- Realize an extension of the proposed process for other languages
- More robust corpora
- Ontology population

Publications

- Sandra Collovini de Abreu, Tiago L. Bonamigo, and Renata Vieira. **A review on relation extraction with an eye on portuguese**. Journal of the Brazilian Computer Society, pages 1–19, 2013.

- Sandra Collovini de Abreu. **Extração de Relações do domínio de Organizações para o Português**. Tese de Doutorado, Faculdade de Informática, PUCRS, 112 p., 2014.

- Sandra Collovini, Lucas Pugens, Aline A. Vanin, and Renata Vieira. **Extraction of Relation Descriptors for Portuguese using Conditional Random Fields**. In: 4th edition of the Ibero-American Conference on Artificial Intelligence - IBERAMIA 2014, Santiago, Chile, 2014.

References

- Banko, M., Etzioni, O.: The tradeoffs between open and traditional relation extraction. In: McKeown, K., Moore, J. D., Teufel, S., Allan, J., Furui, S. (eds) ACL. The Association for Computer, Linguistics, Bulgaria, pp 28–36 (2010)

- Chen, Y., Zheng, Q., Wang, W., Chen, Y.: Knowledge element relation extraction using conditional random fields. In: CSCWD, pp 245–250 (2010)

- Ling, X., Weld, D. S.: Fine-grained entity recognition. In: Proceeding of the Twenty-Sixty AAAI Conference on Artificial Intelligence, AAAI, Toronto, Ontario, Canada (2012)

- Li, Y., Jiang, J., Chieu, H. L., Chai, K. M. A.: Extracting relation descriptors with conditional random fields. In: Proceedings of 5th international joint conference on natural language processing, Chiang Mai, pp 392–400 (2011)

- Bruckshen, M., Souza, J. G. C., Vieira, R., Rigo, S.: Sistema SeRELeP para o Reconhecimento de Relações entre Entidades Mencionadas. In: Mota, C., Santos, D. (eds) Segundo HAREM, Chap 14. Linguateca, pp 247–260 (2008)

- Cardoso, N.: REMBRANDT — Reconhecimento de Entidades Mencionadas Baseado em Relações e ANálise Detalhada do Texto. In: Mota, C., Santos, D. (eds) Segundo HAREM, Chap 11. Linguateca, pp 195–211 (2008)

- Chaves, M. S.: Geo-ontologias e padrões para reconhecimento de locais e de suas relações em textos: o SEI-Geo no Segundo HAREM. In: Mota, C., Santos, D. (eds) Segundo HAREM, Chap 13. Linguateca, pp 231–245 (2008)

References

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

- Xavier, C. C., de Lima, V. L. S.: A semi-automatic method for domain ontology extraction from Portuguese language wikipedia's categories. In: SBIA. pp. 11–20 (2010)

- Bick, E.: The parsing system PALAVRAS. In: Automatic grammatical analysis of Portuguese in a constraint grammar frame- work. University of Arhus, Arhus (2000)

- Lafferty, J. D., McCallum, A., Pereira, F. C. N.: Conditional Random Fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the ICML '01, Morgan Kaufmann, San Francisco, pp. 282–289 (2001)

- Borges, K.A.V.: Uso de uma Ontologia de Lugar Urbano para Reconhecimento e Extração de Evidências Geoespaciais na Web. Diploma thesis, Universidade Federal de Minas Gerais, UFMG (2006)

- Delboni, T. M.: Expressões de Posicionamento como Fonte de Contexto Geográfico na Web. Diploma thesis, Universidade Federal de Minas Gerais, UFMG (2005)

- Chaves, M.S., Silva, M.J., Martins, B.: A geographic knowledge base for semantic web applications. In: Heuser, C.A. (ed.) 20th Brazilian Symposium on Databases. pp. 40–54 (2005)