Language resources for information extraction and semantic computing - NLP at PUCRS

Renata Vieira[1], Daniela do Amaral[1], Sandra Collovini[1], Evandro Fonseca[1], Artur Freitas[1], Larissa Freitas[1], Roger Granada[1], Lucas Hilgert[1], Lucelene Lopes[1], Daniela Schmidt[1], Bernardo Severo[1], Marlo Souza[1], Cassia Trojahn[2]

[1] PUCRS - Porto Alegre, Brazil
{renata.vieira,lucelene.lopes}@pucrs.br
{daniela.amaral,sandra.abreu,evandro.fonseca,artur.freitas,
larissa.freitas,roger.granada,lucas.hilgert,daniela.schmidt,
bernardo.severo,marlo.souza}@acad.pucrs.br
[2] UT2J & IRIT - Toulouse, France
cassia.trojahn@irit.fr

Abstract. In this paper we present an overview of the language resources developed at the Natural Language Processing Lab at PUCRS, making them available to the research community.

Keywords: Information extraction, semantic computing, language resources

1 Introduction

At PUCRS (Pontifícia Universidade Católica do Rio Grande do Sul) NLP Lab we investigate several semantic information extraction related problems, for which we have used and developed a series of corpora, annotations and tools. Our main themes are named entity recognition; terms, concepts, taxonomies and open relation extraction; coreference resolution; sentiment analysis; ontology development and alignment. The currently available resources, developed by our team over the years, related to these research topics are described in this paper.

2 Named Entity Recognition

Named Entity Recognition (NER) consists of the identification and classification of linguistic expressions identification and classification, mostly proper nouns that refer to a specific entity in the text. In general, a NER task is divided into two phases: Named Entities (NEs) identification and NEs classification. NER main challenges in the entities recognition process are the NEs delimitation during the identification phase and the ambiguity of words in the classification phase.

To deal with this task we developed the Named Entities Recognition Portuguese-Conditional Random Fields (NERP-CRF) system [1], its first version was based

on the HAREM corpus[3] and categories. More recently we are investigating geological NEs. We built a reference corpus that contains NEs considering geological classes. The corpus is formed by scientific papers and articles, thesis, and dissertations found in digital libraries. We identified eleven geological classes according to three groups: Habitat of Microfossils, Age of Rocks, and Types of Rocks. The corpus consists of 70,191 words and 3,687 geological NEs checked by a specialist, and is available at `http://www.inf.pucrs.br/linatural/NER.html`.

3 Term Extraction

Term extraction from corpora is the cornerstone of several NLP applications. A particularly interesting application of extracted terms have been developed recently to establish entity profiles [2]. An example of such profiling is available at `http://www.inf.pucrs.br/peg/lucelenelopes/profiler_PPGCC/index.html`.

Our approaches for term extraction rely on both linguistic and statistic-based techniques. The linguistic-based techniques are centered on the recognition of noun phrases from a parser annotation and a set of heuristics to increase the quality of extracted terms [3]. The statistic-based techniques intervene with the use of an index to establish the extracted term relevance, the term frequency-disjoint corpora frequency (*tf-dcf*) index [4], and, finally, with the application of cut-off policies [5]. ExATO software tool [6] implements these term extraction techniques [7]. The current version of ExATO is capable of dealing with English and Portuguese corpora in several formats of output: a concordancer, tag clouds and concept hierarchies.

To exercise our tools and techniques several domain corpora were created [8] and acquired. The corpora created are avalable at `http://www.inf.pucrs.br/peg/lucelenelopes/ll_crp.html` and lists of the extracted terms are available at `http://www.inf.pucrs.br/peg/lucelenelopes/ll_trm.html`. Additionally, an experiment with English corpora was conducted to illustrate the impact of contrasting corpora choices in our term extraction method [9].

In [10] we proposed a method to build bilingual dictionaries for specific domains from parallel corpora. An evaluation was performed on technical manuals in English and Portuguese. The bilingual dictionaries created from the application of this method are available in `http://www.inf.pucrs.br/~linatural/multilingual/`.

4 Semantic Relation Identification

Semantic similarity could be viewed as an association of two terms, that is, the mental activation of one term when another term is presented. This idea was expressed by Zellig Harris [11] when he formulated the hypothesis that words that occur in the same contexts tend to have similar meanings. Models built on this assumption are called Distributional Similarity Models (DSMs) and take

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

into account the co-occurrence distributions of the words in order to cluster them together [12]. In [13], we perform an evaluation on methods that use different co-occurrence orders to get similarity between terms.

In order to evaluate such methods it is also important to have datasets manually evaluated by domain experts. An important resource for evaluation in English has been defined by Rubenstein and Goodenough [14]. This dataset (RG65) contains judgements scaled from 0 to 4 according to their similarity of meaning from 51 human subjects for 65 word pairs. Following the work by Rubenstein and Goodenough, we translated into Portuguese all pairs from RG65 and evaluate them using 50 human subjects (Granada *et al.* [15]). These lists are available at `http://www.inf.pucrs.br/linatural/wikimodels/similarity.html`. Human scores are compared with previous works and an automatic evaluation is performed by comparing with models generated from Wikipedia articles.

5 Taxonomic Relations Extraction

Many methods have been proposed to extract taxonomic relations from texts. Hearst [16] proposed the extraction of taxonomic relations from texts by using lexico-syntactic patterns in the form of regular expressions, Radford [17] identifies the taxonomic relation between terms using the head of the noun phrase, since it determines the nature of the overall phrase. Using a statistical approach Caraballo [18] uses hierarchical clustering in order to identify hierarchical relations. Weeds *et al.* [19] identify relations based on their distributional inclusion, *i.e.*, two words have taxonomic relation if both share a great number of contexts. Sanderson and Croft [20] present the document subsumption method which identify taxonomic relation based on the probabilities of term co-occurrences in documents. Santus *et al.* [21] used an entropy-based measure for the unsupervised identification of taxonomic relations in DSMs.

We developed the HREx framework (`https://github.com/rogergranada/HREx`) to perform automatic and manual evaluations for relations generated by the methods presented above (Granada [22]). The framework is developed in Python and implements: (i) methods for extraction of taxonomic relations based on rules, such as patterns[16] and head-modifier[17]; and (ii) statistical methods based on hierarchical clustering[18], distributional inclusion[19], document subsumption[20] and entropy[21].

6 Open Relation Extraction

Open relation extraction systems aim at identifying all possible relations from an open-domain corpus, with no pre-specified definition of the relations [23]. These systems aim at extracting relation triples from corpus without requiring human supervision. In relation triples such as (E1, Rel, E2), E1 and E2 denote entities (represented usually by nouns or noun phrases), and Rel denotes a relation holding between E1 and E2.

In [24], we extracted relations between named entities in the Organisation domain, using Conditional Random Fields (CRF). Different feature configurations for CRF based on lexical, syntactic and semantic information have been evaluated. The evaluation was based on a subset of HAREM corpus[4] to which we added an extra annotation layer. Our annotation considered the relation descriptors ocurring between named entities of the following categories: Organisation, Person and Place. Relation descriptors are defined as the text chunks that describe an explicit relation between these entities in a sentence. For example, we have the relation descriptor *"diretor de"* (*"director of"*) that occurs between the named entities *"Ronaldo Lemos"* and *"Creative Commons"* in the sentence *"Ronaldo Lemos,* ***diretor da*** *Creative Commons, [...]"*. The annotation was performed by two linguists. Given two named entities occurring in the same sentence, if there is a text sequence (descriptor) that describes an explicit relation between these entities, it is annotated.

Based on this data, in [25], we evaluated a CRF classifier for the extraction of relation descriptors between pairs of named entities (organisations and persons - organisations and places), and also the extraction of pre-defined relation types between these entities (*"affiliation"* and *"placement"*). The resources produced in this work, texts and corresponding manually annotated triples (NE1, relation descriptor, NE2), are available at `http://www.inf.pucrs.br/linatural/data_set_RE.html`.

7 Coreference Resolution

Coreference resolution is the process of identifying mentions to the same entity in a text. In other words, this process consists of identifying the set of expressions that refer to a specific entity. For example, "The opinion is from Miguel Guerra. The agronomist...". In this case, the noun phrase "The agronomist" is coreferent with "Miguel Guerra". In [26] we propose a rule-based approach to solve coreference in Portuguese. Basically, our model is an adaptation of the system by Lee *et al.* [27], solving coreference for nominal nps, using plain texts as input. In a more recent work [28] we investigate sematic knowledge (Hyponymy and Synonymy) based on the relations provided by Onto.PT[29]. Our new semantic model is available at `http://ontolp.inf.pucrs.br/corref/`.

As part of this research we also developed Summ-it++[30], a new enriched version of the Summ-it corpus [31]. This new version adds two annotation layers to the previous coreference annotation: named entities and relations between named entities (based on the works described in Sections 2 and 6). Besides, the annotation format was changed to a well-known and widely used standard, the SemEval [32]. Summ-it++ is available at `http://www.inf.pucrs.br/linatural/summit_plus_plus.html`.

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

8 Sentiment analysis

Sentiment analysis studies methods to analyze people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities and their aspects. This field is also known as: opinion mining, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining, *etc.* [33].

In [34], we proposed a lexicon-based approach to sentiment analysis in short text media, applying it to analyse Twitter messages. This approach has been later extended, in [35], to perform entity-centric sentiment analysis in Twitter messages. The process consists of identifying to which named entity in the message, each opinion-bearing expressions refers to. The reference disambiguation is achieved using a SVM machine. As part of this work we developed the opinion lexicon OpLexicon [36], available in `http://ontolp.inf.pucrs.br/Recursos/downloads-OpLexicon.php`

In Freitas [37], we propose a sentiment analysis methodology based on features and ontologies. Initially, the method receives as input a set of reviews, which are preprocessed. After, features are identified in the preprocessed reviews using a domain ontology. The polarity is identified in the reviews considering features and using available Portuguese sentiment lexicons and linguistic rules. Finally, a summary with features and their respective polarities is generated. In [38], we analysed three different POS tagger tools to choose the best one for our experiments. We also analysed four different sentiment lexicons: SentiLex [39], Brazilian Portuguese Linguistic Inquiry and Word Count dictionary[5], synsets with polarities of Onto.PT [40] and the one we developed, OpLexicon [36].

9 Ontology development and alignment

On the area of ontology development we are studying ontology and multi-agent technologies. In this research direction, we aim to provided a tool for engineering multi-agent systems (MAS) using an ontology as a meta-model [41]. That work extends our ideas towards models of MAS represented as abstractions in ontologies [42,43]. A video that briefly demonstrates our multi-agent system engineering tool based on ontologies can be found in `https://www.youtube.com/watch?v=Lt5ZVG1cgBQ`.

We are also dealing with ontology alignment in two main fronts (i) alignment between top-level and domain ontology and (ii) ontology alignment visualization. In the first case, we are analysing the behavior of state-of-the-art matching systems to align different kinds of ontologies (domain and top-level). A top-level ontology is a high-level and domain independent ontology. The concepts expressed are intended to be basic and universal to ensure generality and expressivity for a wide range of domains [44]. Our goal is to improve the process of matching top-level and domain ontologies. In the second case, we built an environment

[5] http://www.nilc.icmc.usp.br/portlex/index.php/pt/projetos/liwc

for handling ontology alignments with a visual approach: VOAR (Visual Ontology Alignment Environment) [45], available at (`http://voar.inf.pucrs.br`). Within this graphical environment, users can manually create, suppress and edit correspondences and apply a set of operations on alignments (filtering, merge, difference, etc.). Evaluation of multiple alignments, against a reference one, can be carried out with both qualitative and quantitative metrics. Finally, in its most recent version [46], VOAR allows the visualization of multiple alignments together from a set of previously loaded or manually created alignments.

10 Conclusion

In this paper we presented an overview of currently available language resources related to research in information extraction and semantic computing that we have produced at our research lab. A summary of these resources with their access links is given below.

- Named Entity Recognition
  - NE annotated corpus - geological entities: `http://www.inf.pucrs.br/linatural/NER.html`
- Term Extraction
  - Domain corpora: `http://www.inf.pucrs.br/peg/lucelenelopes/ll_crp.html`
  - List of concepts: `http://www.inf.pucrs.br/peg/lucelenelopes/ll_trm.html`
  - English-Portuguese IT dictionary and parallel corpora: `http://www.inf.pucrs.br/~linatural/multilingual`
- Semantic relation identification
  - List of semantically related pairs: `http://www.inf.pucrs.br/linatural/wikimodels/similarity.html`
- Taxonomic relations extraction
  - HREx framework: `https://github.com/rogergranada/HREx`
- Open Relation Extraction
  - Corpus and relation triples: `http://www.inf.pucrs.br/linatural/data_set_RE.html`
- Coreference resolution
  - CORP: `http://ontolp.inf.pucrs.br/corref/`
  - Summ-it++: `http://www.inf.pucrs.br/linatural/summit_plus_plus.html`
- Sentiment analysis
  - OPLexicon: `http://ontolp.inf.pucrs.br/Recursos/downloads-OpLexicon.php`
- Ontologies
  - VOAR - alignment visualization: `http://voar.inf.pucrs.br`

We are happy to share the above research resources with the community. Our current and future efforts are related to the improvement, integration and visualization of the information provided in these resources.

Acknowledgments. This work is partially supported by CNPq, CAPES and FAPERGS.

References

1. do Amaral, D.O.F., Vieira, R.: NERP-CRF: uma ferramenta para o reconhecimento de entidades nomeadas por meio de conditional random fields. linguamatica **6**(1) (2014) 41–49
2. Lopes, L., Vieira, R.: Building and applying profiles through term extraction. In: Proceedings of the 10th Brazilian Symposium in Information and Human Language Technology. STIL 2015, Natal, RN, Brazil, IEEE Press (2015) 193–196
3. Lopes, L., Vieira, R.: Heuristics to improve ontology term extraction. In: PROPOR 2012 – International Conference on Computational Processing of Portuguese Language. LNCS vol. 7243 (2012) 85–92
4. Lopes, L., Fernandes, P., Vieira, R.: Estimating term domain relevance through term frequency, disjoint corpora frequency - tf-dcf. Knowledge-Based Systems **97** (2016) 237 – 249
5. Lopes, L., Vieira, R.: Evaluation of cutoff policies for term extraction. Journal of the Brazilian Computer Society **21**(1) (2015)
6. Lopes, L., Fernandes, P., Vieira, R., Fedrizzi, G.: ExATO lp – An Automatic Tool for Term Extraction from Portuguese Language Corpora. In: Proceedings of the 4th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC '09), Poznan, Poland (2009) 427–431
7. Lopes, L.: Extração automática de conceitos a partir de textos em língua portuguesa. PhD thesis, PUCRS University - Computer Science Department, Porto Alegre, Brazil (2012)
8. Lopes, L., Vieira, R.: Building domain specific parsed corpora in portuguese language. In: Proceedings of the X National Meeting on Artificial and Computational Intelligence (ENIAC). (2013) 1–12
9. Lopes, L., Fernandes, P., Granada, R., Vieira, R.: The impact of contrastive corpora for term relevance measures. In: 2015 Brazilian Conference on Intelligent Systems, IEEE (2015) 146–151
10. Hilgert, L.W., Lopes, L., Freitas, A., Vieira, R., Hogetop, D.N., Vanin, A.A.: Building domain specific bilingual dictionaries. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014. (2014) 2772–2777
11. Harris, Z.: Distributional structure. Words **10**(23) (1954) 146–162
12. Grefenstette, G.: Explorations in automatic thesaurus discovery. Kluwer Academic Publishers Norwell (1994)
13. Granada, R., Vieira, R., Lima, V.: Evaluating co-occurrence order for automatic thesaurus construction. In: Proceedings of the IEEE 13th International Conference on Information Reuse and Integration. IRI 2012 (2012) 474–481
14. Rubenstein, H., Goodenough, J.B.: Contextual correlates of synonymy. Commun. ACM **8**(10) (October 1965) 627–633
15. Granada, R., dos Santos, C.T., Vieira, R.: Comparing semantic relatedness between word pairs in portuguese using wikipedia. In: Proceedings of 11th PROPOR, (PROPOR 2014). (2014) 170–175

16. Hearst, M.: Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 14th conference on Computational linguistics - Volume 2, Stroudsburg, PA, USA, Association for Computational Linguistics (1992) 539–545
17. Radford, A.: Syntax: A minimalist introduction. Cambridge University Press (1997)
18. Caraballo, S.: Automatic construction of a hypernym-labeled noun hierarchy from text. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics. (1999) 120–126
19. Weeds, J., Weir, D., McCarthy, D.: Characterising measures of lexical distributional similarity. In: Proceedings of the 20th International Conference on Computational Linguistics. COLING-2004 (2004) 1015–1021
20. Sanderson, M., Croft, B.: Deriving concept hierarchies from text. In: Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval. (1999) 206–213
21. Santus, E., Lenci, A., Lu, Q., Schulte im Walde, S.: Chasing hypernyms in vector spaces with entropy. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics (2014) 38–42
22. Granada, R.: Evaluation of methods for taxonomic relation extraction from text. PhD thesis, Pontifícia Universidade Católica do Rio Grande do Sul – Université Toulouse III – Paul Sabatier (2015)
23. 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.: Proceedings of ACL-08: HLT, Columbus, Ohio, Association for Computational Linguistics (2008) 28–36
24. Collovini, S., Pugens, L., Vanin, A.A., Vieira, R.: 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, Santiago de Chile, Chile (2014) 108–119
25. Collovini, S., Machado, G., Vieira, R.: A sequence model approach to relation extraction in portuguese. In: Proceedings of the 10th Annual Conference on Language Resources and Evaluation (LREC 2016). (In Press)
26. Fonseca, E.B., Vieira, R., Vanin, A.: Adapting an entity centric model for portuguese coreference resolution. In: Proceedings of the 10th Annual Conference on Language Resources and Evaluation (LREC 2016), In Press (2016)
27. Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., Jurafsky, D.: Deterministic coreference resolution based on entity-centric, precision-ranked rules. Computational Linguistics **39**(4) (2013) 885–916
28. Fonseca, E.B., Vieira, R., Vanin, A.: Improving coreference resolution with semantic knowledge. In: Proceedings of the 12th International Conference on the Computational Processing of Portuguese (PROPOR 2016), In Press (2016)
29. Oliveira, H.G., Gomes, P.: Eco and onto. pt: a flexible approach for creating a portuguese wordnet automatically. Language Resources and Evaluation **48**(2) (2014) 373–393
30. Antonitsch, A., Figueira, A., Amaral, D., Fonseca, E., Vieira, R., Collovini, S.: Summ-it++: an enriched version of the summ-it corpus. In: Proceedings of 10th edition of the Language Resources and Evaluation Conference (LREC 2016), In Press (2016)
31. Collovini, S., Carbonel, T.I., Fuchs, J.T., Coelho, J.C., Rino, L., Vieira, R.: Summ-it: Um corpus anotado com informaes discursivas visando a sumarizao automática.

In: Proceedings of V Workshop em Tecnologia da Informao e da Linguagem Humana. (2007) 1605–1614
32. Recasens, M., Màrquez, L., Sapena, E., Martí, M.A., Taulé, M., Hoste, V., Poesio, M., Versley, Y.: Semeval-2010 task 1: Coreference resolution in multiple languages. In: Proceedings of the 5th International Workshop on Semantic Evaluation, Association for Computational Linguistics (2010) 1–8
33. Liu, B.: Sentiment Analysis and Opinion Mining (Synthesis Lectures on Human Language Technologies). California: Morgan & Claypool Publishers (2012)
34. Souza, M., Vieira, R.: Sentiment analysis on twitter data for portuguese language. In: Computational Processing of the Portuguese Language. Springer Berlin Heidelberg (2012) 241–247
35. Souza, M., Vieira, R.: Entity-centric sentiment analysis on twitter data for the potuguese language. In: 9th Brazilian Symposium in Information and Human Language Technology (STIL 2013), Fortaleza, Brazil. (2013)
36. Souza, M., Vieira, R., Chishman, R., Alves, I.M.: Construction of a portuguese opinion lexicon from multiple resources. In: 8th Brazilian Symposium in Information and Human Language Technology. (2011) 59–66
37. Freitas, L.: Feature-level sentiment analysis applied to brazilian portuguese reviews. PhD thesis, Pontifícia Universidade Católica do Rio Grande do Sul (2015)
38. Freitas, L., Vieira, R.: Comparing portuguese opinion lexicons in feature-based sentiment analysis. International Journal of Computational Linguistics and Applications **1**(4) (2013) 147–158
39. Silva, M.J., Carvalho, P., Sarmento, L.: Building a sentiment lexicon for social judgement mining. In: 10th International Conference Computational Processing of the Portuguese Language. (2012) 218–228
40. Oliveira, H.G., Santos, A.P., Gomes, P.: Assigning polarity automatically to the synsets of wordnet-like resource. In: 3rd Symposium on Languages, Applications and Technologies. (2014) 169–184
41. Freitas, A., Hilgert, L., Marczak, S., Meneguzzi, F., Bordini, R.H., Vieira, R.: A multi-agent systems engineering tool based on ontologies. In: 34th International Conference on Conceptual Modeling, Stockholm, Sweden. Lecture Notes in Computer Science, Springer (2015)
42. Freitas, A., Bordini, R.H., Meneguzzi, F., Vieira, R.: Towards integrating ontologies in multi-agent programming platforms. In: 2013 IEEE/WIC/ACM International Conference on Intelligent Agent Technology, IAT 2013, Atlanta, Georgia, USA. (2013)
43. Freitas, A., Schmidt, D., Panisson, A., Meneguzzi, F., Vieira, R., Bordini, R.H.: Applying ontologies and agent technologies to generate ambient intelligence applications. In: Joint Proceedings Collaborative Agents – Research & Development, CARE for Intelligent Mobile Services & Agents, Virtual Societies and Analytics. (2014) 22–33
44. Semy, S.K., Pulvermacher, M.K., Obrst, L.J.: Toward the use of an upper ontology for u.s. government and u.s. military domains: An evaluation. Technical report, Submission to Workshop on Information Integration on the Web (IIWeb-04) (2004)
45. Severo, B., Trojahn, C., Vieira, R.: VOAR: A visual and integrated ontology alignment environment. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014. (2014) 3671–3677
46. Severo, B., Trojahn, C., Vieira, R.: A gui for visualising and manipulating multiple ontology alignments. In: International Semantic Web Conference (Posters & Demos). Volume 1486. (2015)