Beyond the automatic construction of a lexical ontology for Portuguese: resources developed in the scope of Onto.PT

Hugo Gonçalo Oliveira

CISUC, Dept. of Informatics Engineering, University of Coimbra, Portugal
hroliv@dei.uc.pt

Abstract. Besides the lexical ontology itself, during the Onto.PT project other resources were developed. Those included handcrafted grammars for extracting semantic relations, a term-based lexical-semantic network extracted from dictionaries, a thesaurus with fuzzy memberships, polarities assigned to the Onto.PT synsets, as well as resources used for evaluation, such as manual mappings between words and synsets or the manual classification of synsets and relations as correct or incorrect. This abstract enumerates these resources, which are freely available from the Onto.PT website.

Keywords: lexical knowledge bases, words, extraction grammars, semantic relations, thesaurus, sentiment analysis, evaluation

1 Introduction

Onto.PT [1, 2] is a public domain wordnet-like lexical ontology for Portuguese. Similarly to Princeton WordNet [3], it is structured in synsets – groups of synonymous word senses that can be seen as possible lexicalisations of a natural language concept – and semantic relations connecting synsets – including not only hypernymy (a concept is a kind of another) and part-of (a concept is part of another), but also others, such as causation (a concept causes another) or purpose-of (a concept is used for another). Recently, dictionary definitions were also assigned to part of the synsets, to work as glosses.

The main difference between Onto.PT and other wordnets is that it is created automatically, from available resources. This is an alternative to the time-consuming manual creation and leads to a larger wordnet, in a trade-off on the lower reliability. Onto.PT is thus not a static resource, as either improvements to the creation approach or the exploitation of different resources can lead to new versions.

Besides the wordnet itself, during the Onto.PT project several other resources were developed. These include handcrafted grammars for extracting semantic relations from text (described in sec. 2.1); the output of some of the creation steps, namely a term-based lexical network (sec. 2.2) and a fuzzy thesaurus (sec. 2.3), recent experiments performed towards the assignment of a polarity to the synsets

(sec. 2.4); and the results of several manual evaluations performed along the way (sec. 2.5). All these resources might be useful for other researcher and are thus freely available from Onto.PT's website[1].

2 Resources description

The current version of Onto.PT is 0.6, but older versions are also available. This resource is represented in a RDF/OWL model, based on a similar representation of Princeton WordNet [4]. It is available in two common formats for these models, namely RDF/XML and the more compact N3, which enable it to be loaded in a triple store, thus providing tools such as querying and inferencing.

Onto.PT is created following the ECO approach, which consists of the following steps: (i) relation extraction from text; (ii) synset discovery by synonym clustering; (iii) relation arguments mapping to synsets (ontologisation); (iv) definition assignment. ECO and the contents of Onto.PT are explained elsewhere [1, 2]. For the current version, in the relation extraction step, two public Portuguese dictionaries were exploited, namely Wiktionary.PT[2] and Dicionário Aberto (DA) [5], then merged with the relations from the term-based lexical network PAPEL [6]. Moreover, in the the synset discovery step, an available Portuguese thesaurus is used as a starting point, TeP [7], merged with the synsets of a Portuguese wordnet, OpenWordNet-PT [8], and then augmented with the synonymy relations obtained in the previous step and also those from another thesaurus, OpenThesaurus.PT[3]. In the remaining of this section, other available resources are described.

2.1 Relation extraction grammars

The first step of ECO focused on the extraction of semantic relations from available Portuguese dictionaries, namely Wiktionary.PT and DA. Given that many regularities are preserved across different dictionaries, these relations were acquired using the handcrafted grammars developed in the scope of PAPEL[4], which extract relations between the definiendum and words in the definition. These grammars are editable text files that work with the chart parser PEN[5].

Moreover, in the scope of Onto.PT and using the grammars of PAPEL as a starting point, other extraction grammars were developed, initially for extracting semantic relations from raw text. However, their development was made towards Wikipedia.PT abstracts, which were exploited for extracting synonymy, hypernymy, part-of, causation and purpose-of relations between nouns [9]. Due to the poorer quality of the obtained results and to the scope of the Wikipedia

[1] Check `http://ontopt.dei.uc.pt/index.php?sec=downloads` → Outros recursos
[2] Available from `http://pt.wiktionary.org`
[3] Previously available from `http://openthesaurus.caixamagica.pt/`
[4] Available from `http://www.linguateca.pt/PAPEL/`
[5] Available from `https://code.google.com/p/pen/`

abstracts, the relations obtained with these grammars ended up not being integrated in Onto.PT. But the grammars are available and can be used as a starting point for a relation extraction system from raw text, or for comparison purposes with alternative approaches.

2.2 CARTÃO: term-based semantic relations

The relations extracted from Wiktionary and DA are also available in the same triple format as PAPEL – *word*$_1$ RELATION_PREDICATE *word*$_2$ – where each predicate connects a pair of words of a specific part-of-speech. The three relation sets combined make up the large term-based lexical network CARTÃO [10], which contains about 147,000 terms (≈93,000 nouns, ≈31,800 verbs, ≈31,000 adjectives, ≈3,500 adverbs), connected by about 331,000 relation instances that cover a broad range of types, distributed as follows: ≈135,400 synonymy (SINONIMO_N_DE, SINONIMO_V_DE, SINONIMO_ADJ_DE, SINONIMO_ADV_DE); ≈95,700 hypernymy (HIPERONIMO_DE); ≈9,600 part-of (PARTE_DE, PARTE_DE_ALGO_COM_PROPRIEDADE, PROPRIEDADE_DE_ALGO_PARTE_DE); ≈8,500 member-of (MEMBRO_DE, MEMBRO_DE_ALGO_COM_PROPRIEDADE, PROPRIEDADE_DE_ALGO_MEMBRO_DE); ≈680 contained-in (CONTIDO_EM, CONTIDO_EM_ALGO_COM_PROPRIEDADE); ≈900 material-of (MATERIAL_DE); ≈12,800 causation-of (CAUSADOR_DE, CAUSADOR_DE_ALGO_COM_PROPRIEDADE, PROPRIEDADE_DE_ALGO_QUE_CAUSA, ACCAO_QUE_CAUSA, CAUSADOR_DA_ACCAO); ≈2,400 producer-of (PRODUTOR_DE, PRODUTOR_DE_ALGO_COM_PROPRIEDADE, PROPRIEDADE_DE_ALGO_PRODUTOR_DE); ≈16,600 purpose-of (FAZ_SE_COM, FAZ_SE_COM_ALGO_COM_PROPRIEDADE, FINALIDADE_DE, FINALIDADE_DE_ALGO_COM_PROPRIEDADE); ≈2,400 quality-of (TEM_QUALIDADE, DEVIDO_A_QUALIDADE); ≈600 state-of (TEM_ESTADO, DEVIDO_A_ESTADO); ≈1,700 place-of (LOCAL_ORIGEM_DE); ≈4,100 manner-of (MANEIRA_POR_MEIO_DE, MANEIRA_COM_PROPRIEDADE); ≈270 manner-without (MANEIRA_SEM, MANEIRA_SEM_ACCAO); ≈37,700 property-of (DIZ_SE_SOBRE, DIZ_SE_DO_QUE); ≈1,500 antonymy (ANTONIMO_N_DE, ANTONIMO_V_DE, ANTONIMO_ADJ_DE, ANTONIMO_ADV_DE). In the manual evaluation of a previous extraction, we concluded that accuracy depended on the kind of relation and in the dictionary. It ranged from 98-100% (synonymy between nouns) to 69% (purpose-of in PAPEL) and 69-75% (property-of in Wiktionary.PT).

2.3 CLIP: a fuzzy thesaurus

In ECO's synset discovery step, clusters of synonymous words are identified in the synonymy network extracted from text. Ideally, polysemous words will belong to more than one cluster and, according to the configuration of the network, the membership of a word to a cluster can have different values, which is why we call our synsets fuzzy (check [11] for more information). Although the current version of Onto.PT uses TeP and OpenWordNet-PT as a starting point, in previous version of Onto.PT, we have discovered synsets from the synonymy networks extracted from PAPEL, DA, Wiktionary.PT, TeP and OpenThesaurus.PT. The

resulting fuzzy thesaurus, CLIP, is available from Onto.PT's website. The manual evaluation of this resource lead to an accuracy of about 83%.

Having in mind that word senses are not discrete [12], the fuzzy thesaurus representation is closer to reality than a simple thesaurus. Moreover, in word sense disambiguation, choosing the synset where the target word has higher membership might be used as a baseline.

2.4 Synset polarity

To enable its use in sentiment analysis tasks, we have recently applied a polarity assignment and propagation procedure to Onto.PT [13], where we have exploited SentiLex-PT [14], a public lexicon with the typical polarity of Portuguese words towards human subjects, to assign polarities to the Onto.PT synsets. The result of this procedure can be seen as sentiment wordnet, with some similarities to SentiWordNet [15]. It consists of 14,000 Onto.PT synsets with negative (≈8,000), positive (≈4,200) and neutral assigned polarities (≈1,650), also available. In an evaluation with a previous version of Onto.PT, the polarity accuracy was between 70% and 79%.

2.5 Evaluation package

In addition to all the previous resources, we made available the result of several manual evaluations and gold collections, most performed by two human judges. The evaluation package includes samples of: (i) term-relations from PAPEL and extracted from DA and Wiktionary.PT – classified as correct (2), related but wrong relation (1) or incorrect (0). (ii) synonymy pairs from PAPEL and their suitable TeP synsets, from those containing one of the words; (iii) synonymy pairs from the same discovered synset – classified as correct (1) or incorrect (0); (iv) complete synsets – classified as correct (1), if all the words in the same synset share a meaning, or incorrect (0) otherwise; (v) term-relations from PAPEL and suitable TeP synsets for mapping their arguments; (vi) the classification of synsets and synset relations – both either as correct (1) or incorrect (0) – the final evaluation of Onto.PT. These resources are of great utility for evaluating further improvements of Onto.PT and the steps of ECO, and might be useful for other researchers working on this area and willing to evaluate their results.

3 Conclusion

This abstracted is a brief description of the freely available resources developed in the scope of the Onto.PT project, apart from the wordnet lexical ontology. They include grammars for relation extraction from Portuguese raw text, semantic relation instances extracted from dictionaries, a fuzzy thesaurus, a synset-oriented sentiment lexicon, and several datasets that were used for evaluating the creation steps of Onto.PT. We recall that all of these resources are freely available and we sincerely hope that they can be useful for the Portuguese NLP community.

References

1. Gonçalo Oliveira, H., Gomes, P.: ECO and Onto.PT: A flexible approach for creating a Portuguese wordnet automatically. Language Resources and Evaluation **48**(2) (2014) 373–393
2. Gonçalo Oliveira, H., Gomes, P.: Onto.PT: recent developments of a large public domain Portuguese wordnet. In: Proceedings of the 7th Global WordNet Conference. GWC'14, Tartu, Estonia (2014) 16–22
3. Fellbaum, C., ed.: WordNet: An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press (1998)
4. van Assem, M., Gangemi, A., Schreiber, G.: RDF/OWL representation of WordNet. W3c working draft, World Wide Web Consortium (June 2006)
5. Simões, A., Sanromán, A.I., ao Almeida, J.J.: Dicionário-Aberto: A source of resources for the Portuguese language processing. In: Procs of 10th International Conference on Computational Processing of the Portuguese Language (PROPOR 2012). Volume 7243 of LNCS., Coimbra, Portugal, Springer (April 2012) 121–127
6. Gonçalo Oliveira, H., Gomes, P., Santos, D., Seco, N.: PAPEL: a dictionary-based lexical ontology for Portuguese. In: Procs of 8th International Conference on Computational Processing of the Portuguese Language (PROPOR 2008). Volume 5190., Aveiro, Portugal, Springer (September 2008) 31–40
7. Maziero, E.G., Pardo, T.A.S., Felippo, A.D., Dias-da-Silva, B.C.: A Base de Dados Lexical e a Interface Web do TeP 2.0 - Thesaurus Eletrônico para o Português do Brasil. In: VI Workshop em Tecnologia da Informação e da Linguagem Humana. TIL (2008) 390–392
8. de Paiva, V., Rademaker, A., de Melo, G.: OpenWordNet-PT: An open Brazilian Wordnet for reasoning. In: Procs of 24th International Conference on Computational Linguistics: Demonstration Papers. COLING, Mumbai, India, The COLING 2012 Organizing Committee (2012) 353–360
9. Gonçalo Oliveira, H., Santos, D., Gomes, P.: Extracção de relações semânticas entre palavras a partir de um dicionário: o PAPEL e sua avaliação. Linguamática **2**(1) (May 2010) 77–93
10. Gonçalo Oliveira, H., Antón Pérez, L., Costa, H., Gomes, P.: Uma rede léxico-semântica de grandes dimensões para o português, extraída a partir de dicionários electrónicos. Linguamática **3**(2) (December 2011) 23–38
11. Gonçalo Oliveira, H., Gomes, P.: Automatic discovery of fuzzy synsets from dictionary definitions. In: Proceedings of 22nd International Joint Conference on Artificial Intelligence (IJCAI), Barcelona, Spain, IJCAI/AAAI (2011) 1801–1806
12. Kilgarriff, A.: Word senses are not bona fide objects: implications for cognitive science, formal semantics, NLP. In: Proceedings of 5th International Conference on the Cognitive Science of Natural Language Processing. (1996) 193–200
13. Gonçalo Oliveira, H., Paulo-Santos, A., Gomes, P.: Assigning polarity automatically to the synsets of a wordnet-like resource. In: 3rd Symposium on Languages, Applications and Technologies (SLATE 2014) - Bragança, Portugal. OASICS, Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik (June 19-20 2014) 169–184
14. Silva, M.J., Carvalho, P., Sarmento, L.: Building a sentiment lexicon for social judgement mining. In: Procs of 10th International Conference on Computational Processing of the Portuguese Language (PROPOR 2012). Volume 7243 of LNCS., Coimbra, Portugal, Springer (April 2012) 218–228
15. Esuli, A., Sebastiani, F.: SentiWordNet: A publicly available lexical resource for opinion mining. In: Proceedings of the 5th Conference on Language Resources and Evaluation. LREC 2006 (2006) 417–422