The Portuguese B^{2}SG: A Semantic Test for Distributional Thesaurus

Rodrigo Wilkens(✉), Leonardo Zilio, Eduardo Ferreira, and Aline Villavicencio

Institute of Informatics, UFRGS, Porto Alegre, Brazil
{rodrigo.wilkens,lzilio,eduardo.ferreira,avillavicencio}@inf.ufrgs.br

Abstract. The lack of availability of gold standards for evaluation of distributional thesauri is a stumbling block that prevents a direct comparison of alternative approaches in a uniform way. Here we present B^{2}SG, a TOEFL-like task for Portuguese that contains 2,875 tests with semantic relations (synonyms, antonyms and hypernyms) for nouns and verbs. The resource is validated by comparing it with lexical resources and by human judgment. The resource was used for evaluating two distributional thesauri: one built from lemmata and the other from surface forms. The evaluation of thesauri demonstrated that the use of lemmata is slightly more accurate than the use surface forms for building distributional thesauri. B^{2}SG is readily available for download (http://www.inf.ufrgs.br/pln/resource/B2SG.zip).

Keywords: Gold standard · Semantic relations · Distributional thesauri

1 Introduction

The importance of resources such as WordNet [5] can be measured by the number of initiatives dedicated to (re)produce them in other languages, such as the EuroWordNet [15] and the Global WordNet Association [3]. For Portuguese, such initiatives include Onto.PT [7], OpenWN-PT [13], WordNet.PT [10], WordNet.Br [4]. Manual construction of such resources is costly and time-consuming, and normally ends up with low coverage. An alternative is the automatic construction of distributional thesauri, that present semantic similarity between words and are both language independent and applicable to any domain [9].

The automatic evaluation of such thesauri is a complex task, because of the lack of resources with information on the similarity of words. Moreover, due to the large scale of the resulting thesauri the evaluation cannot be done manually by judges as it would consume too much time. An alternative is the evaluation of thesauri based on their performance on a particular task. For instance, we can approximate the concept of similarity by presenting an explicit semantic relation between words, such as the TOEFL [8] and the WordNet-Based Synonymy Test (WBST) [6] for English. For Portuguese, there are no specific gold standards for the evaluation of distributional thesauri. Here we present the BabeNet-Based

J. Silva et al. (Eds.): PROPOR 2016, LNAI 9727, pp. 333–339, 2016.
DOI: 10.1007/978-3-319-41552-9_33

Semantic Gold Standard (B^2SG)[1], that builds upon the WBST, and we also use this resource as a means to evaluate two distributional thesauri.

We discuss the methodology used for developing and validating B^2SG in Sect. 2. The creation and evaluation of distributional thesauri is then described in Sects. 3 and 4. Conclusions and future work are presented in Sect. 5.

2 Constructing B²SG

The BabelNet-Based Semantic Gold Standard (B²SG) contains nouns and verbs involving antonym, hypernym and synonym relations. Similar to Toefl [8] and WBST [6], for each target word the B²SG lists 4 alternatives: one semantically related word, and 3 unrelated words. For instance, for the target noun *concorrente* and the synonym relation, there are four alternatives: *competidor*, *cortina*, *amurada*, and *carmesim*, among which the correct alternative is the first.

The data set was generated in 3 steps:

1. **Selection of target words:** we used a word frequency list from AC/DC project[2] to avoid low-frequency words. The number of senses of each word was annotated using BabelNet [12], and words not found on BabelNet were excluded from B²SG.
2. **Selection of semantically related words:** for each word in the frequency list from step one we selected a set of semantically related candidates from BabelNet. We then selected the candidate with closest frequency and number of senses regarding the target word. A total of 10,000 nouns and 5,000 verbs were chosen for synonymy and hypernymy, abiding to the restriction that they were the closest in frequency[3]. The antonym category for both verbs and nouns did not present the respective minimum of 10,000 and 5,000 words, so we used all candidates, without applying a frequency filter.
3. **Selection of unrelated words**: using the list of words from Step 2 we selected, for each target word, only words without explicit relation to it. These selected words were randomly divided in groups of 3 words, and we then selected the group with closest mean frequency and mean number of senses in regard to the target word.

Using this process, we ensured that the target, related and unrelated words were close in terms of frequency and polysemy. It is also important to clarify that the same group of words were used in multiple test items, either as target, related or unrelated word. After going through these three steps, we selected a list of test items containing 4,734 target words (1,200 verbs and 3,534 nouns), as shown in Table 1.

For a validation of B²SG, the data set was first evaluated against Onto.PT [7], a thesaurus for Portuguese. As a second step, any relation that was not found in

[1] A preliminary version of B^2SG, yet without validation, was presented in [16].
[2] Available at http://www.linguateca.pt/ACDC.
[3] This list of 10,000 words include both target and related words.

Table 1. B²SG per relation

	Synonyms	Hypernyms	Antonyms	*Total*
Verbs	500	500	200	*1,200*
Nouns	1,667	1,667	200	*3,534*
Total	*2167*	*2167*	*400*	*4,734*

Onto.PT was manually evaluated by two native speaker human judges. As a result 25.4 % of the resource was found in Onto.PT, and another 35.3 % were validated by human judges. From the initial 4,734 relations, 60.7 % in total were considered valid, resulting in a gold standard with 2,875 validated relations.

Table 2. Semi-automatic validation

	Antonym		Synonym		Hypernym		Total
	N	V	N	V	N	V	
Initial	200	200	1667	500	1667	500	4734
Onto.PT	40	51	676	244	191	0	1202
Human Judges	105	116	495	191	568	198	1673
Total Validated	145	167	1171	435	759	198	2875
% Correct	*72,5 %*	*83,5 %*	*70,2 %*	*87,0 %*	*45,5 %*	*39,6 %*	*60,7 %*

As can be seen in Table 2, the validation resulted in more true positive relations for verbs than for nouns, and more for synonyms and antonyms than hypernyms. In the case of nouns, many of the false positive candidates were proper nouns (e.g. *Martinho*), letters (e.g. *c*), abbreviations (e.g. *sr.*), and foreign words (e.g. *punch* and *eau*) present in BabelNet. When these words were among unrelated alternatives, they were replaced by other candidates following the same criteria for frequency and number of senses as before. However, when they were either among the target or related words, the whole relation was removed from the resource. For hypernyms, many of the false positives were candidates evaluated by the judges as synonyms, and, as they lacked the more general meaning of a hypernym, they were removed from the resource.

3 Generating Distributional Thesauri

Since one of the possible applications of resources such as B²SG is the evaluation of distributional thesauri, we decided to make an experiment that could be seen as a baseline for future distributional thesauri for Portuguese. So here we describe the corpus and methodology we used to generate two distributional thesauri that were later evaluated against B²SG.

To obtain a large representative corpus from which to generate the distributional thesauri we combined different Portuguese corpora and used their surface forms and lemmata (Table 3) for training[4]. The whole corpus was parsed with PALAVRAS [2], so that we had access to the lemmata.

Both distributional thesauri were then generated with *word2vec* [11], using Skip-Gram with the following parameters: a vector size of 300 dimensions, a context window of size 5, a downsampling threshold of 1e-5, a sampling of 5 for the negative training algorithm, and a minimum frequency of 10 in the corpus.

Table 3. Corpus information

Corpus	Surface		Lemma	
	Types	Tokens	Types	Tokens
brWaC	812 K	166.7 M	618 K	166.5 M
EuroPal	132 K	47.8 M	67 K	47.9 M
CETEN Folha	206 K	21.2 M	120 K	21.5 M
PLN-BR	582 K	34.1 M	479 K	34.1 M
CETEM Publico	611 K	166.6 M	330 K	138.9 M
Corpus Brasileiro	2.8 M	1 G	-	-
Total	3.7 M	1.5 G	1.5 M	409 M

4 Distributional Thesauri Evaluation

The evaluation of both distributional thesauri was made by obtaining the similarity values between the target word and alternatives for each test item in B^2SG. If the related word (correct alternative) had the highest similarity score among the alternatives, the answer was considered correct.

We evaluated both thesauri using 2 criteria: a strict one, in which all 5 words in a test item (target and all alternatives) had to be in the thesauri; and a non-strict one, in which at least the target and related alternative had to be present.[5] The results of the evaluation are shown in Table 4, for the surface form thesaurus, and in Table 5, for the lemmatized thesaurus.

The results obtained with both thesauri were comparable, and the thesaurus built from the lemmatized corpus performed slightly better in general, even though the corpus was more than three times smaller[6].

[4] The parsed corpus does not include the *Corpus Brasileiro* [1] because the lemma information is different from the one in the other corpora, since it is not parsed with PALAVRAS.

[5] The results for the 2 criteria were very similar, because both thesauri had good coverage in relation to the test items.

[6] As Pennington, Socher and Manning [14] point out, larger corpora lead to better statistics, so we can assume that a larger corpus of lemmatized forms would present even better results.

Table 4. Evaluation of the surface form thesaurus: strict and non-strict criterium

Test	Type	Strict criterium			Non-strict criterium		
		Coverage	Correct	% Correct	Coverage	Correct	% Correct
Antonym	Noun	105	90	85.7 %	145	126	86.9 %
	Verb	143	100	69.9 %	167	118	70.7 %
Hypernym	Noun	545	432	79.3 %	756	606	80.2 %
	Verbs	167	115	68.9 %	198	138	69.7 %
Synonym	Noun	861	726	84.3 %	1167	997	85.4 %
	Verb	366	275	75.1 %	433	332	76.7 %

Table 5. Evaluation of the lemmatized thesaurus: strict and non-strict criterium

Test	Type	Strict criterium			Non-strict criterium		
		Coverage	Correct	% Correct	Coverage	Correct	% Correct
Antonym	Noun	98	82	83.7 %	143	123	86,0 %
	Verb	141	110	78.0 %	167	132	79,0 %
Hypernym	Noun	525	425	81.0 %	753	615	81,7 %
	Verb	166	118	71.1 %	198	141	71,2 %
Synonym	Noun	832	721	86.7 %	1162	1025	88,2 %
	Verb	366	267	73.0 %	433	320	73,9 %

5 Conclusions and Future Work

In this paper, we described the development of B²SG, a TOEFL-like task for Portuguese. We used a lexical resource for Portuguese to extract target words that were similar in frequency and polysemy, resulting in almost five thousand test items including nouns and verbs distributed among three relation types: synonymy, antonymy and hypernymy. After automatic and manual evaluation of the resource, B²SG presents 2,875 validated test items. Using an existing lexical resource for the validation of B²SG made the evaluation faster, since more than one fourth of the gold standard could be automatically validated before passing on to human judgments.

We also used B²SG in an experiment that serves as a baseline for Portuguese distributional thesauri. The test of B²SG was applied for the evaluation of two Portuguese distributional thesauri: one from surface forms and the other from lemmatized forms. On our results, the latter was in general slightly more accurate, even if smaller, than the former.

As future work we plan to apply this methodology to build the resource for other languages, and to extend the test items to include also adjectives and adverbs.

Acknowledgments. This research was partially developed in the context of the project *Text Simplification of Complex Expressions*, sponsored by Samsung Eletrônica da Amazônia Ltda., in the terms of the Brazilian law n. 8.248/91. This work was also partly supported by CNPq (482520/2012- 4, 312114/2015-0) and FAPERGS AiMWEst.

References

1. Berber Sardinha, T., Moreira Filho, J., Alambert, E.: O corpus brasileiro. Comunicaçao ao VII Encontro de Lingüistica de Corpus (2008)
2. Bick, E.: The parsing system Palavras. Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework (2000)
3. Bond, F., Paik, K.: A survey of wordnets and their licenses. In: Proceedings of the 6th Global WordNet Conference. pp. 64–71 (2012)
4. Dias-da-Silva, B.C., Felippo, A.D., das Graças Volpe Nunes, M.: The automatic mapping of Princeton WordNet lexical-conceptual relations onto the brazilian portuguese wordnet database. In: Proceedings of LREC 2008, European Language Resources Association, Marrakech, Morocco (2008)
5. Fellbaum, C.: WordNet. Wiley Online Library, New York (1998)
6. Freitag, D., Blume, M., Byrnes, J., Chow, E., Kapadia, S., Rohwer, R., Wang, Z.: New experiments in distributional representations of synonymy. In: Proceedings of the Ninth Conference on Computational Natural Language Learning. pp. 25–32. Association for Computational Linguistics (2005)
7. Gonçalo Oliveira, H., Gomes, P.: Towards the automatic creation of a wordnet from a term-based lexical network. In: Proceedings of the ACL Workshop TextGraphs-5: Graph-based Methods for Natural Language Processing. pp. 10–18. ACL Press (July 2010). http://eden.dei.uc.pt/~hroliv/pubs/GoncaloOliveira_Gomes2010_TextGraphs5_postconf.pdf
8. Landauer, T.K., Dumais, S.T.: A solution to plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychol. Rev. **104**(2), 211 (1997)
9. Lin, D.: Automatic retrieval and clustering of similar words. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - vol. 2. pp. 768–774. ACL 1998, Association for Computational Linguistics (1998)
10. Marrafa, P.: WordNet do Português: uma base de dados de conhecimento linguístico. Instituto de Camões, Lisboa (2002)
11. Mikolov, T., Karafiát, M., Burget, L., Cernockỳ, J., Khudanpur, S.: Recurrent neural network based language model. In: 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010, Makuhari, Chiba, Japan, pp. 1045–1048, 26–30 September 2010
12. Navigli, R., Ponzetto, S.P.: Babelnet: building a very large multilingual semantic network. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. pp. 216–225. Association for Computational Linguistics (2010)
13. de Paiva, V., Rademaker, A., de Melo, G.: OpenWordNet-PT: an open Brazilian WordNet for reasoning. In: Proceedings of the 24th International Conference on Computational Linguistics (2012). http://www.coling2012-iitb.org (Demonstration Paper). Published also asTechreport http://hdl.handle.net/10438/10274
14. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. EMNLP **14**, 1532–1543 (2014)

15. Vossen, P. (ed.): EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Kluwer Academic Publishers, Norwell (1998)
16. Wilkens, R., Zilio, L., Gonçalves, G., Ferreira, E., Villavicencio, A.: Tesauros distribucionais para o português: avaliação de metodologias. In: Proceedings of STIL 2015. Sociedade Brasileira de Computação (2015)