B^{2}SG: a TOEFL-like Task for Portuguese

Rodrigo Wilkens, Leonardo Zilio, Eduardo Ferreira, Aline Villavicencio
Institute of Informatics, Federal University of Rio Grande do Sul (Brazil)
rodrigo.wilkens, lzilio@inf.ufrgs.br, eduardo.ferreira, avillavicencio@inf.ufrgs.br

Abstract

Resources such as WordNet are useful for NLP applications, but their manual construction consumes time and personnel, and frequently results in low coverage. One alternative is the automatic construction of large resources from corpora like distributional thesauri, containing semantically associated words. However, as they may contain noise, there is a strong need for automatic ways of evaluating the quality of the resulting resource. This paper introduces a gold standard that can aid in this task. The BabelNet-Based Semantic Gold Standard (B^{2}SG) was automatically constructed based on BabelNet and partly evaluated by human judges. It consists of sets of tests that present one target word, one related word and three unrelated words. B^{2}SG contains 2,875 validated relations: 800 for verbs and 2,075 for nouns; these relations are divided among synonymy, antonymy and hypernymy. They can be used as the basis for evaluating the accuracy of the similarity relations on distributional thesauri by comparing the proximity of the target word with the related and unrelated options and observing if the related word has the highest similarity value among them. As a case study two distributional thesauri were also developed: one using surface forms from a large (1.5 billion word) corpus and the other using lemmatized forms from a smaller (409 million word) corpus. Both distributional thesauri were then evaluated against B^{2}SG, and the one using lemmatized forms performed slightly better.

Keywords: distributional thesauri, gold standard, lexical resource

1. Introduction

The importance of resources such as WordNet (Fellbaum, 1998), that represent semantic relations between words, can be measured by the number of initiatives dedicated to (re)producing them in other languages, such as the EuroWordNet[1] (Vossen, 1998) and the Global WordNet Association[2] (Bond and Paik, 2012). These resources have been windely used in numerous NLP applications, such as systems for Q&A, text simplification, and sentiment analysis. For Portuguese, some such initiatives include Onto.PT[3] (Gonçalo Oliveira and Gomes, 2010), OpenWN-PT[4] (de Paiva et al., 2012), MultiWordnet of Portuguese[5], WordNet.PT[6] (Marrafa, 2002), WordNet.Br[7] (Dias-da-Silva et al., 2008).

However, the manual construction of this type of resource is costly and much time-consuming, in addition to having low coverage and being applicable to only one domain. Moreover, its availability is limited or non-existent for many languages. A popular alternative is the automatic construction of distributional thesauri from corpora, resulting in a resource with semantic association among words. These techniques are language independent and applicable to any domain (Lin, 1998). As a consequence there is much attention being devoted to the systematic construction, evaluation and enhancement of distributional thesauri.

The automatic evaluation of such thesauri, in particular, is a complex task, because of the lack of resources with information about the similarity between words. Moreover, due to the large scale of the resulting thesauri a manual evaluation by human judges is prohibitively expensive and would consume too much time. An alternative is the extrinsic evaluation of the quality of a thesaurus, where performance on a particular task would indirectly reflect the quality of a resource. For instance, we can approximate the concept of similarity by presenting an explicit semantic relation between words, as done in the TOEFL test (Landauer and Dumais, 1997) and the WordNet-Based Synonymy Test (WBST) (Freitag et al., 2005) for English. For Portuguese, automatically generated ontologies include BabelNet (Navigli and Ponzetto, 2010) and Onto.PT (Gonçalo Oliveira and Gomes, 2010), but there are no specific gold standards for the evaluation of distributional thesauri. Having this in mind, we developed the BabeNet-Based Semantic Gold Standard (B^{2}SG) for Portuguese, based on the WBST.[8] The main difference between them is that B^{2}SG is based on an automatically generated resource, while WBST is based on a manually constructed one, WordNet.

In this paper we discuss related work in Section 2, and the methodology used for developing B^{2}SG in Section 3. We describe an intrinsic evaluation of the quality of the test items proposed in Section 4 and of their application as part of an extrinsic evaluations in 5. Finally, we present conclusions and future work in Section 6.

[1] http://www.illc.uva.nl/EuroWordNet/

[2] http://globalwordnet.org/wordnets-in-the-world/

[3] http://ontopt.dei.uc.pt

[4] https://github.com/arademaker/openWordnet-PT

[5] http://mwnpt.di.fc.ul.pt/

[6] http://www.clul.ul.pt/clg/wordnetpt/index.html

[7] http://143.107.183.175:21380/wordnetbr

[8] A preliminary version of this study was presented in (Wilkens et al., to appear). In this paper we present an expanded version of the resource, with partial evaluation of the data by human judges. Moreover, we discuss a case study of evaluation of two distributional thesauri using B^{2}SG. B^{2}SG is readily available at http://www.inf.ufrgs.br/pln/resource/B2SG.zip

2. Related Work

For English, there are several datasets for the evaluation of distributional thesauri, such as:

- 65 noun pairs (Rubenstein and Goodenough, 1965)
- 80 test items in TOEFL (Landauer and Dumais, 1997)
- 353 noun pairs (WordSim (Finkelstein et al., 2001))
- 2003 pairs and context sentences (SCWS (Huang et al., 2012))
- 3,000 pairs (Bruni et al., 2014)

Many of these are available as part of Word Vector Evaluation suite (Faruqui and Dyer, 2014).[9] The TOEFL dataset in particular presents, for each target word, four alternatives, and the task is to select which among them is the more semantically related to the target word than the others. Other examples of TOEFL-like tasks include the WordNet-Based Synonymy Test (WBST) (Freitag et al., 2005), which is an extension of the TOEFL test that was automatically generated from WordNet. The dataset presented here adapts the methodology of the latter to BabelNet, selecting 4 alternatives for each target word to create test items automatically for resource limited languages like Portuguese.

3. Building *B²SG*

The BabelNet-Based Semantic Gold Standard contains nouns and verbs involving antonym, hypernym and synonym relations. Like TOEFL (Landauer and Dumais, 1997) and WBST (Freitag et al., 2005), for each target word it lists 4 alternatives: one semantically related word, and 3 potentially unrelated words. For instance, for the target noun *tenderness* and synonym relation, it provides four alternatives: *affection*, *partner*, *inconstancy*, and *lap*, from which the correct alternative for synonym is the first.
The dataset was generated in 3 steps:

1. **Selection of target words:** we used a word frequency list from the AC/DC project[10] to avoid low frequency words. Each word was annotated with information about the number of senses from BabelNet (Navigli and Ponzetto, 2010), and words not found on BabelNet were not included among the target words.
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[11]. 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**: we adopted the list produced in Step 2, but, for each target word, only words without an explicit relation to it were selected as candidates. 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.

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*

4. Validation

The semi-automatic validation was done in two stages: the dataset was first automatically validated against Onto.PT (Gonçalo Oliveira and Gomes, 2010), a thesaurus for Portuguese. As a second step, any relation that was not found in Onto.PT was manually validate by two native speaker human judges. From the set of relations, 25.4% of the resource was found in Onto.PT, and another 35.3% was validated by human judges. From the initial 4,734 relations from B²SG, 60.7% in total were considered valid, resulting in a gold standard with 2,875 relations.
The methodology adopted resulted in more true positive relations for verbs than for nouns, and more for synonyms and antonyms than hypernyms. For nouns, many of the 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 included among one of the 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 targets or related words, they were simply 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.

5. Thesaurus quality evaluation

B²SG was also used to evaluate two automatically constructed distributional thesauri for Portuguese. The idea of using it in an extrinsic thesaurus evaluation is that the performance on the test reflects thesaurus quality. In this case, the accuracy of the answers of the distributional thesauri for the relation test presented by B²SG is measured by the rank of the possible candidates answers, which must place the related work in the top. The thesauri were generated using *word2vec* (Mikolov et al., 2010), using Skip-Gram with

[9] http://wordvectors.org/suite.php
[10] Available at http://www.linguateca.pt/ACDC.
[11] This list of 10,000 words include both target and related words.

Table 2: 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%*

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.

To obtain a large representative corpus we combined different Portuguese corpora and used their surface forms for training (Table 3a). Additionally, we created a corpus using lemmatized forms from the parsed corpora, Table 3b[12]. The parsing information comes from the PALAVRAS parser (Bick, 2000).

Table 3: Corpus information

(a) Surface Form Corpus

Corpus	Types	Tokens
brWaC	812K	166.7M
euroParl	135K	47.8M
CETENfolha	206K	21.2M
PLN-BR	582K	34.1M
CETEMpúblico	611K	166.6M
Corpus Brasileiro	2.8M	1G
Total	3.7M	1.5G

(b) Parsed/Lemmatized Corpus

Corpus	Types	Tokens
brWaC	618K	166.5M
euroParl	67K	47.9M
CETENfolha	120K	21.5M
PLN-BR	479K	34.1M
CETEMpúblico	330K	138.9M
Total	1.5M	409M

The evaluation of both models was made by obtaining the similarity values between the target and alternatives for each test item in B^2SG. If the similarity value of the related word was the highest among the alternatives, the answer was considered correct. The results of the evaluation are in Table 4, for the model constructed with the surface form corpus, and Table 5, for the model with the lemmatized corpus. We evaluated using 2 criteria: a strict one, in which all 5 words in a test item (target and all alternatives) had to be in the thesaurus; and a non-strict one, in which at least the target and related alternative had to be in the thesaurus.[13]

[12]The parsed corpus does not include the *Corpus Brasileiro* (Berber Sardinha et al., 2008) because lemmatized information in this corpus is different from the other corpora, since it is not parsed with PALAVRAS.

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

Table 4: Evaluation of the thesaurus built from surface forms

(a) Strict evaluation

Test	Type	Coverage	Correct	% Correct
Antonym	**Noun**	105	90	85.7%
	Verb	143	100	69.9%
Hypernym	**Noun**	545	432	79.3%
	Verbs	167	115	68.9%
Synonym	**Noun**	861	726	84.3%
	Verb	366	275	75.1%

(b) Non-strict evaluation

Test	Type	Coverage	Correct	% Correct
Antonym	**Noun**	145	126	86.9%
	Verb	167	118	70.7%
Hypernym	**Noun**	756	606	80.2%
	Verbs	198	138	69.7%
Synonym	**Noun**	1167	997	85.4%
	Verb	433	332	76.7%

Table 5: Evaluation of the distributional thesaurus built from lemmatized corpus

(a) Strict evaluation

Test	Type	Coverage	Correct	% Correct
Antonym	**Noun**	98	82	83.7%
	Verb	141	110	78.0%
Hypernym	**Noun**	525	425	81.0%
	Verb	166	118	71.1%
Synonym	**Noun**	832	721	86.7%
	Verb	366	267	73.0%

(b) Non-strict evaluation

Test	Type	Coverage	Correct	% Correct
Antonym	**Noun**	143	123	86,0%
	Verb	167	132	79,0%
Hypernym	**Noun**	753	615	81,7%
	Verb	198	141	71,2%
Synonym	**Noun**	1162	1025	88,2%
	Verb	433	320	73,9%

The results obtained with both thesauri were comparable, and those obtained with the thesaurus built from the lemmatized corpus were slightly better in general, even though the corpus was smaller, as it did not include the Corpus Brasileiro.

6. Conclusions and Future Work

In this paper, we described the development of B^2SG, a TOEFL-like task for Portuguese. The 2,875 test items involve synonyms, antonyms and hypernyms for nouns and verbs. A partly automatic validation of the resource was done using Onto.PT and human judgments. As a case study we used B^2SG in the evaluation of distributional thesauri. We built two thesauri: one from surface forms and another from lemmatized forms, and the latter was slightly more accurate, even if smaller, than the former.

As future work we are going to apply this methodology to build gold standards such as B^2SG 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.

7. Bibliographical References

Tony Berber Sardinha, JL Moreira Filho, and E Alambert. 2008. O corpus brasileiro. *Comunicaçao ao VII Encontro de Lingüıstica de Corpus*.

Eckhard Bick. 2000. The parsing system palavras. *Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*.

Francis Bond and Kyonghee Paik. 2012. A survey of wordnets and their licenses. In *Proceedings of the 6th Global WordNet Conference*, pages 64–71.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *J. Artif. Intell. Res.(JAIR)*, 49:1–47.

Valeria de Paiva, Alexandre Rademaker, and Gerard de Melo. 2012. Openwordnet-pt: An open brazilian wordnet for reasoning. In *Proceedings of the 24th International Conference on Computational Linguistics*. See at http://www.coling2012-iitb.org (Demonstration Paper). Published also as Techreport http://hdl.handle.net/10438/10274.

Bento Carlos Dias-da-Silva, Ariani Di Felippo, and Maria das Graças Volpe Nunes. 2008. The automatic mapping of princeton wordnet lexical-conceptual relations onto the brazilian portuguese wordnet database. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*. European Language Resources Association.

Manaal Faruqui and Chris Dyer. 2014. Community evaluation and exchange of word vectors at wordvectors.org. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Baltimore, USA, June. Association for Computational Linguistics.

Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414. ACM.

Dayne Freitag, Matthias Blume, John Byrnes, Edmond Chow, Sadik Kapadia, Richard Rohwer, and Zhiqiang Wang. 2005. New experiments in distributional representations of synonymy. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, pages 25–32. Association for Computational Linguistics.

Hugo Gonçalo Oliveira and Paulo Gomes. 2010. 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*, pages 10–18. ACL Press, July.

Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 873–882. Association for Computational Linguistics.

Thomas K Landauer and Susan T Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

Dekang Lin. 1998. 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 - Volume 2*, ACL '98, pages 768–774. Association for Computational Linguistics.

Palmira Marrafa. 2002. *WordNet do Portugułs: uma base de dados de conhecimento lingustico*. Instituto de Cames, Lisboa.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, pages 1045–1048.

Roberto Navigli and Simone Paolo Ponzetto. 2010. Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 216–225. Association for Computational Linguistics.

Herbert Rubenstein and John B Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Piek Vossen, editor. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Norwell, MA, USA.

Rodrigo Wilkens, Leonardo Zilio, Gabriel Gonalves, Eduardo Ferreira, and Aline Villavicencio. to appear. Tesauros distribucionais para o portuguls: avaliao de metodologias. In *Proceedings of STIL 2015*. Sociedade Brasileira de Computao.