Assessing Wordnets with WordNet Embeddings

Ruben Branco[1] and **João Rodrigues**[1] and **Chakaveh Saedi**[1,2] and **António Branco**[1]

[1]University of Lisbon
NLX-Natural Language and Speech Group, Department of Informatics
Faculdade de Ciências, Campo Grande, 1749-016 Lisboa, Portugal

[2]Macquarie University
Department of Computing
Sydney, Australia

{ruben.branco,jrodrigues,chakaveh.saedi,ahb}@di.fc.ul.pt

Abstract

An effective conversion method was proposed in the literature to obtain a lexical semantic space from a lexical semantic graph, thus permitting to obtain WordNet embeddings from WordNets. In this paper, we propose the exploitation of this conversion methodology as the basis for the comparative assessment of WordNets: given two WordNets, their relative quality in terms of capturing the lexical semantics of a given language, can be assessed by (i) converting each WordNet into the corresponding semantic space (i.e. into WordNet embeddings), (ii) evaluating the resulting WordNet embeddings under the typical semantic similarity prediction task used to evaluate word embeddings in general; and (iii) comparing the performance in that task of the two word embeddings, extracted from the two WordNets. A better performance in that evaluation task results from the word embeddings that are better at capturing the semantic similarity of words, which, in turn, result from the WordNet that is of higher quality at capturing the semantics of words.

1 Introduction

Lexical semantics studies the semantic properties of lexical units, and is often defined as the study of word meaning. Given its importance, the computational representation of lexical meaning is a core challenge in natural language processing (NLP).

Since the meaning of a word is strongly related to the meaning of other words, the relations between words are a key ingredient for the representation of their meaning. There have been different types of representations proposed for lexical semantics, which, in general, can be viewed as pertaining to one of three main family of representations, namely semantic networks (Quillan, 1966), feature-based models (Minsky, 1975; Bobrow and Norman, 1975), and semantic spaces (Harris, 1954; Osgood et al., 1957).

Semantic networks are a type of approach for lexical semantics that is based on graphs. In a nutshell, a lexical unit, typically a word, is recorded as a node in a graph while the semantic relations among words, such as hyponymy or synonymy, etc., are recorded as labeled edges among the nodes of the graph. One of the most popular semantic networks is WordNet (Fellbaum, 1998). It stands out as being a lexical semantics network based on non trivial linguistic intuitions of human experts.

Feature-based models representing lexical semantics, in turn, resort to a hash table that stores the lexical units as keys, and the semantically related units as the respective values. Small World of Words (De Deyne et al., 2013) is an example of such a model. In its development, the semantic features (related words) of a lexical entry can be obtained straightforwardly from laypersons by using the lexical entry as a cue to evoke possible words associated to it.

Finally, in semantic spaces, the meaning of a lexical unit is represented as a vector in a high dimension space — also known as word embedding —, typically obtained on the basis of the frequency of its co-occurrence with other lexical units, resorting to a large collection of documents. Word2vec (Mikolov et al., 2013) is an example of a method to obtain semantic spaces.

Bridging between these different types of lexical meaning representations is instrumental for a wider use of all the existing lexical semantics resources. Unifying this knowledge in one lexical semantic representation would carry an immediate impact across a range of NLP tasks.

An existing form of (partial) bridging is obtained with the conversion of one type of representation to another as in (Saedi et al., 2018), with the wnet2vec methodology. Wnet2vec permits the conversion from lexical semantic networks to lexical semantic spaces, termed as WordNet embeddings.

The success of this type conversion can be measured by using the typical semantic space evaluation process. That is obtained by comparing the semantic similarity scores between the vectors of words arranged in pairs against the gold scores of semantic similarity among the words in the pairs, which were obtained from human subjects.

The evaluation of the semantic similarity task based on the semantic space wnet2vec used the SimLex-999 (Hill et al., 2016), a mainstream semantic similarity data set composed of 999 pairs of words with a correspondent similarity strength value. Semantic similarity detection with wnet2vec (Saedi et al., 2018) shows an almost 20% superior result against a strong baseline, namely Google's word2vec semantic space, which is trained on a very large collection of 100 Billion token texts.

Our goal in the present paper is to propose the exploitation of this conversion methodology as the basis for the comparative assessment of WordNets: Given two WordNets, for the same language, their relative quality in terms of capturing the lexical semantics of that language, can be assessed by (i) converting each WordNet into the corresponding semantic space (i.e. WordNet embeddings), (ii) evaluating the resulting embeddings in the semantic similarity prediction task; and (iii) comparing the performance in that task of the two word embeddings, extracted from the two WordNets. A better performance results from the word embeddings that better capture the semantic similarity of words, which, in turn, results from the WordNet that is of higher quality at capturing the semantics of words.

In order to illustrate this proposed methodology for the comparative assessment of WordNets with a first exercise with its application, we resort to two WordNets of the same language, Portuguese, developed under two distinct methodologies, namely MWN.PT — hand-crafted — and OWN-PT — built (semi-)automatically.

The next Section 2 reports on the conversion of the hand-crafted WordNet to the respective WordNet embeddings and on the performance of the latter in the semantic similarity prediction task. The following Section the same exercise is undertaken but now with the WordNet built (semi-)automatically. Sections 4 and 5 present, respectively, the discussion of the results and the related work. The conclusions are presented in Section 6.

2 Embeddings from hand-crafted WordNet

The MultiWordnet of Portuguese (MWN.PT) is developed under the same methodological principles as the seminal Princeton English Wordnet — including the resorting to manually validated representations. Its synsets are aligned with the translationally equivalent synsets in Princeton WordNet. It is available from ELRA-European Language Resources Association.[1] Besides the difference in the language covered, MWN.PT differentiates to Princeton WordNet by being smaller, encompassing 17k concepts/synsets (against over 120k of Princeton), by encoding only synonymy and hyponymy/hypernymy (against some 25 semantics relations in Princeton WordNet), and by including only nouns (against all open categories), and includes mostly the sub-ontologies of Person, Organization, Event, Location and Art works. Hence, it offered interesting contrasting conditions to proceed with an empirical study of the strength of the wnet2vec methodology when applied to quite different and more challenging empirical settings than the one originally resorted to in (Saedi et al., 2018) to convert the Princeton WordNet into its WordNet embeddings.

To obtain word embeddings, the mainstream methods have used the frequency of co-occurrence in large corpora between the target word and its neighboring words to construct the respective vector. Instead of texts and the frequency of co-occurrence between words, wnet2vec resorts to lexical semantics graphs and the knowledge encoded in them, using the semantic networks as the empirical source to obtain the vectors of the corresponding semantic space. The key insight in the conversion process is that a stronger semantic affinity between two lexical units is found between nodes that are closer and have a higher number of connecting paths.

[1] MWNT.PT was obtained from http://catalogue-old.elra.info/product_info.php?cPath=42_45&products_id=1101&language=en

In a nutshell, the wnet2vec methodology starts by creating a matrix with all of the possible semantic relations between all the words, resulting in an adjacency matrix M. Then it populates each cell M_{ij} of the matrix resorting to a WordNet, in the present experiment MWN.PT, as the semantic graph G. Each cell M_{ij} is set to 1 if and only if there is a direct edge between synsets including the two words $word_i$ and $word_j$ the cell encodes/represents. Words present in the same synset have a synonym relation and thus are assigned a value of 1. If there is no edge between the two words that cell is set to 0.

For all nodes not directly connected, that is connected through other nodes in between, the representation of their affinity strength is obtained by following the cumulative iteration:

$$M_G^n = I + \alpha M + \alpha^2 M^2 + \ldots + \alpha^n M^n \quad (1)$$

M^n is the matrix where every two words, $word_i$ and $word_j$, are transitively related by n edges. I represents the identity matrix and α is used as a decay factor for longer paths.

The iteration converges into the matrix M_G, obtained by an inverse matrix operation:

$$M_G = \sum_{e=0}^{\infty} (\alpha M)^e = (I - \alpha M)^{-1} \quad (2)$$

After the convergence, a Positive Point-wise Mutual Information transformation (PMI+) is applied to reduce the frequency bias, followed by an L2-norm to normalize each line of M_G, and finally, a Principal Component Analysis (PCA) is applied to reduce the dimension of the vectors. Further details on this conversion can be found in (Saedi et al., 2018).

2.1 From the semantic graph to a corresponding semantic space

When converted to a semantic space and the resulting semantic space is evaluated on the semantic similarity task with SimLex-999, Princeton WordNet supports a wnet2vec whose performance has an accuracy score of 0.50 in terms of Spearman's coefficient (Saedi et al., 2018). On the same task and testing dataset, Google's word2vec semantic space, used as the baseline, obtains 0.44 accuracy score.

While the semantic space obtained from the English WordNet was evaluated with the original SimLex-999 dataset, given we are handling here Portuguese instead, we resort to LX-SimLex-999 (Querido et al., 2017), which resulted from the translation of SimLex-999 into Portuguese.[2]

And while the corpus-based baseline for English was the Google's word2vec semantic space, for the corpus-based baseline here, we resort LX-DSemVectors 2.2b (Rodrigues and Branco, 2018) for Portuguese that also uses word2vec learning tools.[3] This semantic space was trained over a collection of text with more than 2 Billion tokens and obtains state-of-the-art results in a wide range of test datasets, including the LX-SimLex-999. Its best-reported accuracy score with this testing dataset is 0.35, in terms of Spearman's coefficient. All evaluations use the cosine distance measure between the vectors.

We use the same settings as in the experiment with the English WordNet, using here a decay factor of 0.75 and all available semantic relations being taken into account. The dimensions of the embeddings were kept at 850, the best-reported size. In the experiment with English, only 60k of the over 120k synsets in Princeton WordNet were used due to memory footprint limitations. No such reduction was necessary for the MWN.PT conversion due to the smaller size (17k synsets) of this semantic graph.

Our experiments were performed with an Intel Xeon E5-2640 V2 with 2 CPUs, each CPU has 8 cores. The training resorted to an upper bound of 120GB of memory and took 2 days.

2.2 Results

The result obtained with the WordNet embeddings obtained from MWN.PT using wnet2vec methodology can be found in Table 1, together with the score of the baseline. The graph-based semantic space obtained from 15886 words with wnet2vec is 11 percentage points better than the corpus-based baseline obtained from 2.2B words with word2vec.

Given the difference in the size of their vocabularies, the number of similarity pairs with unknown words differs among the two semantic spaces. The LX-DsemVectors 2.2b, trained on more than 2 Billion tokens, covers almost all of the words of the 999 pairs, with only 3.5% pairs with unknown words. The semantic space obtained

[2] Obtained from `https://github.com/nlx-group/LX-DSemVectors`

[3] Obtained from `http://lxcenter.di.fc.ul.pt/datasets/models/2.2b/`

with the MWN.PT has a coverage with 74.9% pairs with unknown words.

Lexical Semantic Model	Similarity
MWN.PT (wnet2vec)	0.4643
LX-DSemVectors 2.2b (word2vec)	0.3502

Table 1: Performance in the semantic similarity task over the LX-SimLex-999, given by Spearman's coefficient (higher score is better).

3 Embeddings from (semi-)automatic WordNet

Given the lessons learned with the creation of a semantic space from the MWN.PT, in the second phase of our experiments we applied the same conversion methodology to another Portuguese WordNet, the OpenWordnet-PT (OWN-PT) (de Paiva et al., 2012).

While MWN.PT was built manually by resorting to human experts labor, OWN-PT is different in that it resorts to (semi-)automatic and machine learning methodologies, and has a dimension that is over three times larger — over 54k words (against over 15k in MWN.PT) —, thus offering an interesting case for empirical study.

We resorted to OpenWordnet-PT in the LMF format (Vossen et al., 2013), whose last release in this format we found is from October 2018. To reuse the scripts ready for the conversion to wordnet2vec, we converted this LMF format into a Princeton WNDB format, having retained 54390 words.[4] This conversion was done by iterating over the lexicon and keeping track of lexical entries and their lemmas and senses, according to a unique id to differentiate between them, and also keeping a log of the semantic relations between synsets. Only two semantic relations present in OWN-PT are not represented in the final converted network, due to them not being present in the Princeton WNDB format[5]. Those two semantic relations are ”exemplifies” and ”is_exemplified_by”.[6]

For the sake of comparability, and given the different sizes of the two WordNets for Portuguese, three experiments were performed with OWN-PT.

In a first experiment, the 54390 words of OWN-PT in the WNDB format were used.

In a second experiment, a subset of OWN-PT was selected with the same number of words of MWN.PT (15886). The words that are common to both WordNets were selected. Given that not all of the MWN.PT words exist in the OWN-PT, further words were selected from OWN-PT to attain the aimed dimension. Remaining synsets were ordered from the ones with more outgoing edges to less outgoing edges and the words from the more connected synsets were selected until the intended dimension was reached. In previous experiments with English (Saedi et al., 2018), it became apparent that selecting words from synsets with more outgoing edges leads to semantic spaces with better performance in the semantic similarity task.

In a third experiment, a subset of equal dimension to the MWN.PT set was again extracted, this time with the simpler methodology of the second part of the selection undertaken in the second experiment: synsets were ordered from the ones with more to less outgoing edges and the words from the more connected synsets were selected until the intended dimension was reached.

Table 2 presents the scores obtained in these experiments.

4 Discussion

The result of these experiments with MWN.PT is in line with the results of the experiments with English (Saedi et al., 2018), even though now the experiment was with another language and with WordNets that are quite different in dimension and coverage than the English one. When evaluated in the semantic similarity task with a mainstream test dataset, the semantic space obtained from a concept-based semantic network with wnet2vec methodology outperforms the strong baseline consisting of a semantic space obtained from mainstream corpus-based methods, namely with word2vec trained with a very large collection of text, with 2.2B tokens in the present case.

The results of the subsequent experiments with OWN-PT are also in line with those findings. Even though it was built with a methodology resorting to heuristics and (semi-)automatics methods, the semantic space obtained from a second concept-based semantic network of Portuguese with wnet2vec methodology also outperforms the same strong baseline.

[4] https://wordnet.princeton.edu/documentation/wndb5wn

[5] All semantic relations from Princeton WNDB (`https://wordnet.princeton.edu/documentation/wninput5wn`) were resorted to

[6] This script is available from `https://github.com/nlx-group/WordNet-Format-Conversion`

WordNet	Similarity	Words
MWN.PT	0.4643	15886
OWN-PT All words (1st experiment)	0.3124	54390
OWN-PT Same size, common words w/ MWN.PT (2nd exp.)	0.4060	15886
OWN-PT Same size, synsets w/ more relations (3rd exp.)	0.4020	15886

Table 2: Performance of the models obtained from the conversion of MWN.PT and OWN-PT WordNets over LX-SimLex-999 given by Spearman's coefficient (higher score is better).

In this connection, we offer the observation that when a subset of the English WordNet was experimented using a number of synsets (25k) that is closer to our experiments reported here, a 0.45 score was obtained (against 0.53 with 60k synsets) (Saedi et al., 2018). This may indicate that improving the existing WordNets of Portuguese with a larger number of lexical units and relations may bring even better performance.

Additionally, the results of the experiments reported above suggest that when using the wnet2vec methodology to obtain a semantic space from a semantic graph, under comparable experimental circumstances (i.e. over 15k words that hold higher number of relations), 15% better semantic similarity performance scores are obtained with a manually crafted WordNet — 0.46 with MWN.PT — than with a WordNet obtained (semi-) automatically — 0.40 with OWN-PT.

This is in line with what is expected given the noise introduced by the (semi-)automatic methods used in the construction of WordNets. What is new with respect to the methodology proposed here is that there is now a quantitative way to assess the difference between WordNets in what concerns their different quality at capturing the semantics of words.

5 Related work

A proposal for the conversion from the Princeton WordNet to a semantic space different from the one used here can be found in (Goikoetxea et al., 2015). That is different in that in this other proposal the conversion from semantic graph to semantic space is not direct. First, a synthetic corpus is generated by a random walk in the WordNet. Then on the basis of that artificial text, common corpus-based techniques are used to obtain the word embeddings.

In (Gonçalo Oliveira, 2018), in turn, another approach was used to obtain a semantic space from Portuguese semantic networks also resorting to a random walk but, differently from the approach mentioned above, via direct conversion. Instead of using a concept-based semantic network (WordNet) as in our study reported here, semantic networks based on words only (no synsets) were used and converted to semantic spaces. Also, a different method than ours was used, a random walk with 30 iterations.

Although these differences render the results not comparable, it may be still interesting to draft some observations with the necessary caution and grains of salt. The best accuracy score reported in (Gonçalo Oliveira, 2018) with LX-SimLex-999 is 0.61 in terms of Spearman's coefficient. This score is obtained with a network with more than 200k words, more than ten times larger than the network used in our study reported here, with approximately 17k synsets.

The system that, in turn, is reported there as having a performance score of 0.45, in line with the score of 0.46 we found here for a 17k network, was trained over a network five times larger than the one used here. This may be another sign of the higher quality of (hand-crafted) WordNets at recording the lexical semantics of words.

In future work, it will be interesting to undertake further experiments to try to understand to what extent the strength of the findings reported here are due to intrinsic strength of the conversion algorithm adopted here or to the intrinsic quality of the semantic networks used, or just of a bit of both factors and of their combination.

6 Conclusions

In a previous study in the literature (Saedi et al., 2018), a conversion method (wnet2vec) was explored to obtain a semantic space (aka word embeddings) from a semantic graph, by applying it to the English Princeton WordNet. The WordNet embeddings wnet2vec thus generated, on the basis of 60k synsets, outperforms a strong baseline which is a corpus-based word embedding word2vec, based on 100B words. It outperforms in the semantic similarity detection task over the mainstream SimLex-999 test dataset, with an accuracy score of 0.50 against 0.44 in terms of Spearman's coefficient, for wnet2vec and word2vec respectively .

In the present paper, we experimented with this conversion method under further empirical conditions. We applied it over a WordNet manually built under the same construction principles as Princeton WordNet (over 120k synsets) only that it is more than seven times smaller (17k synsets) and is for another language, namely Portuguese. We experimented also with another WordNet for Portuguese but constructed under an alternative approach that resorts to (semi-) automatic methods.

The WordNet embeddings obtained were tested under the semantic similarity task over the Portuguese translation of the mainstream SimLex-999 test dataset (Querido et al., 2017). The baseline was the word embeddings obtained with the corpus-based word2vec procedure over a 2.2B words corpus of Portuguese (Rodrigues and Branco, 2018).

The results obtained are in line with earlier findings. The wnet2vec conversion method to obtain a semantic space from a semantic network is very effective.

The semantic similarity detectors based on word embeddings wnet2vec — obtained from any of the WordNets experimented with in this paper — outperform the strong baseline detector based on the corpus-based word embeddings word2vec.

The semantic similarity detector based on the manually built WordNet, in turn, outperformed the detector based on the WordNet that was built (semi-) automatically. These results suggest that, when using the wnet2vec methodology to obtain a semantic space from a semantic graph, under comparable experimental circumstances, better semantic similarity performance scores are obtained with a manually crafted WordNet rather than with a WordNet obtained (semi-) automatically. This is as expected given the noise introduced by automatic methods. What is new with respect to the assessment methodology proposed here is that it offers a new quantitative way to evaluate the difference between WordNets in what concerns their different quality at capturing the meaning of words.

7 Acknowledgments

The present research was partly supported by the Infrastructure for the Science and Technology of Language (PORTULAN CLARIN) by the National Infrastructure for Distributed Computing (INCD) of Portugal, and by the ANI/3279/2016 grant.

References

[Bobrow and Norman1975] Daniel G. Bobrow and Donald Arthur Norman. 1975. Some principles of memory schemata. In *Representation and Understanding: Studies in Cognitive Science*, page 131–149. Elsevier.

[De Deyne et al.2013] Simon De Deyne, Daniel J Navarro, and Gert Storms. 2013. Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. *Behavior Research Methods*, 45(2):480–498.

[de Paiva et al.2012] Valeria de Paiva, Alexandre Rademaker, and Gerard de Melo. 2012. Openwordnet-pt: An open Brazilian Wordnet for reasoning. In *Proceedings of COLING 2012: Demonstration Papers*, pages 353–360, Mumbai, India, December. The COLING 2012 Organizing Committee. Published also as Techreport http://hdl.handle.net/10438/10274.

[Fellbaum1998] Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

[Goikoetxea et al.2015] Josu Goikoetxea, Aitor Soroa, and Eneko Agirre. 2015. Random walks and neural network language models on knowledge bases. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics and Human Language Technologies Conference (NAACL-HLT25)*, pages 1434–1439. Association for Computational Linguistics.

[Gonçalo Oliveira2018] Hugo Gonçalo Oliveira. 2018. Distributional and knowledge-based approaches for computing portuguese word similarity. *Information*, 9(2):35.

[Harris1954] Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

[Hill et al.2016] Felix Hill, Roi Reichart, and Anna Korhonen. 2016. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41:665–695.

[Mikolov et al.2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Googlenews-vectors-negative300.bin.gz - efficient estimation of word representations in vector space. *arXiv preprint* *arXiv:1301.3781*. https://code.google.com/archive/p/word2vec/.

[Minsky1975] Marvin Minsky. 1975. A framework for representing knowledge. In *Psychology of Computer Vision*. McGraw-Hill.

[Osgood et al.1957] Charles E Osgood, George J Suci, and Percy H Tannenbaum. 1957. The measurement of meaning. *Urbana: University of Illinois Press*.

[Querido et al.2017] Andreia Querido, Rita de Carvalho, João Rodrigues, Marcos Garcia, Catarina Correia, Nuno Rendeiro, Rita Valadas Pereira, Marisa Campos, João Silva, and António Branco. 2017. LX-LR4DistSemEval: A collection of language resources for the evaluation of distributional semantic models of Portuguese. *Revista da Associação Portuguesa de Linguística*, 3.

[Quillan1966] M Ross Quillan. 1966. Semantic memory. Technical report, Bolt Beranek and Newman Inc., Cambridge MA.

[Rodrigues and Branco2018] João Rodrigues and António Branco. 2018. Finely tuned, 2billion token based word embeddings for portuguese. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*.

[Saedi et al.2018] Chakaveh Saedi, António Branco, João António Rodrigues, and João Silva. 2018. Wordnet embeddings. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 122–131. Association for Computational Linguistics.

[Vossen et al.2013] Piek Vossen, Claudia Soria, and Monica Monachini. 2013. Wordnet-lmf: A standard representation for multilingual wordnets. *LMF Lexical Markup Framework*, pages 51–66.