


Evaluation of Distributional Models with the Outlier Detection Task

Pablo Gamallo

Centro de Investigación en Tecnologías da Información (CiTIUS)

University of Santiago de Compostela, Galiza

pablo.gamallo@usc.es

 <https://orcid.org/0000-0002-5819-2469>

Abstract

In this article, we define the outlier detection task and use it to compare neural-based word embeddings with transparent count-based distributional representations. Using the English Wikipedia as text source to train the models, we observed that embeddings outperform count-based representations when their contexts are made up of bag-of-words. However, there are no sharp differences between the two models if the word contexts are defined as syntactic dependencies. In general, syntax-based models tend to perform better than those based on bag-of-words for this specific task. Similar experiments were carried out for Portuguese with similar results. The test datasets we have created for outlier detection task in English and Portuguese are released.

2012 ACM Subject Classification Computing methodologies → Unsupervised learning

Keywords and phrases distributional semantics, dependency analysis, outlier detection, similarity

Digital Object Identifier 10.4230/OASICS.SLATE.2018.13

Funding This work was supported by a 2016 BBVA Foundation Grant for Researchers and Cultural Creators, and by Project TELEPARES, Ministry of Economy and Competitiveness (FFI2014-51978-C2-1-R). It has received financial support from the Consellería de Cultura, Educación e Ordenación Universitaria (accreditation 2016-2019, ED431G/08) and the European Regional Development Fund (ERDF).

1 Introduction

Intrinsic evaluations of distributional models are based on word similarity tasks. The most popular intrinsic evaluation is to calculate the correlation between the similarity scores obtained by a system using a word vector representation and a gold standard of human-assigned similarity scores. Recent critics to intrinsic evaluation claim that inter-annotator agreement at the word similarity task is considerably lower compared to other tasks such as document classification or textual entailment [3]. To overcome this problem, Camacho-Collados and Navigli [7] propose an alternative evaluation relying on the outlier detection task, which tests the capability of vector space models to create semantic clusters. More precisely, given a set of words, for instance *car*, *train*, *bus*, *apple*, *bike*, the goal of the task is to identify the word that does not belong to a semantically homogeneous group. In this case, the outlier is *apple*, which is not a vehicle. The main advantage of this task is to provide a clear gold standard with, at least, two properties: high inter-annotator agreement and easy method to increase the test size by adding new groups.

On the other hand, recent works comparing count-based word distributions with word embeddings (i.e., dense representations obtained with neural networks) to compute word similarity show mixed results. Some claim that embeddings outperform transparent and



© Pablo Gamallo;

licensed under Creative Commons License CC-BY

7th Symposium on Languages, Applications and Technologies (SLATE 2018).

Editors: Pedro Rangel Henriques, José Paulo Leal, António Leitão, and Xavier Gómez Guinovart

Article No. 13; pp. 13:1–13:8



OpenAccess Series in Informatics

OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

explicit count-based models [2, 24], while others show that there are no significant differences between them [20, 23], in particular if the hyperparameters are configured and set in a similar way [22]. Other works report heterogeneous results since the performance of the two models varies according to the task to be evaluated [5, 9, 14, 19].

In this paper, we make use of the outlier detection task defined in Camacho-Collados and Navigli [7] to compare different types of embeddings and count-based word representations. In particular, we compare the use of bag-of-words and syntactic dependencies in both embeddings and count-based models. We observed that there are no clear differences if the two models rely on syntactic dependencies; yet embeddings seem to perform better than transparent models when the dimensions are made up of bag-of-words. In addition, we contribute to enlarge the test dataset by adding more groups of semantically homogeneous words and outliers (50% larger), and by manually translating the expanded dataset to Portuguese.

Next, we will describe the outlier datasets (Section 2), a count-based model with filtering (Section 3), and several experiments to compare different distributional models using the outlier datasets (Section 4). Conclusions will be addressed in Section 5.

2 The Datasets for Outlier Detection

For the outlier detection task, Camacho-Collados and Navigli [7] provided the *8-8-8* dataset¹, which consists of eight different topics each containing a cluster of eight words and eight outliers which do not belong to the given topic. For instance, one of the topics is “European football teams”, which was defined with a set of eight nouns:

FC Barcelona, Bayern Munich, Real Madrid, AC Milan, Juventus, Atletico Madrid, Chelsea, Borussia Dortmund

and a set of eight outliers:

Miami Dolphins, McLaren, Los Angeles Lakers, Bundesliga, football, goal, couch, fridge

In order to expand the number of examples, two annotators were asked to create four new topics, and for each topic to provide a set of eight words belonging to the chosen topic, and a set of eight heterogeneous outliers. One of them proposed all the words in less than 15 minutes, and the other annotator just agreed with all the decisions taken by the first one. This 100% inter-annotator agreement is in contrast with the low inter-annotator levels achieved in the standard word similarity datasets, for instance in WordSim353 [10] the average pair-wise Spearman correlation among annotators is merely 0.61. The new expanded dataset, called *12-8-8*, contains 12 topics, each made up of 8+8 topic words and outliers. In addition, in order to simplify the comparison with systems that do not identify multiwords, we also changed the multi-words found in the *8-8-8* dataset by one-word terms denoting similar entities. For instance: the terms “Celtic” and “Betis” were used instead of “Atletico Madrid” and “Bayern Munich”, all referring to football teams. The *12-8-8* dataset contains 50% more test examples than the original one. Finally, we also created a new dataset by translating *12-8-8* into Portuguese.

In Camacho-Collados and Navigli [7], the outlier detection task is defined on the basis of a generic concept of *compactness score*. Here, we propose to define a more specific *compactness score* by assuming that the similarity coefficient is symmetrical (e.g. Cosine). Intuitively,

¹ <http://lcl.uniroma1.it/outlier-detection/>

given a set of 9 words consisting of 8 words belonging to the same group and one outlier, the *compactness score* of each word of the set is the result of averaging the pair-wise similarity scores of the target word with the other members of the set.

Formally, given a set of words $W = w_1, w_2, \dots, w_n, w_{n+1}$, where w_1, w_2, \dots, w_n belongs to the same cluster and w_{n+1} is the outlier, we define the compactness score $c(w)$ of a word $w \in W$, and assuming a symmetrical similarity coefficient sim , as follows:

$$c(w) = \frac{1}{n} \sum_{\substack{w_i \in W \\ w \neq w_i}} sim(w, w_i) \quad (1)$$

An outlier is correctly detected if the compactness score of the outlier word is lower than the scores of the cluster words. So, *accuracy* measures how many outliers were correctly detected by the system divided by the number of total detections: 12x8 in our *12-8-8* dataset. Camacho-Collados and Navigli [7] also define Outlier Position Percentage (OPP) which takes into account the position of the outlier in the list of $n + 1$ words ranked by the compactness score, which ranges from 0 to n (position 0 indicates the lowest overall score among all words in W , and position n indicates the highest overall score).

3 A Filtered-Based Distributional Model

The outlier datasets will be used to compare count-based distributional models with embeddings. The count-based model we propose is based on a filtering approach and dependency contexts.

As co-occurrence matrices representing context distribution are sparse, most entries of a sparse matrix are zeros that do not need to be stored explicitly. In fact, highly dispersed matrices are computationally easy to work with [9, 16]. A possible storage mode for a sparse matrix is a hash table where keys are word-context pairs with non-zero values [16]. To reduce the number of keys in a hash table representing word-context co-occurrences, we apply a technique to filter out contexts by relevance [6]. The compressing technique consists in computing an *informativeness* measure -e.g., loglikelihood [8]- between each word and their contexts. For each word, only the R (relevant) contexts with highest loglikelihood scores are kept in the hash table. R is a global, arbitrarily defined constant whose usual values range from 10 to 1000 [4, 26]. In short, we keep the R most relevant contexts for each target word. Context filtering allows us to dramatically reduce the context space and, unlike embeddings, makes the word model transparent, fully interpretable and easily readable by humans.

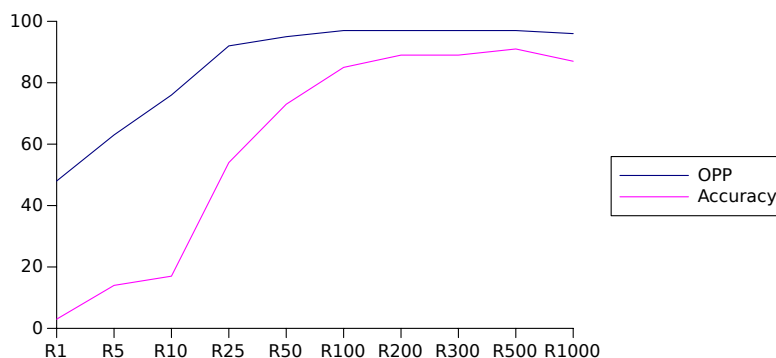
Syntactic-based word contexts can be derived from the dependency relations the words participate in (e.g. subject, direct object, modifier). To extract contexts from dependencies, we use the co-compositional methodology defined in Levy and Goldberg [21] and Gamallo [15]. Notice that syntax-based models are fully interpretable as each dimension is an explicit lexico-syntactic context.

4 Experiments and Evaluation

We performed three experiments. The first one using the original *8-8-8* dataset. The second one comparing more approaches against the expanded *12-8-8* dataset. And the third one comparing the best approaches of the previous experiments using the Portuguese *12-8-8* dataset.

■ **Table 1** Outlier Position Percentage (OPP) and Accuracy of different word models on the 8-8-8 outlier detection dataset using Wikipedia.

<i>System</i>	<i>Strategy</i>	<i>OPP</i>	<i>Accuracy</i>
Dep500	count+syntax	97.3	90.6
CBOW	embed+bow	95.3	73.4
Skip-Gram	embed+bow	93.8	70.3
Glove	embed+bow	91.8	56.3



■ **Figure 1** Accuracy and OPP of our count-based strategy across different filtering thresholds: from $R = 1$ to $R = 1000$.

4.1 The 8-8-8 Dataset

The goal of the experiment is to compare the basic count-based model defined in the previous section (3) with the results obtained by different versions of embeddings, which were reported in Camacho-Collados and Navigli [7].

Table 1 shows the results obtained by the count-based strategy we developed, *Dep500*, which is a count-based model with contexts represented as syntactic dependencies and a relevance filter $R = 500$. The contexts of the model were built by making use of a rule-based dependency parser, *DepPattern* [13]. The method outperforms the results obtained by three standard embedding models: the CBOW and Skip-Gram models of Word2Vec [24] and GloVe [28], which are based on bag-of-words contexts², and whose results were reported in Camacho-Collados and Navigli [7]. The dimensionality of the dense vectors was set to 300 for the three embedding models. Context-size 5 for CBOW and 10 for Skip-Gram and GloVe; hierarchical softmax for CBOW and negative sampling for Skip-Gram and GloVe. In all experiments, the corpus used to build the vector space was the 1.7B-tokens English Wikipedia (dump of November 2014).

The growth curve depicted in Figure 1 shows the evolution of accuracy and OPP over different R values. We can observe that the curve stabilizes at $R = 200$ and starts going down before $R = 1000$. It means that small count-based distributional models with relevant contexts perform better than large models made up of many noisy syntactic contexts.

² We use *bow* to refer to linear bag-of-word contexts, which must be distinguished from CBOW (continuous bag-of-words) [22]

4.2 The 12-8-8 Expanded Dataset

The main goal of the next experiments is to use the outlier detection task to compare the performance of different types of dependency parsers (rule-based and transition-based) to build both count-based distributions and neural embeddings. Additionally, we also compare the use of syntactic dependencies and bag-of-words in the same task. We require a dataset without multiwords since some of the tools we used for building distributions only tokenize unigrams. For this purpose, we defined the following six strategies:

Count₁ A count-based model with rule-based dependencies.

Count₂ A count-based model with transition-based dependencies.

Count₃ A count-based model with bag-of-words.

Emb₁ Embeddings with rule-based dependencies.

Emb₂ Embeddings with transition-based dependencies.

Emb₃ Embeddings with bag-of-words.

The three count-based models were built with the filter $R = 300$, whereas the dimensionality of the three embeddings was set to 300. The three embeddings were based on the continuous *skip-gram* neural embedding model [24], with negative-sampling parameter set at 15. The two bag-of-words models were generated using a window of size 10: 5 words to the left and 5 to the right of the target word. Both Emb₂ and Emb₃ are the models described in Levy and Goldberg [21], which are publicly available³. To create the dependency-based models, the corpus was parsed with a very specific configuration of the arc-eager transition-based dependency parser described in [17]⁴. The performance of the parser for English is about 89% UAS (unlabeled attachment score) obtained on the CoNLL 2007 dataset. To build Emb₂, we made use of `word2vecf`⁵, a modified version of `word2vec`, which is suited to build embeddings with syntactic dependencies [21]. Rule-based dependencies were obtained using `DepPattern` [13].

Even though the strategies are very different using very different software, we tried to use the same hyperparameters in order to minimize differences that are not due to the word models themselves. As Levy and Goldberg [23] suggest, much of the difference between vectorial models are due to certain system design choices and hyperparameter optimizations (e.g., subsampling frequent words, window size, etc.) rather than to the algorithms themselves. The authors revealed that seemingly minor variations in external parameters can have a large impact on the success of word representation methods.

Table 2 shows the results obtained on the *12-8-8* dataset by the six models built from the English Wikipedia. The four syntax-based methods (with rules or transitions, count-based or embeddings) give very similar scores. However, they tend to perform better than those based on bag-of-words (as in the previous experiment in Subsection 4.1). This is in accordance with a great number of previous works which evaluate and compare syntactic contexts (usually dependencies) with bag-of-words techniques [11, 12, 18, 21, 25, 27, 29]. All of them state that syntax-based methods outperform bag-of-words techniques, in particular when the objective is to compute semantic similarity between functional (or paradigmatic) equivalent words, such as detection of co-hyponym/hypernym word relations. By contrast, *bow*-based models tend to perform better in tasks oriented to identify semantic relatedness

³ <https://levyomer.wordpress.com/2014/04/25/dependency-based-word-embeddings/>

⁴ We are very grateful to the authors for sending us the English Wikipedia syntactically analyzed with their parser.

⁵ <https://bitbucket.org/yoavgo/word2vecf>

■ **Table 2** Outlier Position Percentage (OPP) and Accuracy of different word models on the 12-8-8 outlier detection dataset using Wikipedia.

<i>System</i>	<i>Strategy</i>	<i>OPP</i>	<i>Accuracy</i>
Count ₁	rules	94.92	75.0
Count ₂	transitions	93.48	71.87
Count ₃	bow	86.71	60.41
Emb ₁	rules	93.09	76.04
Emb ₂	transitions	94.27	72.91
Emb ₃	bow	93.88	69.79

■ **Table 3** Outlier Position Percentage (OPP) and Accuracy of two distributional models on the 12-8-8 outlier detection dataset using Portuguese Wikipedia.

<i>System</i>	<i>Strategy</i>	<i>OPP</i>	<i>Accuracy</i>
Count ₁	rules	91.40	48.95
Emb ₁	rules	84.375	39.58

and analogies. We may conclude the following: First, the outlier detection task is suited to search for similarity and not for semantic relatedness [1], and second, the type of context (dependency-based or bag-of-words) is more determinant than the type of model (count-based or embeddings) for that task. Finally, embeddings clearly outperform count-based representations when the contexts are defined with bag-of-words (see score of Emb₃ against Count₃ in Table 2).

4.3 Portuguese 12-8-8 Dataset

The 12-8-8 Expanded Dataset was translated into Portuguese in order to make new tests in this language. The Portuguese experiments were carried out with the two best strategies, according to the previous experiments: count-based model with rule-based dependencies (Count₁) and embeddings with rule-based dependencies (Emb₁). As in the previous experiment, the count-based model was built with the filter $R = 300$, whereas the dimensionality of the embeddings was set to 300. The latter was implemented with *skip-gram* and negative-sampling parameter set at 15. Table 3 shows the results obtained on the Portuguese 12-8-8 dataset by the two models evaluated.

In these experiments, the count-based strategy clearly outperforms embeddings. This may be partially explained by the fact that the Portuguese Wikipedia is almost 10 times smaller than the English one, and neural networks require large corpus to make better predictions.

5 Conclusions

We have used the outlier detection task for intrinsic evaluation of distributional models in English and Portuguese. Unlike standard gold-standards for similarity tasks, the construction of datasets for outlier detection requires low human cost with very high inter-annotator agreement. Our very preliminary experiments show that the use of syntactic contexts in traditional count-based models and embeddings leads the two models to similar performance on the outlier detection task, even if count-based strategies seem to perform better with less training corpus.

In future work, we intend to develop outlier detection datasets for many other languages in order to make it possible multilingual word similarity evaluation. The software required to build the count-based models as well as the 12-8-8 datasets are publicly available⁶.

References

- 1 Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27, 2009.
- 2 Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 238–247, 2014.
- 3 Miroslav Batchkarov, Thomas Kober, Jeremy Reffin, Julie Weeds, and David Weir. A critique of word similarity as a method for evaluating distributional semantic models. In *Proceedings of the ACL Workshop on Evaluating Vector Space Representations for NLP*, pages 7–12, 2016.
- 4 Biemann, C., and Riedl M. Text: Now in 2D! A framework for lexical expansion with contextual similarity. *Journal of Language Modelling*, 1(1):55–95, 2013.
- 5 William Blacoe and Mirella Lapata. A comparison of vector-based representations for semantic composition. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 546–556, 2012.
- 6 Stefan Bordag. A comparison of co-occurrence and similarity measures as simulations of context. In *Computational Linguistics and Intelligent Text Processing (CICLing)*, pages 52–63, 2008.
- 7 José Camacho-Collados and Roberto Navigli. Find the word that does not belong: A framework for an intrinsic evaluation of word vector representations. In *Proceedings of the ACL Workshop on Evaluating Vector Space Representations for NLP*, pages 43–50, 2016.
- 8 Ted Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, 1993.
- 9 Manaal Faruqui and Chris Dyer. Non-distributional word vector representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 464–469, 2015.
- 10 Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppín. Placing search in context: the concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131, 2002.
- 11 Pablo Gamallo. Comparing window and syntax based strategies for semantic extraction. In *Computational processing of the Portuguese language*, pages 41–50, 2008.
- 12 Pablo Gamallo. Comparing different properties involved in word similarity extraction. In *14th Portuguese Conference on Artificial Intelligence (EPIA '09)*, pages 634–645, 2009.
- 13 Pablo Gamallo. Dependency parsing with compression rules. In *Proceedings of the 14th International Workshop on Parsing Technology (IWPT)*, pages 107–117, 2015.
- 14 Pablo Gamallo. Comparing explicit and predictive distributional semantic models endowed with syntactic contexts. *Language Resources and Evaluation*, 51(3):727–743, 2017.
- 15 Pablo Gamallo, Alexandre Agustini, and Gabriel Lopes. Clustering syntactic positions with similar semantic requirements. *Computational Linguistics*, 31(1):107–146, 2005.

⁶ <http://gramatica.usc.es/~gamallo/prototypes/Word2Model.tgz>

- 16 Pablo Gamallo and Stefan Bordag. Is singular value decomposition useful for word similarity extraction. *Language Resources and Evaluation*, 45(2):95–119, 2011.
- 17 Yoav Goldberg and Joakim Nivre. A dynamic oracle for arc-eager dependency parsing. In *24th International Conference on Computational Linguistics Proceedings of the Conference (COLING)*, pages 959–976, 2012.
- 18 Gregory Grefenstette. Evaluation techniques for automatic semantic extraction: Comparing syntactic and window-based approaches. In *Workshop on Acquisition of Lexical Knowledge from Text (SIGLEX)*, pages 205–216, 1993.
- 19 Eric Huang, Richard Socher, and Christopher Manning. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 873–882, 2012.
- 20 Rémi Lebret and Ronan Collobert. Rehabilitation of count-based models for word vector representations. In *Computational Linguistics and Intelligent Text Processing (CICLing)*, volume 9041, pages 417–429, 2015.
- 21 Omer Levy and Yoav Goldberg. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 302–308, 2014.
- 22 Omer Levy and Yoav Goldberg. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning (CoNLL)*, pages 171–180, 2014.
- 23 Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225, 2015.
- 24 Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, 2013.
- 25 Sebastian Padó and Mirella Lapata. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199, 2007.
- 26 Muntsa Padró, Marco Idiart, Aline Villavicencio, and Carlos Ramisch. Nothing like good old frequency: Studying context filters for distributional thesauri. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 419–424, 2014.
- 27 Yves Peirsman, Kris Heylen, and Dirk Speelman. Finding semantically related words in Dutch. Co-occurrences versus syntactic contexts. In *CoSMO Workshop*, pages 9–16, 2007.
- 28 Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.
- 29 Violeta Seretan and Eric Wehrli. Accurate collocation extraction using a multilingual parser. In *21st International Conference on Computational Linguistics*, pages 953–960, 2006.