machine learning & knowledge extraction

MDPI

Article

Using the Outlier Detection Task to Evaluate Distributional Semantic Models

Pablo Gamallo

Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS), Campus Vida, Universidade de Santiago de Compostela, 15782 Santiago de Compostela, Galiza, Spain; pablo.gamallo@usc.es; Tel.: +34-881816426

Received: 2 August 2018; Accepted: 19 November 2018; Published: 22 November 2018

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 a 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 the outlier detection task in English and Portuguese are freely available.

Keywords: distributional semantics; dependency analysis; evaluation; word similarity

1. Introduction

This article is an expanded version of a conference paper presented at SLATE-2018 [1].

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 in the word similarity task is considerably lower compared to other tasks such as document classification or textual entailment [2]. To overcome this problem, Camacho-Collados and Navigli [3] proposed 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 an easy method to increase the test size by adding new groups/clusters.

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 explicit count-based models [4,5], while others show that there are no significant differences between them [6,7], in particular if the hyperparameters are configured and set in a similar way [8]. Other works report heterogeneous results since the performance of the two models varies according to the task to be evaluated [9–12].

In this paper, we make use of the outlier detection task defined in Camacho-Collados and Navigli [3] 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 word 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 order to build distributional models with syntactic dependencies, we took into account the strategies defined in Levy and Goldberg [13] and Gamallo [11].

In addition, we contribute to enlarging 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 different intrinsic evaluation methods and their drawbacks (Section 2), the outlier detection method and their datasets (Section 3), as well as a dependency-based model with filtering of relevant contexts (Section 4). Several experiments to compare different distributional models using the outlier datasets will be described in Section 5. Conclusions will be addressed in Section 6.

2. Intrinsic Evaluation of Distributional Models

So far, the most popular intrinsic methods to evaluate distributional models are mainly two: those based on correlation with human similarity and those simulating the classic TOEFL test for learning English as a second language.

2.1. Correlation with Human Similarity

This evaluation method consists of constructing datasets by asking human subjects to rate the degree of semantic similarity or relatedness between two words on a numerical scale. The performance of a computational system is measured in terms of correlation between the scores assigned by humans to the word pairs and the similarity coefficient (cosine, dice, etc.) assigned by the system taking into account the model space. Correlation can be computed with Pearson or Spearman, or even by a mean between the two as in the Cross-Lingual Word Similarity Task at SemEval 2017 [14]. One of the most widely-used datasets is WordSim353 [15], which, as the name suggests, consists of 353 word pairs.

However, the concept of semantic similarity or relatedness is not at all obvious for many word pairs. Table 1 shows a sample of Wordsim353 word pairs with the assigned human score: from zero (totally different) to 10 (totally identical). In some pairs, the relationship or the lack of relationship is very evident, as in the examples *company-stock*, which are seen as very related (7.08), and *stock-jaguar*, which are perceived as not related (0.92). By contrast, there are other cases where the relation is not so clear, which can lead to inconsistencies in the score annotation. For instance, in Table 1, the pair *stock-oil* is perceived as very related (6.34), but *stock-egg* or *stock-CD* is almost not related (1.81 and 1.31, respectively). These inconsistencies are probably due to attention imbalances between different individuals with respect to polysemy. Most words are polysemous, and humans are not able to activate the appropriate sense if the linguistic context is not fully determined. In the case of *stock*, the annotator seems to activate the merchandise sense when it is compared to *oil* (its context word), but this is not the case when it is paired with *egg* or *CD*. There are probably psycholinguistic factors that explain these seemingly inconsistent scorer decisions. In our opinion, these datasets might be useful to measure psycholinguistic issues related to word meaning, but they may be too subjective to evaluate linguistic models of meaning.

Agirre et al. [16] split WordSim353 into two subsets: pairs measured by degree of synonymy (e.g., tiger/cat are close because their referents are similar) and pairs rated by topical relations (e.g., planet/astronomer). However, the problems stated above still persist.

A more recent dataset built with the same methodology is SimLex-999 [17], containing 999 word pairs. Unlike WordSim353, it was aimed to capture degrees of synonymy, independent of topicality. It is worth noticing that the authors of SimLex-999 used the term *similarity* for what we call synonymy and *relatedness/association* for topicality. To better understand these terms, some examples might help. On the one hand, degrees of synonymy include hypernym relations (e.g., *astronaut/person*) and co-hyponyms (e.g., *astronaut/scientist*). On the other hand, topicality and relatedness include any other type of semantic link (e.g., *astronaut/moon*).

Table 1. Correlation with human similarity: sample of the Wordsim353 dataset.

company	*stock*	7.08
stock	*jaguar*	0.92
stock	*egg*	1.81
fertility	*egg*	6.69
stock	*life*	0.92
stock	*live*	3.73
oil	*stock*	6.34
CD	*stock*	1.31

The same types of datasets are also proposed for evaluating cross-lingual models [14], as well as compositional models with intransitive [18] and transitive constructions [19].

2.2. TOEFL-Style Tests

The classic TOEFL (toefl) set, introduced by Landauer and Dumais [20], is constituted by 80 multiple-choice questions that put in relation a target word with our candidates. The four candidates are near-synonyms, but only one of them (the solution) is also a partial synonym of the target word. For example, for the target word *levied*, the learner must choose between *imposed*, which is the correct answer (or solution), and another three words: *believed*, *requested*, and *correlated*:

Target:	*levied*
Choice (a):	*imposed*
Choice (b):	*believed*
Choice (c):	*requested*
Choice (d):	*correlated*
Solution:	(a) *imposed*

To evaluate the performance of a system, it must compute any similarity measure of each candidate vector with the target word and pick the candidate with the highest score as its answer. Performance is evaluated in terms of correct-answer accuracy. The main drawback of this task is the difficulty of creating new datasets, since it requires linguists (not just human annotators) to prepare them. This means that only small datasets that are not very representative are available. In addition, given the characteristics of the questionnaires, which are very dependent on the language, they cannot be easily translated to other languages

There are other TOEFL-style benchmarks. ESL, constituted by 50 questions from the English as a Second Language test [21], is a very similar dataset to TOEFL, also used to evaluate distributional models, and B2SGis a TOEFL-like dataset for Portuguese language [22].

3. The Outlier Detection Task

The outlier detection task is based on standard vocabulary techniques to learn new words and in specific lexical questionnaires of language exams [23]. More precisely, given a group of words, the objective of the task is to identify the word that does not belong to the group. This task is oriented to separate synonyms and co-hyponyms (i.e., clusters of words with similar referents) from those that are neither synonyms nor co-hyponyms. Thus, the task is not suited to identify topical relationships.

3.1. The Compactness Score

In Camacho-Collados and Navigli [3], the outlier detection task was defined on the basis of a generic concept of *compactness score* that considers both symmetrical and asymmetrical measures. Here, we propose to define a more specific *compactness score* by assuming that the similarity/distance coefficient must be symmetrical (e.g., cosine). Intuitively, given a set of nine words consisting of eight 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 $W = w_1, w_2, \ldots, w_n, w_{n+1}$, where $w_1, w_2, \ldots, w_n$ belongs to the same cluster and w_{n+1} is the outlier, we define the compactness score $c(w)$ of a word $w \in W$, assuming a symmetrical similarity coefficient *sim*, as follows:

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

An outlier is correctly detected if the compactness score of the outlier word is lower than the scores of the cluster words. Camacho-Collados and Navigli [3] defined two evaluation coefficients to measure the degree of correction: *Outlier Position Percentage* (*OPP*) and *accuracy*. The former relies on the *Outlier Position* (*OP*), which takes into account the position of the outlier in the set W of $n+1$ words ranked by the compactness score, which ranges from zero to n (0 indicates the lowest overall score among all words in W, and n indicates the highest overall score). To compute the overall score on a dataset D (composed of $|D|$ sets of words), OPP is defined as follows:

$$OPP = \frac{\sum_{W \in D} \frac{OP(W)}{|W|-1}}{|D|} \tag{2}$$

On the other hand, Camacho-Collados and Navigli [3] also defined *accuracy*, which measures how many outliers were correctly detected by the system divided by the number of total detections: 12×8 in our *12-8-8* dataset. More formally, given *Outlier Detection* (*OD*), defined as 1 if the outlier is correctly detected (0 otherwise), accuracy is defined as follows:

$$Accuracy = \frac{\sum_{W \in D} OD(W)}{|D|} \tag{3}$$

3.2. New Benchmarks for the Outlier Detection Task

For the outlier detection task, Camacho-Collados and Navigli [3] provided the *8-8-8* dataset (http://lcl.uniroma1.it/outlier-detection/), 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 (see Table 2) and a set of eight outliers (Table 3)

Table 2. Eight proper nouns belonging to the class of European football teams in the *8-8-8* dataset.

European football teams	*FC Barcelona*
	Bayern Munich
	Real Madrid
	AC Milan
	Juventus
	Atletico Madrid
	Chelsea
	Borussia Dortmund

To help improve and expand the first dataset, we have developed an extended version. 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 min, and the other annotator just agreed with all the decisions made 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 [15], 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 multi-words, 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.

Table 3. Eight nouns that are not in the class of European football teams (*8-8-8* dataset).

Outliers of European football teams	*Miami Dolphins* *McLaren* *Los Angeles Lakers* *Bundesliga* *football* *goal* *couch* *fridge*

One of the main problems in creating new clusters and outliers is the difficulty of finding a set of words that is quiet similar to the cluster, but which does not belong to it. For instance, take one of the new topics, name of colors, included in the *12-8-8* dataset and shown in the first column in Table 4. There should be no disagreement about belonging to this class. However, to make the outliers' search more challenging, it would be necessary to find words semantically related to the topic without belonging to the class of colors. Some of these words are close hyperonyms such as *color*, *property*, or *substance* (see the second column of Table 4). Annotators were instructed to use at least four or more words semantically closely related to the target topic. In Table 4, the five first outliers (in bold) are semantically related to the class of colors. The more words of this type there are among the outliers, the more complicated the detection task becomes.

Table 4. One of the new topics in expanded *12-8-8* dataset: 8 proper nouns belonging to the class of Colors and 8 that do not belong.

Colors	Outliers
red	***color***
blue	***property***
white	***feature***
green	***substance***
violet	***material***
yellow	*image*
black	*bottle*
rose	*water*

Given the characteristics of the task, it is easy to exploit all kinds of taxonomies in any domain of knowledge, for example zoological knowledge. Table 5 shows the topic of ruminants and their outliers. Notice that the set of outliers also contains very difficult cases, namely the names of other similar animals to ruminants that are not in the zoological category (in bold in the second column).

The outlier detection task is conceptually closer to the TOEFL-style test than to the task relying on correlation with human similarity. However, unlike the TOEFL test, the outlier task is conceived to build new test datasets by non-professional annotators in an easy way. It also allows comparing a word (the outlier) with a larger set of words (not just three or four candidates per target word). The outlier is, in fact, compared against a cluster of words belonging to the same lexical class, e.g., mammals, colors, football teams, prime-ministers, German people, fresh vegetables, or whatever predefined class. This makes it possible to make more word comparisons and, therefore, to increase the coverage of the test.

Table 5. Eight proper nouns belonging to the class of Ruminants and 8 outliers of this class in the *12-8-8* dataset.

Ruminants	Outliers
cow	***fish***
bison	***lion***
bull	***wolf***
buffalo	***mouse***
camel	*grass*
sheep	*glass*
goat	*month*
gazelle	*south*

Finally, as the groups are homogeneous and the word meanings are contextualized, and then are not ambiguous, there is much less subjectivity than in the correlation test. This facilitates and favors inter-annotator agreement. This also makes it easier to translate into other languages, without requiring too much adaptation.

4. A Filtered-Based Distributional Model

One of the objectives of the current paper is to use the outlier datasets to compare count-based distributional models with embeddings. The count-based model we propose is based on a filtering approach and dependency contexts. In the following subsections, our approach is defined.

4.1. Contexts from Syntactic Dependencies

The semantic model is based on extracting contexts from the syntactic dependencies between lemmas (e.g., modifier, prepositional object, nominal subject, adjunct, etc). The extraction of contexts from syntactic dependencies is done by using the co-compositional strategy reported in in [24], as well as more recently in [13]. Given a lemma l linked to a set of dependent lemmas $dep_1, ..., dep_k$ and to one *head* (as in dependency grammar, each lemma only depends on just one head), the following contexts are extracted:

$$(dep_1, \downarrow r_1) ..., (dep_k, \downarrow r_k), (head, \uparrow r_h)$$

where r is a dependency label and $\downarrow r$ represents a unary relation derived from a binary dependency, which requires a dependent lemma. The inverse relation is $\uparrow r$, which stands for a unary relation requiring a head lemma. Therefore, $(dep_i, \downarrow r_i)$ stands for a lexico-syntactic context of lemma l in the head position, while $(head, \uparrow r_h)$ is a lexico-syntactic context of l as dependent. For instance, in *"Jane smiled"*, $(jane, \downarrow subject)$ represents a lexico-syntactic context of the verb *smiled*, whereas $(smile, \uparrow subject)$ stand for a context of *Jane*.

In our model, prepositions are not words, but dependency relations, so they are added to the dependency label. For instance, in *"smiled at Mary"*, $(mary, \downarrow prep\_at)$ is a lexico-syntactic context of *smiled*, whereas $(smile, \uparrow prep\_at)$ stands for the inverse context of *Mary*. According to Baroni and Lenci [25], this type of dependency-based contexts gives rise to the *word by link-word* vector model, where vectors are lemmas (or words) and dimensions are tuples of two elements: a relation label and a lemma.

4.2. Context Filtering

Co-occurrence matrices learned from text corpora are sparse due to the Zipf law distribution. This makes most of the cells in the matrix zeros, which are not required to be stored when using sparse representations. In fact, matrices with a large number of zero values can be stored with computationally-efficient representations [10,26]. One of these storage strategies suited for sparse matrices is comprised of hash tables containing lemma-context pairs with only non-zero values [26].

More precisely, the computational structure we used is a hash of hashes. Keys are lemmas whose values are context-value hashes in which contexts are keys and frequencies or other lexical association scores are their values.

The number of contexts can be reduced by applying a method to filter out those contexts that are not relevant [27]. This technique reduces the size of the hash table by removing irrelevant contexts. It consists of computing a *relevance* measure (in our experiments, we use log-likelihood [28]) between each lemma and their co-occurring contexts. More precisely, for each lemma, the R most relevant contexts are selected, i.e., only those contexts with the highest log-likelihood values are stored in the table. R is a global constant that is often declared with values ranging from 10–1000 [29,30]. As will be shown later, it is not required to assign R with high values. In many cases, it is enough to select the 100 or 200 most relevant contexts. Filtering out non-relevant contexts allows us to reduce the distributional model significantly, which is represented as a hash table. Unlike word embeddings, this representation makes the semantic model transparent and fully interpretable in linguistic terms.

Let us give an example. Once the contexts have been sorted by likelihood values, the first context of the lemma *cow* is the following:

$$cow \quad => \{ (dung, \text{nmod}\downarrow) \quad => \quad 12.008 \ \}$$

This means that its most relevant context is (*dung*, nmod↓), as it is assigned a 12.008 log-likelihood score. This high value is due to the fact that word *cow* co-occurs with *dung* as the noun modifier ("*cow dung*") in many sentences. By contrast, the 300$^{\text{th}}$ most relevant context of *cow* in the same table is:

$$cow \quad => \{ (sacrifice, \text{dobj}\downarrow) \quad => \quad 0.049 \ \}$$

As shown by the likelihood value, the direct object of the verb *sacrifice* represents a much less relevant context of *cow*. Indeed, the type of entities that can be sacrificed is much broader than the types of dungs.

Notice that syntax-based models are fully interpretable as each dimension (or key) is an explicit lexico-syntactic context. Methods based on dimensionality reduction and dense embeddings, by contrast, make the vector space more compact with dimensions that are not transparent for linguistic purposes.

5. Experiments and Evaluation

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

5.1. The 8-8-8 Dataset

The goal of the experiment was to compare the basic count-based model defined in the previous Section 4 with the results obtained by different versions of embeddings, which were reported in Camacho-Collados and Navigli [3].

Table 6 shows the results obtained by the count-based strategy we developed, Dep500, which was 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 [31]. The method outperformed the results obtained by three standard embedding models: the Continuous Bag-of-Words (cbow) and skip-gram models of Word2Vec [5], and GloVe [32], which are based on bag-of-words contexts (we use *bow* to refer to linear bag-of-word contexts, which must be distinguished from CBOW [8]), and whose results were reported in Camacho-Collados and Navigli [3]. The dimensionality of the dense vectors was set to 300 for the three embedding models; context-size of 5 for cbow and 10 for skip-gram and GloVe; hierarchical softmax for cbow and negative sampling for skip-gram. In all experiments, the corpus used to build the vector space was the 1.7B-tokens English Wikipedia (dump of November 2014).

Table 6. Outlier Position Percentage (OPP) and accuracy of different word models on the *8-8-8* outlier detection dataset using Wikipedia.

System	Strategy	OPP	Accuracy
Dep500	count + syntax	97.3	90.6
Word2Vec	bow (cbow)	95.3	73.4
Word2Vec	bow (skip-gram)	93.8	70.3
Glove	bow	91.8	56.3

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 $R = 1000$. This means that small count-based distributional models with relevant contexts perform better than large models made up of many noisy syntactic contexts. The best score, however, was achieved at $R = 500$, namely concerning the OPP measure. This is why we chose this configuration to be compared with the other systems.

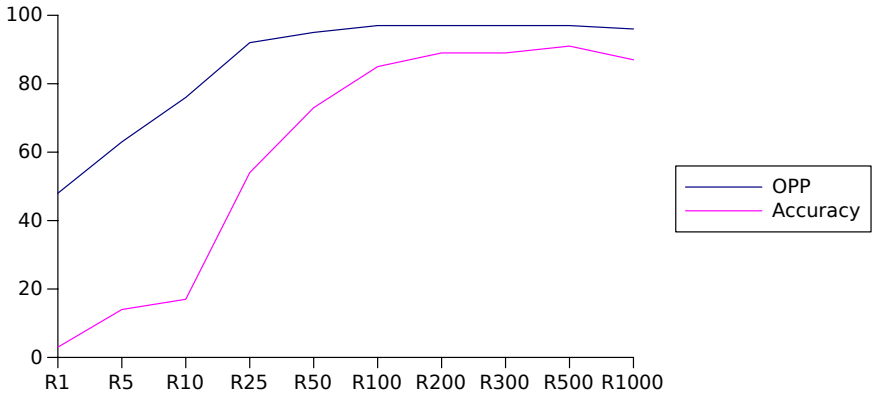

Figure 1. Accuracy and OPP of our count-based strategy across different filtering thresholds: from $R = 1$–$R = 1000$.

5.2. The 12-8-8 Expanded Dataset

The main goal of the next experiments was 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 compared the use of syntactic dependencies and bag-of-words in the same task. We required a dataset without multi-words since some of the tools we used for building distributions only tokenized unigrams. For this purpose, we defined the following six strategies:

Count$_1$: A count-based model with rule-based dependencies.
Count$_2$: A count-based model with transition-based dependencies.
Count$_3$: A count-based model with bag-of-words.
Emb$_1$: Embeddings with rule-based dependencies.
Emb$_2$: Embeddings with transition-based dependencies.
Emb$_3$: Embeddings with bag-of-words (skip-gram algorithm).

The three count-based models were built with the filter $R = 300$. In this case, we chose $R = 300$ because is more efficient than $R = 500$, yielding very similar results. The dimensionality of the three embeddings was set to 300, and the algorithm to build them was based on the continuous *skip-gram* neural embedding model [5], with the 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 Emb$_2$ and Emb$_3$ were the models described in Levy and Goldberg [13], which are publicly

available (https://levyomer.wordpress.com/2014/04/25/dependency-based-word-embeddings/). 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 Goldberg and Nivre [33] (we are very grateful to the authors for sending us the English Wikipedia syntactically analyzed with their parser). The performance of the parser for English is about 89% UAS (Unlabeled Attachment Score) obtained on the CoNLL 2007 dataset. To build Emb_2, we made use of `Word2Vecf` (https://bitbucket.org/yoavgo/word2vecf), a modified version of Word2Vec, which is suited to build embeddings with syntactic dependencies [13] (code.google.com/p/word2vec/). Rule-based dependencies were obtained using DepPattern [34].

Even though the strategies were very different using very different software, we tried to use the same hyperparameters in order to minimize differences that were not due to the word models themselves. As Levy et al. [7] suggested, much of the difference between vectorial models were 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 7 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) gave very similar scores. However, they tended to perform better than those based on bag-of-words (as in the previous experiment in Section 5.1). This was in accordance with a great number of previous works that evaluated and compared syntactic contexts (usually dependencies) with bag-of-words techniques [13,35–40]. All of them stated 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 (synonymy) word relations. By contrast, *bow*-based models tend to perform better in tasks oriented to identify semantic relatedness and analogies [11]. By taking into account that syntax-based methods work better on synonymy while bow-based models on relatedness, we may conclude the following: First, the outlier detection task was suited to search for synonymy/co-hyponymy and not for semantic relatedness or topicality [16], and second, the type of context (dependency-based or bag-of-words) was more determinant than the type of model (count-based or embeddings) for that task. Finally, embeddings clearly outperformed count-based representations when the contexts were defined with bag-of-words (see the score of Emb_3 against $Count_3$ in Table 7).

Table 7. Outlier Position Percentage (OPP) and accuracy of different word models on the *12-8-8* outlier detection dataset using Wikipedia.

System	Strategy	OPP	Accuracy
$Count_1$	syntactic rules	94.92	75.0
$Count_2$	syntactic transitions	93.48	71.87
$Count_3$	bow	86.71	60.41
Emb_1	syntactic rules	93.09	76.04
Emb_2	syntactic transitions	94.27	72.91
Emb_3	bow (skip-gram)	93.88	69.79

5.3. 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 three best strategies, according to the previous experiments: count-based model with rule-based dependencies ($Count_1$), embeddings with rule-based dependencies (Emb_1), and bow-based embeddings (Emb_3). 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 two embeddings were implemented with skip-gram and the negative-sampling

parameter set at 15. In all experiments, the corpus used to build the three models was the 250M-tokens Portuguese Wikipedia (dump of September 2015). Table 8 shows the results obtained on the Portuguese *12-8-8* dataset by the three models evaluated.

Table 8. Outlier Position Percentage (OPP) and accuracy of several distributional models on the *12-8-8* outlier detection dataset. The three first models were trained using Portuguese Wikipedia, while the rest of models consist of pre-trained embeddings derived from a representative Portuguese corpus compiled by the NILC group.

System	Strategy	OPP	Accuracy	Source
$Count_1$	syntactic rules	91.56	**50.00**	PT Wikipedia
Emb_1	syntactic rules	84.37	39.58	PT Wikipedia
Emb_3	bow (skip-gram)	83.72	40.62	PT Wikipedia
Word2Vec	cbow	88.67	*58.33*	NILC corpus
Word2Vec	skip-gram	*91.01*	*62.5*	NILC corpus
FastText	cbow	*89.06*	*57.29*	NILC corpus
FastText	skip-gram	*90.62*	*62.5*	NILC corpus
Wang2Vec	cbow	*90.75*	*62.5*	NILC corpus
Wang2Vec	skip-gram	*91.01*	*62.5*	NILC corpus
GloVe	no neural bow	*91.01*	*62.5*	NILC corpus

In these experiments, the count-based strategy obtained higher OPP and accuracy scores than the two embeddings. This may be partially explained by the fact that the Portuguese Wikipedia was seven-times smaller than the English one, and neural networks require a large corpus to make better predictions. In addition, we also evaluated freely-available pre-trained embeddings derived from a representative Portuguese corpus collected by the NILC group [41] and containing 1.4B-tokens (http://nilc.icmc.usp.br/embeddings). Four systems were evaluated: Word2vec and GloVe, and two extensions of the former, namely FastText (https://fasttext.cc/) and Wang2Vec, by considering the two variations: cbow and skip-gram . Some pre-processing tasks were performed on the corpus: stop word removal and stemming. In all cases, the accuracy was higher than in the three previous experiments. This is probably because they were trained with a corpus representative of the Portuguese language, whose size is also much larger (over six-times bigger) than that of Wikipedia. However, it is worth noting that none of them outperformed the OPP value reached by the $Count_1$ system (count and syntax based), despite the fact that we used the smallest corpus (the NILC corpus is not available).

6. 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 though count-based strategies seem to perform better with less training corpus, as is the case of the experiment carried out with the Portuguese Wikipedia.

As the outlier detection task is aimed at identifying synonyms and co-hyponyms, we can infer that these results confirm previous experiments and conclusions. Namely, it confirms that models built with syntactic dependencies are better suited to identify synonymy and co-hyponymy [11,13,16].

In future work, we intend to develop outlier detection datasets for many other languages in order to make possible cross-lingual word similarity evaluation. The new multilingual benchmark could be seen as complementary to the one used in the Cross-Lingual Word Similarity Task at SemEval 2017 [14].

The software required to build the count-based models, as well as the *12-8-8* datasets are publicly available (http://gramatica.usc.es/~gamallo/prototypes/Word2Model.tgz).

Funding: This research received no external funding.

Acknowledgments: 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).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gamallo, P. Evaluation of Distributional Models with the Outlier Detection Task. In *7th Symposium on Languages, Applications and Technologies (SLATE 2018)*; Henriques, P.R., Leal, J.P., Leitão, A.M., Guinovart, X.G., Eds.; OpenAccess Series in Informatics (OASIcs); Schloss Dagstuhl–Leibniz-Zentrum Fuer Informatik: Dagstuhl, Germany, 2018; Volume 62, pp. 13:1–13:8, doi:10.4230/OASIcs.SLATE.2018.13.
2. Batchkarov, M.; Kober, T.; Reffin, J.; Weeds, J.; Weir, D. 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, Berlin, Germany, 7–12 August 2016; pp. 7–12.
3. Camacho-Collados, J.; Navigli, R. 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, Berlin, Germany, 7–12 August 2016; pp. 43–50.
4. Baroni, M.; Dinu, G.; Kruszewski, G. 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 (Volume 1: Long Papers), Baltimore, MA, USA, 23–25 June 2014; pp. 238–247.
5. Mikolov, T.; Yih, W.-T.; Zweig, G. 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, Atlanta, GA, USA, 9–14 June 2013; pp. 746–751.
6. Lebret, R.; Collobert, R. Rehabilitation of Count-Based Models for Word Vector Representations. In *CICLing-2015*; Gelbukh, A.F., Ed.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2015; Volume 9041, pp. 417–429.
7. Levy, O.; Goldberg, Y.; Dagan, I. Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Trans. Assoc. Comput. Linguist.* **2015**, *3*, 211–225.
8. Levy, O.; Goldberg, Y. Linguistic Regularities in Sparse and Explicit Word Representations. In Proceedings of the Eighteenth Conference on Computational Natural Language Learning (CoNLL 2014), Baltimore, MD, USA, 26–27 June 2014; pp. 171–180.
9. Blacoe, W.; Lapata, M. A comparison of vector-based representations for semantic composition. In Proceedings of the Empirical Methods in Natural Language Processing-EMNLP-2012, Jeju Island, Korea, 12–14 July 2012; pp. 546–556.
10. Faruqui, M.; Dyer, C. Non-distributional Word Vector Representations. In Proceedings of ACL, Beijing, China, 26–31 July 2015; pp. 464–469.
11. Gamallo, P. Comparing Explicit and Predictive Distributional Semantic Models Endowed with Syntactic Contexts. *Lang. Resour. Eval.* **2017**, *51*, 727–743.
12. Huang, E.; Socher, R.; Manning, C. Improving word representations via global context and multiple word prototypes. In Proceedings of the ACL-2012, Jeju Island, Korea, 8–14 July 2012; pp. 873–882.
13. Levy, O.; Goldberg, Y. Dependency-Based Word Embeddings. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014), Baltimore, MD, USA, 22–27 June 2014; pp. 302–308.
14. Camacho-Collados, J.; Pilehvar, M.; Collier, N.; Navigli, R. SemEval-2017 Task 2: Multilingual and Cross-lingual Semantic Word Similarity. In Proceedings of the SemEval, Vancouver, BC, Canada, 3–4 August 2017.
15. Finkelstein, L.; Gabrilovich, E.; Matias, Y.; Rivlin, E.; Solan, Z.; Wolfman, G.; Ruppin, E. Placing search in context: The concept revisited. *ACM Trans. Inf. Syst.* **2002**, *20*, 116–131.

16. Agirre, E.; Alfonseca, E.; Hall, K.; Kravalova, J.; Paşca, M.; Soroa, A. A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches. In *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL '09)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2009; pp. 19–27.
17. Hill, F.; Reichart, R.; Korhonen, A. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Comput. Linguist.* **2015**, *41*, 665–695.
18. Mitchell, J.; Lapata, M. Vector-based models of semantic composition. In Proceedings of the ACL-08: HLT, Columbus, OH, USA, 19–20 June 2008; pp. 236–244.
19. Grefenstette, E.; Sadrzadeh, M. Experimenting with Transitive Verbs in a DisCoCat. In Proceedings of the Workshop on Geometrical Models of Natural Language Semantics (EMNLP-2011), Edinburgh, UK, 31 July 2011.
20. Landauer, T.; Dumais, S. A solution to Plato's problem: The Latent Semantic Analysis theory of acquision, induction and representation of knowledge. *Psychol. Rev.* **1997**, *10*, 211–240.
21. Turney, P. Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In Proceedings of the 12th European Conference of Machine Learning, Freiburg, Germany, 5–7 September 2001; pp. 491–502.
22. Wilkens, R.; Zilio, L.; Ferreira, E.; Villavicencio, A. B2SG: A TOEFL-like Task for Portuguese. In Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, 23–28 May 2016.
23. Rodríguez, M.A.; Egenhofer, M.J. The role of vocabulary teaching. *TESOL Q.* **1976**, *10*, 77–89.
24. Gamallo, P.; Agustini, A.; Lopes, G. Clustering Syntactic Positions with Similar Semantic Requirements. *Comput. Linguist.* **2005**, *31*, 107–146.
25. Baroni, M.; Lenci, A. Distributional Memory: A General Framework for Corpus-based Semantics. *Comput. Linguist.* **2010**, *36*, 673–721.
26. Gamallo, P.; Bordag, S. Is Singular Value Decomposition Useful for Word Simalirity Extraction. *Lang. Resour. Eval.* **2011**, *45*, 95–119.
27. Bordag, S. A Comparison of Co-occurrence and Similarity Measures as Simulations of Context. In Proceedings of the 9th CICLing, Haifa, Israel, 17–23 February 2008; pp. 52–63.
28. Dunning, T. Accurate Methods for the Statistics of Surprise and Coincidence. *Comput. Linguist.* **1993**, *19*, 61–74.
29. Biemann, C.; Riedl, M. Text: Now in 2D! A Framework for Lexical Expansion with Contextual Similarity. *J. Lang. Model.* **2013**, *1*, 55–95.
30. Padró, M.; Idiart, M.; Villavicencio, A.; Ramisch, C. 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 2014, A Meeting of SIGDAT, a Special Interest Group of the ACL), Doha, Qatar, 25–29 October 2014; pp. 419–424.
31. Gamallo, P. Dependency Parsing with Compression Rules. In Proceedings of the 14th International Workshop on Parsing Technology (IWPT 2015), Association for Computational Linguistics, Bilbao, Spain, 5–7 July 2015; pp. 107–117.
32. Pennington, J.; Socher, R.; Manning, C.D. GloVe: Global Vectors for Word Representation. In Proceedings of the Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
33. Goldberg, Y.; Nivre, J. A Dynamic Oracle for Arc-Eager Dependency Parsing. In Proceedings of the 24th International Conference on Computational Linguistics: Technical Papers (COLING 2012), Mumbai, India, 8–15 December 2012; pp. 959–976.
34. Gamallo, P.; Garcia, M. Dependency parsing with finite state transducers and compression rules. *Inf. Process. Manag.* **2018**, *54*, 1244–1261.
35. Gamallo, P. Comparing Window and Syntax Based Strategies for Semantic Extraction. In *PROPOR-2008*; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2008; pp. 41–50.
36. Gamallo, P. Comparing Different Properties Involved in Word Similarity Extraction. In Proceedings of the 14th Portuguese Conference on Artificial Intelligence (EPIA'09), LNCS, Aveiro, Portugal, 12–15 October 2009; Volume 5816, pp. 634–645.

37. Grefenstette, G. Evaluation techniques for automatic semantic extraction: Comparing syntactic and window-based approaches. In Proceedings of the Workshop on Acquisition of Lexical Knowledge from Text SIGLEX/ACL, Columbus, OH, USA, 21 June 1993; pp. 205–216.
38. Padó, S.; Lapata, M. Dependency-Based Construction of Semantic Space Models. *Comput. Linguist.* **2007**, *33*, 161–199.
39. Peirsman, Y.; Heylen, K.; Speelman, D. Finding semantically related words in Dutch. Co-occurrences versus syntactic contexts. In Proceedings of the CoSMO Workshop, Roskilde, Denmark, 20 August 2007; pp. 9–16.
40. Seretan, V.; Wehrli, E. Accurate Collocation Extraction Using a Multilingual Parser. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the ACL, Sydney, Australia, 17–18 July 2006; pp. 953–960.
41. Hartmann, N.; Fonseca, E.R.; Shulby, C.; Treviso, M.V.; Silva, J.; Aluísio, S.M. Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks. In Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology (STIL 2017), Uberlândia, Brazil, 2–5 October 2017; pp. 122–131.