Find the word that does not belong:
A Framework for an Intrinsic Evaluation of Word Vector Representations

José Camacho-Collados and Roberto Navigli
Department of Computer Science
Sapienza University of Rome
{collados,navigli}@di.uniroma1.it

Abstract
We present a new framework for an intrinsic evaluation of word vector representations based on the outlier detection task. This task is intended to test the capability of vector space models to create semantic clusters in the space. We carried out a pilot study building a gold standard dataset and the results revealed two important features: human performance on the task is extremely high compared to the standard word similarity task, and state-of-the-art word embedding models, whose current shortcomings were highlighted as part of the evaluation, still have considerable room for improvement.

1 Introduction
Vector Space Models have been successfully used on many NLP tasks (Turney and Pantel, 2010) such as automatic thesaurus generation (Crouch, 1988; Curran and Moens, 2002); word similarity (Deerwester et al., 1990; Turney et al., 2003; Radinsky et al., 2011) and clustering (Pantel and Lin, 2002), query expansion (Xu and Croft, 1996), information extraction (Laender et al., 2002), semantic role labeling (Erk, 2007; Pennacchiotti et al., 2008), spelling correction (Jones and Martin, 1997), and Word Sense Disambiguation (Navigli, 2012). These models are in the main based on the distributional hypothesis of Harris (1954) claiming that words that occur in the same contexts tend to have similar meanings. Recently, more complex models based on neural networks going beyond simple co-occurrence statistics have been developed (Mikolov et al., 2013; Pennington et al., 2014) and have proved beneficial on key NLP applications such as syntactic parsing (Weiss et al., 2015), Machine Translation (Zou et al., 2013), and Question Answering (Bordes et al., 2014).

Word similarity, which numerically measures the extent to which two words are similar, is generally viewed as the most direct intrinsic evaluation of these word vector representations (Baroni et al., 2014; Levy et al., 2015). Given a gold standard of human-assigned scores, the usual evaluation procedure consists of calculating the correlation between these human similarity scores and scores calculated by the system. While word similarity has been shown to be an interesting task for measuring the semantic coherence of a vector space model, it suffers from various problems. First, the human inter-annotator agreement of standard datasets has been shown to be relatively too low for it to be considered a reliable evaluation benchmark (Batchkarov et al., 2016). In fact, many systems have already surpassed the human inter-annotator agreement upper bound in most of the standard word similarity datasets (Hill et al., 2015). Another drawback of the word similarity evaluation benchmark is its simplicity, as words are simply viewed as points in the vector space. Other interesting properties of vector space models are not directly addressed in the task.

As an alternative we propose the outlier detection task, which tests the capability of vector space models to create semantic clusters (i.e. clusters of semantically similar items). As is the case with word similarity, this task aims at evaluating the semantic coherence of vector space models, but providing two main advantages: (1) it provides a clear gold standard, thanks to the high human performance on the task, and (2) it tests an interesting language understanding property of vector space models not fully addressed to date, and this is their ability to create semantic clusters in the vector space, with potential applications to various NLP tasks.
2 Outlier Detection Task

The proposed task, referred to as outlier detection henceforth, is based on a standard vocabulary question of language exams (Richards, 1976). Given a group of words, the goal is to identify the word that does not belong in the group. This question is intended to test the student’s vocabulary understanding and knowledge of the world. For example, book would be an outlier for the set of words apple, banana, lemon, book, orange, as it is not a fruit like the others. A similar task has already been explored as an ad-hoc evaluation of the interpretability of topic models (Chang et al., 2009) and word vector dimensions (Murphy et al., 2012; Fyshe et al., 2015; Faruqui et al., 2015).

In order to deal with the outlier detection task, vector space models should be able to create semantic clusters (i.e. fruits in the example) compact enough to detect all possible outliers. A formalization of the task and its evaluation is presented in Section 2.1 and some potential applications are discussed in Section 2.2.

2.1 Formalization

Formally, given a set of words \( W = \{w_1, w_2, \ldots, w_n, w_{n+1}\} \), the task consists of identifying the word (outlier) that does not belong to the same group as the remaining words. For notational simplicity, we will assume that \( w_1, \ldots, w_n \) belong to the same cluster and \( w_{n+1} \) is the outlier. In what follows we explain a procedure for detecting outliers based on semantic similarity.

We define the compactness score \( c(w) \) of a word \( w \in W \) as the compactness of the cluster \( W \setminus \{w\} \), calculated by averaging all pair-wise semantic similarities of the words in \( W \setminus \{w\} \):

\[
c(w) = \frac{1}{k} \sum_{w_i \in W \setminus \{w\}} \sum_{w_j \in W \setminus \{w\}, w_j \neq w_i} \text{sim}(w_i, w_j) \tag{1}
\]

where \( k = n(n-1) \). We propose two measures for computing the reliability of a system in detecting an outlier given a set of words: Outlier Position (OP) and Outlier Detection (OD). Given a set \( W \) of \( n+1 \) words, OP is defined as the position of the outlier \( w_{n+1} \) according to 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). OD is, instead, defined as 1 if the outlier is correctly detected (i.e. \( OP(w_{n+1}) = n \)) and 0 otherwise. To estimate the overall performance on a dataset \( D \) (composed of \( |D| \) sets of words), we define the Outlier Position Percentage (OPP) and Accuracy measures:

\[
OPP = \frac{\sum_{W \in D} OP(W)}{|D|} \times 100 \tag{2}
\]

\[
Accuracy = \frac{\sum_{W \in D} OD(W)}{|D|} \times 100 \tag{3}
\]

The compactness score of a word may be expensive to calculate if the number of elements in the cluster is large. In fact, the complexity of calculating \( OP \) and \( OD \) measures given a cluster and an outlier is \( (n+1) \times n \times (n-1) = O(n^3) \). However, this complexity can be effectively reduced to \( (n+1) \times 2n = O(n^2) \). Our proposed calculations and the proof are included in Appendix A.

2.2 Potential applications

In this work we focus on the intrinsic semantic properties of vector space models which can be inferred from the outlier detection task. In addition, since it is a task based partially on semantic similarity, high-performing models in the outlier detection task are expected to contribute to applications in which semantic similarity has already shown its potential: Information Retrieval (Hliaoutakis et al., 2006), Machine Translation (Lavie and Denkowski, 2009), Lexical Substitution (McCarty and Navigli, 2009), Question Answering (Mohler et al., 2011), Text Summarization (Mohammad and Hirst, 2012), and Word Sense Disambiguation (Patwardhan et al., 2003), to name a few. Furthermore, there are other NLP applications directly connected with the semantic clustering proposed in the outlier detection task. Ontology Learning is probably the most straightforward application, as a meaningful cluster of items is expected to share a common hypernym, a property that has already been exploited in recent studies using embeddings (Fu et al., 2014; Espinosa-Anke et al., 2016). In fact, building ontologies is a time-consuming task and generally relies on automatic or semi-automatic steps (Velardi et al., 2013; Alfarone and Davis, 2015). Ontologies are one of the basic components of the Semantic Web (Berners-Lee et al., 2000) and have already proved their importance in downstream applications like Question Answering (Mann, 2002).
which in the main rely on large structured knowledge bases (Bordes et al., 2014).

In this paper we do not perform any quantitative evaluation to measure the correlation between the performance of word vectors on the outlier detection task and downstream applications. We argue that the conclusions drawn by recent works (Tsvetkov et al., 2015; Chiu et al., 2016) as a result of measuring the correlation between standard intrinsic evaluation benchmarks (e.g. word similarity datasets) and downstream task performances are hampered by a serious methodological issue: in both cases, the sample set of word vectors used for measuring the correlation is not representative enough, which is essential for this type of statistical study (Patton, 2005). All sample vectors came from corpus-based models trained on the same corpus and all perform well on the considered intrinsic tasks, which constitute a highly homogeneous and not representative sample set. Moreover, using only a reduced selected set of applications does not seem sufficient to draw general conclusions about the quality of an intrinsic task, but rather about its potential on those specific applications. Further work should focus on these issues before using downstream applications to measure the impact of intrinsic tasks for evaluating the quality of word vectors. However, this is out of the scope of this paper.

Table 1: First four clusters (including outliers) of the 8-8-8 outlier detection dataset.

<table>
<thead>
<tr>
<th>Cluster elements</th>
<th>Big cats</th>
<th>European football teams</th>
<th>Solar System planets</th>
<th>Months</th>
</tr>
</thead>
<tbody>
<tr>
<td>tiger</td>
<td>FC Barcelona</td>
<td>Mercury</td>
<td>January</td>
<td></td>
</tr>
<tr>
<td>lion</td>
<td>Bayern Munich</td>
<td>Venus</td>
<td>March</td>
<td></td>
</tr>
<tr>
<td>cougar</td>
<td>Real Madrid</td>
<td>Earth</td>
<td>May</td>
<td></td>
</tr>
<tr>
<td>jaguar</td>
<td>AC Milan</td>
<td>Mars</td>
<td>July</td>
<td></td>
</tr>
<tr>
<td>leopard</td>
<td>Juventus</td>
<td>Jupiter</td>
<td>September</td>
<td></td>
</tr>
<tr>
<td>cheetah</td>
<td>Atletico Madrid</td>
<td>Saturn</td>
<td>November</td>
<td></td>
</tr>
<tr>
<td>wildcat</td>
<td>Chelsea</td>
<td>Uranus</td>
<td>February</td>
<td></td>
</tr>
<tr>
<td>lynx</td>
<td>Borussa Dortmund</td>
<td>Neptune</td>
<td>June</td>
<td></td>
</tr>
</tbody>
</table>

1st Outlier: dog - Miami Dolphins - Sun - Wednesday
2nd Outlier: mouse - McLaren - Moon - winter
3rd Outlier: dolphin - Los Angeles Lakers - Triton - date
4th Outlier: shark - Bundesliga - Comet Halley - year
5th Outlier: savanna - football - eclipse - astrology
6th Outlier: jungle - goal - astronaut - birthday
7th Outlier: day - couch - lunch - ball
8th Outlier: car - fridge - window - paper

3 Pilot Study

We carried out a pilot study on the outlier detection task. To this end, we developed a new dataset, 8-8-8 henceforth. The dataset consisted of eight different topics each made up of a cluster of eight words and eight possible outliers. Four annotators were used for the creation of the dataset. Each annotator was asked to first identify two topics, and for each topic to provide a set of eight words belonging to the chosen topic (elements in the cluster), and a set of eight heterogeneous outliers, selected varying their similarity to and relatedness with the elements of the cluster\(^2\). In total, the dataset included sixty-four sets of 8 + 1 words for the evaluation. Tables 1 and 2 show the eight clusters and their respective outliers of the 8-8-8 outlier detection dataset.

When we consider the time annotators had to spend creating the relatively small dataset for this pilot study, the indications are that building a large-scale dataset may not need to be very time-consuming. In our study, the annotators spent most of their time reading and understanding the guidelines, and then thinking about suitable topics. In fact, with a view to constructing a large-scale dataset, this topic selection step may be carried out prior to giving the assignments to the annotators, providing topics to annotators according to their

\(^1\)In the case of Chiu et al. (2016) all word vectors in the sample come from the Skip-Gram model of Word2Vec (Mikolov et al., 2013).

\(^2\)We release the full dataset and guidelines for the creation of the topics at [http://lcl.uniroma1.it/outlier-detection](http://lcl.uniroma1.it/outlier-detection)
### Table 2: Last four clusters (including outliers) from the 8-8-8 outlier detection dataset.

<table>
<thead>
<tr>
<th>Cluster elements</th>
<th>IT companies</th>
<th>German car manufacturers</th>
<th>Apostles of Jesus Christ</th>
<th>South American countries</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st Outlier</td>
<td>Opel</td>
<td>Michelin</td>
<td>Noah</td>
<td>Bogotá</td>
</tr>
<tr>
<td>2nd Outlier</td>
<td>Boeing</td>
<td>Bridgestone</td>
<td>Mary</td>
<td>Rio de Janeiro</td>
</tr>
<tr>
<td>3rd Outlier</td>
<td>Nestlé</td>
<td>Boeing</td>
<td>Pope Benedict XVI</td>
<td>New York</td>
</tr>
<tr>
<td>4th Outlier</td>
<td>Adidas</td>
<td>Samsung</td>
<td>Ambrose</td>
<td>Madrid</td>
</tr>
<tr>
<td>5th Outlier</td>
<td>computer</td>
<td>Michael Schumacher</td>
<td>crucifixion</td>
<td>town</td>
</tr>
<tr>
<td>6th Outlier</td>
<td>software</td>
<td>Angela Merkel</td>
<td>church</td>
<td>government</td>
</tr>
<tr>
<td>7th Outlier</td>
<td>chair</td>
<td>Capri</td>
<td>airplane</td>
<td>bottle</td>
</tr>
<tr>
<td>8th Outlier</td>
<td>plant</td>
<td>pineapple</td>
<td>Microsoft</td>
<td>telephone</td>
</tr>
</tbody>
</table>

3.1 Human performance

We assessed the human performance of eight annotators in the task via accuracy. To this end, each annotator was given eight different groups of words, one for each of the topics of the 8-8-8 dataset. Each group of words was made up of the set of eight words comprising the cluster, plus one additional outlier. All the words were shuffled and given to the annotator without any additional information (e.g. annotators did not know the topic of the cluster). The task for the annotators consisted of detecting the outlier in each set of nine words. To this end, each annotator was asked to provide two different answers: one without any external help, and a second one in which the annotator could use the Web as external help for three minutes before giving his answer. This human performance in the outlier detection task may be viewed as equivalent to the inter-annotator agreement in word similarity, which is used to measure the human performance in the task.

The results of the experiment were the following: an accuracy of 98.4% for the first task in which annotators did not use any external help, and an accuracy of 100% for the second task in which annotators were allowed to use external help. This contrasts with the evaluation performed in word similarity, which is based on human-assigned scores with a relatively low inter-annotator agreement. For example, the inter-annotator agreements in the standard WordSim-353 (Finkelstein et al., 2002) and SimLex-999 (Hill et al., 2015) word similarity datasets were, respectively, 0.61 and 0.67 according to average pair-wise Spearman correlation. In fact, both upper-bound values have already been surpassed by automatic models (Huang et al., 2012; Wieting et al., 2015).

3.2 Word embeddings performance

We tested the performance of three standard word embedding models in the outlier detection task: the CBOW and Skip-Gram models of Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014). We report the results of each of the models trained on the 3B-words UMBC webbase corpus

http://ebiquity.umbc.edu/blogger/2013/05/01/umbc-webbase-corpus-of-3b-english-words/

We used the Wikipedia dump of November 2014.

The dimensionality of the vectors was set to 300 for the three 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.
<table>
<thead>
<tr>
<th>Model</th>
<th>Corpus</th>
<th>OPP</th>
<th>Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>CBOW</td>
<td>UMBC</td>
<td>93.8</td>
<td>73.4</td>
</tr>
<tr>
<td></td>
<td>Wikipedia</td>
<td>95.3</td>
<td>73.4</td>
</tr>
<tr>
<td>Skip-Gram</td>
<td>UMBC</td>
<td>92.6</td>
<td>64.1</td>
</tr>
<tr>
<td></td>
<td>Wikipedia</td>
<td>93.8</td>
<td>70.3</td>
</tr>
<tr>
<td></td>
<td>Google News</td>
<td>94.7</td>
<td>70.3</td>
</tr>
<tr>
<td>GloVe</td>
<td>UMBC</td>
<td>81.6</td>
<td>40.6</td>
</tr>
<tr>
<td></td>
<td>Wikipedia</td>
<td>91.8</td>
<td>56.3</td>
</tr>
</tbody>
</table>

Table 3: Outlier Position Percentage (OPP) and Accuracy (Acc.) of different word embedding models on the 8-8-8 outlier detection dataset.

as similarity measure ($sim$ in Equation 1).

Table 3 shows the results of all the word embedding models on the 8-8-8 outlier detection dataset. Outliers, which were detected in over 40% of cases by all models, were consistently given high compactness scores. This was reflected in the OPP results (above 80% in all cases), which proves the potential and the capability of word embeddings to create compact clusters. All the models performed particularly well in the Months and South American countries clusters. However, the best model in terms of accuracy, i.e. CBOW, achieved 73.4%, which is far below the human performance, estimated in the 98.4%-100% range.

In fact, taking a deeper look at the output we find common errors committed by these models. First, the lack of meaningful occurrences for a given word, which is crucial for obtaining an accurate word vector representation, seems to have been causing problems in the cases of the wildcat and lynx instances of the Big cats cluster, and of Alpina from the German car manufacturers cluster. Second, the models produced some errors on outliers closely related to the words of the clusters, incorrectly considering them as part of the cluster. Examples of this phenomenon are found in the outliers Bundesliga from the European football teams cluster, and software from the IT companies cluster. Third, the ambiguity, highlighted in the word Smart from the German car manufacturers cluster and in the Apostles of Jesus Christ cluster, is an inherent problem of all these word-based models. Finally, we encountered the issue of having more than one lexicalization (i.e. synonyms) for a given instance (e.g. Real, Madrid, Real Madrid, or Real Madrid CF), which causes the representations of a given lexicalization to be ambiguous or not so accurate and, in some cases, to miss a representation for a given lexicalization if that lexicalization is not found enough times in the corpus\(^6\). In order to overcome these ambiguity and synonymy issues, it might be interesting for future work to leverage vector representations constructed from large lexical resources such, as FreeBase (Bordes et al., 2011; Bordes et al., 2014), Wikipedia (Camacho-Collados et al., 2015a), or BabelNet (Iacobacci et al., 2015; Camacho-Collados et al., 2015b).

4 Conclusion

In this paper we presented the outlier detection task and a framework for an intrinsic evaluation of word vector space models. The task is intended to test interesting semantic properties of vector space models not fully addressed to date. As shown in our pilot study, state-of-the-art word embeddings perform reasonably well in the task but are still far from human performance. As opposed to the word similarity task, the outlier detection task achieves a very high human performance, proving the reliability of the gold standard. Finally, we release the 8-8-8 outlier detection dataset and the guidelines given to the annotators as part of the pilot study, and an easy-to-use Python code for evaluating the performance of word vector representations given a gold standard dataset at http://lcl.uniroma1.it/outlier-detection.

Acknowledgments

The authors gratefully acknowledge the support of the ERC Starting Grant MultiJEDI No. 259234.

We would like to thank Claudio Delli Bovi, Ilenia Giambruno, Ignacio Iacobacci, Massimiliano Mancini, Tommaso Pasini, Taher Pilehvar, and Alessandro Raganato for their help in the construction and evaluation of the outlier detection dataset. We would also like to thank Jim McManus for his comments on the manuscript.

References

Daniele Alfarone and Jesse Davis. 2015. Unsupervised learning of an is-a taxonomy from a limited domain-specific corpus. In Proceedings of IJCAI.

\(^6\)This last issue was not present in this evaluation as for the multiword instances we carefully selected the lexicalizations which were covered by the pre-trained Word2Vec vectors, which ensured a full coverage of all models.


Alona Fyshe, Leila Wehbe, Partha P Talukdar, Brian Murphy, and Tom M Mitchell. 2015. A compositional and interpretable semantic space. In Proc. of NAACL.


Michael Quinn Patton. 2005. *Qualitative research*. Wiley Online Library.


John Wiecing, Mohit Bansal, Kevin Gimpel, Karen Livescu, and Dan Roth. 2015. From paraphrase database to compositional paraphrase model and back. *TACL*.


A Proposition 1

The complexity for calculating \( OP(w) \) can be reduced to \( 2n \) by calculating the following pseudo-inverted compactness score\(^7\) \( p(w) \) instead of the compactness score \( c(w) \) of Equation 1, and defining \( OP_p(w) \) as the position of the outlier in \( W \) according to the inverted pseudo-inverted compactness score:

\[
p(w) = \frac{1}{k'} \left( \sum_{w_i \in W} \text{sim}(w, w_i) + \sum_{w_i \in W} \text{sim}(w_i, w) \right)
\]

where \( k' = 2(|W| - 1) \).

Proof. Since \( OP(w) \) is given by the position of \( c(w) \) with respect to the remaining words in \( W \) and \( \leq \) represents a relation of total order, we only have to prove the following statement:

\[
c(w) \leq c(w') \iff p(w') \leq p(w), \forall w, w' \in W
\]

Given any \( w \in W \), we can calculate the sum of all pair-wise similarities of the words in \( W \) (i.e. \( \mu \)) as follows:

\[
\mu = \sum_{w_i \in W \setminus \{w\}} \sum_{w_j \in W \setminus \{w\}} \text{sim}(w_i, w_j) + \sum_{w_i \in W \setminus \{w\}} \text{sim}(w, w_i)
\]

\[
= k \cdot c(w) + k' \cdot p(w)
\]

where \( k = (|W| - 1)(|W| - 2) \). Therefore,

\[
\mu = k \cdot c(w) + k' \cdot p(w), \forall w \in W
\]

Since \( k, k' \) (being both \( k \) and \( k' \) positive values) and \( \mu \) are all fixed values only depending on \( W \), we can trivially infer the following statement from Equation 7 given any \( w, w' \in W \):

\[
c(w) \leq c(w') \iff p(w') \leq p(w)
\]

\(^7\)In this proposition we do not assume any special property to the function \( \text{sim}(., .) \) for generalization. If \( \text{sim}(., .) \) were symmetrical (e.g. cosine similarity is symmetrical), we could simply define the pseudo-inverted compactness score as \( p(w) = \sum_{w_i \in W} \text{sim}(w_i, w) \), which would lead to a complexity of \( n \).