Meaning_space at SemEval-2018 Task 10: Combining explicitly encoded knowledge with information extracted from word embeddings

Pia Sommerauer **Antske Fokkens** **Piek Vossen**
Computational Lexicology & Terminology Lab (CLTL)
Vrije Universiteit Amsterdam, the Netherlands
{pia.sommerauer,antske.fokkens,piek.vossen}@vu.nl

Abstract

This paper presents the two systems submitted by the meaning_space team in Task 10 of the SemEval competition 2018 entitled *Capturing discriminative attributes*. The systems consist of combinations of approaches exploiting explicitly encoded knowledge about concepts in WordNet and information encoded in distributional semantic vectors. Rather than aiming for high performance, we explore which kind of semantic knowledge is best captured by different methods. The results indicate that WordNet glosses on different levels of the hierarchy capture many attributes relevant for this task. In combination with exploiting word embedding similarities, this source of information yielded our best results. Our best performing system ranked 5th out of 13 final ranks. Our analysis yields insights into the different kinds of attributes represented by different sources of knowledge.

1 Introduction

SemEval Task 10 "Capturing Discriminative Attributes" (Krebs et al., 2018) provides participants with triples of words consisting of two concepts and an attribute. The task is to determine whether the attribute is a distinguishing property of the first concept compared to the second concept. This is the case in triple *shrimp, spinach, pink*, for instance, because shrimp can be pink whereas spinach is usually of a different color. When the first concept does not have a semantic relation with the attribute or both the concepts have the same semantic relation with it, the attribute is considered not to be discriminative.

In general, Task 10 can be understood as detecting whether there is a semantic relation between the concepts and the attribute. The dataset includes a wide range of variation. For instance, the attribute may be a part of the concept (e.g. *tortoise, snail, legs*) or category membership (e.g. *polyurethane, polyester, material*), relations between entities and activities they engage in (e.g. *cheetah, lion, runs*) as well as rather specific relations, for instance the relation between a specialist and the phenomenon they are specialized in (e.g. *optician, dentist, eyes*).

Rather than finding specific solutions for each kind of relation, we investigate different approaches exploiting different sources of knowledge. Both of our systems comprise a component exploiting the glosses and hierarchical structure of WordNet (Fellbaum, 1998) in order to determine whether an attribute applies to a concept. Our underlying assumption is that definitions should provide the most important distinctive attributes of a concept. Since concepts are not necessarily always distinguished on the same level of concreteness, but might also be distinguished on a more abstract level (e.g. *herbs, root, green* v.s. *apse, nightgown, royal*) we exploit the entire WordNet hierarchy.

In both of our systems, the second component exploits information encoded in distributional vector representations of words. Word vectors have not only been shown to capture information about semantic similarity and relatedness but, beyond that, seem to encode information about individual components of word meaning that are necessary to solve analogy tasks such as in the famous example *man* is to *woman* as *king* is to *queen* (Mikolov et al., 2013b). This indicates that the dimensions of the distributional vector representations encode information about specific attributes of words. We experiment with two approaches: a basic approach comparing cosine similarities and an exploratory approach that deducts word vectors from one another to detect meaning differences. Best performance was obtained by the system using cosine similarity. The second approach per-

Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), pages 940–946
New Orleans, Louisiana, June 5–6, 2018. ©2018 Association for Computational Linguistics

forms lower in isolation, but performance is comparable to the first system in combination with the WordNet component.

The main insights gained from our experiments are the following. First, despite the limited coverage of information on attributes in WordNet (as pointed out by Poesio and Almuhareb (2005)), the contribution of the WordNet component to the overall results indicates that definitions yield a valuable source of knowledge with respect to discriminative attributes. Second, we analyze how individual systems perform across different types of attributes. Our analysis shows that similarity performs best on general descriptive properties and WordNet definitions help most for finding specific properties. These observations indicate that more sophisticated methods of combining these components could lead to superior results in future work.

The remainder of this paper is structured as follows: After presenting background and related work (Section 2), our system designs are introduced in Section 3. Section 4 provides an overview of the results achieved by different systems and system components, including our analysis across attribution types. This is followed by a conclusion (Section 5).

2 Background and related work

Solving the task at hand requires both knowledge about lexical relations and the world. We assume that this knowledge cannot be found in one resource alone. Rather, different approaches of representing word meaning may comprise complementary information. In this exploratory work, we exploit explicitly encoded knowledge in a lexical resource and information encoded in the distribution of words in large corpora. While attributes of concepts have been studied before from a cognitive (e.g. McRae et al. (2005)) and computational (e.g. Poesio and Almuhareb (2005)) perspective, this task is, to our knowledge, the first task aiming at detecting discriminative features.

We use WordNet (Fellbaum, 1998) as a source of explicitly represented knowledge. Whereas the WordNet structure contains a vast amount of information about lexical relations (hyponymy, synonymy, meronymy), its definitions constitute a resource of world knowledge. WordNet definitions have been used successfully in approaches to word sense disambiguation (Lesk, 1986) and inferring verb frames (Green et al., 2004). The only study requiring knowledge and reasoning about attributes we are aware of is an exploratory study examining what knowledge in definitions contributes to question-answering tasks (Clark et al., 2008).

Vector representations of word meaning based on the distribution of words in large corpora do not yield explicit information about specific relations, but implicitly encode all kinds of associations between concepts. In contrast to manually constructed resources, their coverage is much larger. More specifically, they have been shown to encode information relevant in solving analogy tasks (Mikolov et al., 2013a; Levy and Goldberg, 2014b; Gladkova et al., 2016; Gábor et al., 2017; Linzen, 2016) and inferring semantic hierarchies (Fu et al., 2014; Pocostales, 2016). This indicates that the dimensions of distributional representations encode information about attributes of concepts (Levy and Goldberg, 2014b, p.177). For instance, in order to find the fourth component in the analogy *man* is to *woman* as *king* is to *queen*, a model has to detect the relation holding between the pairs in the analogy. In this example, the relations are formed by the two features of royalty and gender. One way of solving this is to use the vector offsets resulting from *woman* - *man* + *king*. The result should be closest to the fourth component (*queen*). Thus, the first component for this calculation, B - A, should capture information about the distinguishing features between A and B, as the subtraction eliminates the identical (or very similar) dimensions in both representations, but keeps the features associated with B.

Our first system follows the basic assumption that if there is some kind of association between a concept and an attribute, this should be reflected by the vector representations. We assume that attributes occur in the linguistic contexts of the concepts they apply to and thus appear in proximity to them. In a comparative set-up such as in this task, the attribute should be closer to the concept it applies to. In our second system, we attempt to exploit the operations used for solving analogies in order to determine whether an attribute distinguishes two concepts.

3 System description

Each of our systems[1] consists of a WordNet component and a component exploiting word embed-

[1] Code can be found at `https://github.com/cltl/meaning_space`

ding vectors. If the WordNet component is unable to classify an example, it is passed on to the word embedding component. After presenting the WordNet component, we describe the two embedding-based systems. The word vectors used in all approaches are taken from the Word2Vec Google News model (Mikolov et al., 2013a).[2] The systems are developed using training and validation data and evaluated using test data.[3]

3.1 WordNet glosses and hierarchical structure

We design rules to exploit explicitly encoded knowledge in synset glosses and the hierarchical structure. We assume that (1) the most important discriminative attributes are mentioned in definitions and (2) concepts can be distinguished on different levels of abstraction. Essentially, we check whether the attribute is in any of the definitions of the concepts. We employ two variants of the system. The first variant simply relies on string match (definition string match), whereas the second one employs cosine similarity between the attribute and the words in the glosses (definition similarity). For both variants, we retrieve the glosses of all WordNet synsets containing the concepts and the glosses of all their hypernyms. We preprocess the definitions by tokenizing them and excluding stopwords using NLTK (Bird et al., 2009). In the two best-performing full systems, we used the definition-similarity variant.

3.1.1 Definition string match

This variant employs one rule to detect positive cases and two rules to detect negative cases:

POS The attribute matches a word in the glosses of concept 1 and no word in the glosses of concept 2.

NEG The attribute matches a word in the glosses of both concepts.

NEG The attribute matches a word in the glosses of concept 2 and no word in the glosses of concept 1.

We could also count all cases in which the attribute matches no word as negative cases, but since only a selection of attributes is mentioned in the glosses, this would yield a high number of false negative decisions. Instead, we fall back on one of two distributional approaches in case none of the above-mentioned cases applies. If run in isolation, we label all instances for which none of the conditions apply as negative.

3.1.2 Definition similarity

In the second variant, we replace words by their vectors and measure the cosine similarity between the attribute and the words in the glosses in order to determine positive and negative cases. As a first step, we search for the word in the glosses with the highest cosine similarity to the attribute. Next, we employ an upper and a lower threshold in order to determine whether the attribute is similar or dissimilar enough to the word in the glosses. We assume this strategy allows us to extend the scope from exact pattern match to highly similar words, such as synonyms or hypernyms. The transformed rules are shown below:

POS The similarity between concept 1 and the attribute is above the upper threshold and the similarity between concept 2 and the attribute is below the lower threshold.

NEG The similarity between concept 2 and the attribute is above the upper threshold and the similarity between concept 1 and the attribute is below the lower threshold.

NEG The similarity between concept 1 and the attribute and the similarity between concept 2 and the attribute are above the upper threshold.

NEG The similarity between concept 1 and the attribute and the similarity between concept 2 and the attribute are below the lower threshold.

Here, we do include a condition for cases in which both similarities are below the threshold, as we assume a wider coverage due to the vector representations. The best performing similarity thresholds (0.75 for upper and 0.23 for lower threshold) were determined by testing several configurations on the validation data.

In addition to glosses, several other kinds of semantic relations encoded in the WordNet hierarchy can be expected to increase the performance

[2] Downloaded from `https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit`, Nov. 1, 2017.

[3] `https://github.com/dpaperno/DiscriminAtt/`

on this task. We experimented with meronymy relations, synonymy as well as the hypernyms themselves. Whereas these additions increased the performance in a set-up consisting of only the WordNet component, they harmed the performance in our overall system set-ups where we also use word embeddings.

3.2 Word embeddings

As a second component, we employ two different ways of extracting semantic knowledge from word embeddings.

3.2.1 Vector similarity

This component compares the similarities between the attribute and the concept and counts all cases in which the similarity between concept 1 and the attribute is higher than the similarity between concept 2 and the attribute as positive cases. All other cases are counted as negative. Setting even low thresholds harmed performance.

3.2.2 Vector subtraction

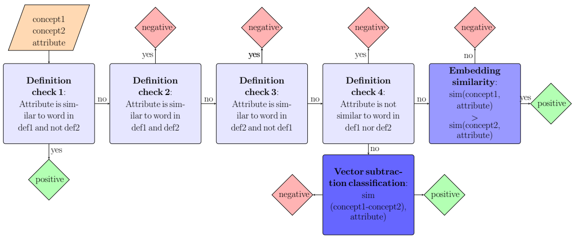
We subtract the vector of concept 2 from the vector of concept 1 and assume that the resulting vector representation is close to the kind of attribute that distinguishes the concepts. If this vector is close to the attribute, we assume it is discriminative. The subtraction should eliminate the shared aspects leaving information about the differences. The vector resulting from for instance *man - woman* cannot be seen as a representation of the specific word *male*, but rather reflects something like for instance 'maleness'.

Rather than setting a definitive similarity threshold, we employ a supervised classification approach in which we use the similarity between the calculated vector and the attribute vector as a sole feature. For reasons of time constraints, we only experimented with a Multi-Layer Perceptron, implemented in the SciKit learn toolkit (Buitinck et al., 2013).[4]

4 Results

Table 1 provides an overview of our different implementations on the test set. The highest performance was reached by the combination of WordNet gloss information and embedding similarity. In the overall SemEval ranking, performance ranges from 0.47 to 0.75. Our similarity-centered systems rank in the upper mid range (performing between 0.69 and 0.64), with the best run achieving 5th rank among 13 ranks (and 21 submitted system).

Figure 1: System overview.


concept1
concept2
attribute
negative
yes
Definition check 1: Attribute is similar to word in def1 and not def2
no
Definition check 2: Attribute is similar to word in def1 and def2
no
Definition check 3: Attribute is similar to word in def2 and not def1
no
Definition check 4: Attribute is not similar to word in def1 nor def2
no
Embedding similarity: sim(concept1, attribute) > sim(concept2, attribute)
yes
positive
yes
positive
no
Vector subtraction classification: sim (concept1-concept2), attribute)
negative
positive

System	P-pos	P-neg	R-pos	R-neg	F1-av
def-emb-sim	0.63	0.75	0.73	0.65	0.69
def-emb-sub	0.69	0.68	0.52	0.82	0.67
sim	0.58	0.73	0.75	0.56	0.64
def-sim	0.65	0.59	0.22	0.90	0.52
def	0.65	0.59	0.22	0.90	0.52
sub	0.45	0.55	0.19	0.82	0.46

Table 1: Performance overview of systems and system components on the test set.

The combination of WordNet gloss information and information from subtracting word vectors performs 2 points lower than our best performing system. When comparing the two word embedding approaches in isolation, we see that the system based on subtraction performs almost 18 points lower than the system using embedding similarity. This indicates that the overlap of correct answers between the Wordnet system and the subtraction system is lower than between the WordNet system and the embedding similarity system. The following sections provide insights into the level at which properties are found in WordNet definitions and the kinds of attributes successfully recognized by the different approaches.

System	P-pos	P-neg	R-pos	R-neg	F1-av
def-emb-sim	0.66	0.69	0.71	0.63	0.67
def-emb-sub	0.74	0.62	0.51	0.82	0.65
sim	0.60	0.64	0.72	0.51	0.61
sub	0.70	0.58	0.39	0.83	0.59
def-sim	0.70	0.54	0.24	0.90	0.52
def	0.70	0.54	0.24	0.90	0.52

Table 2: Performance of the systems and system components on the validation set.

[4] http://scikit-learn.org/stable/index.html, best configuration on the validation set: layer size: (1,1) activation function: logistic.

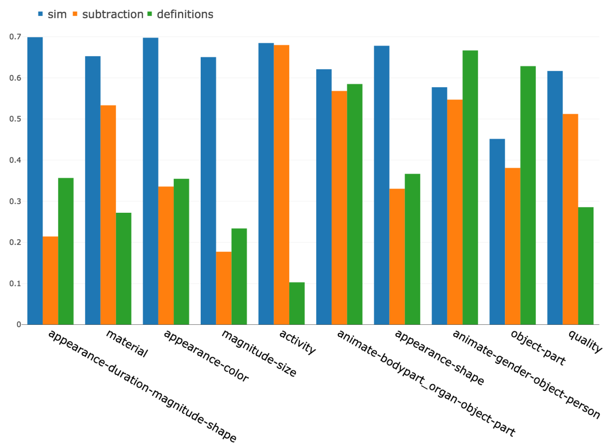


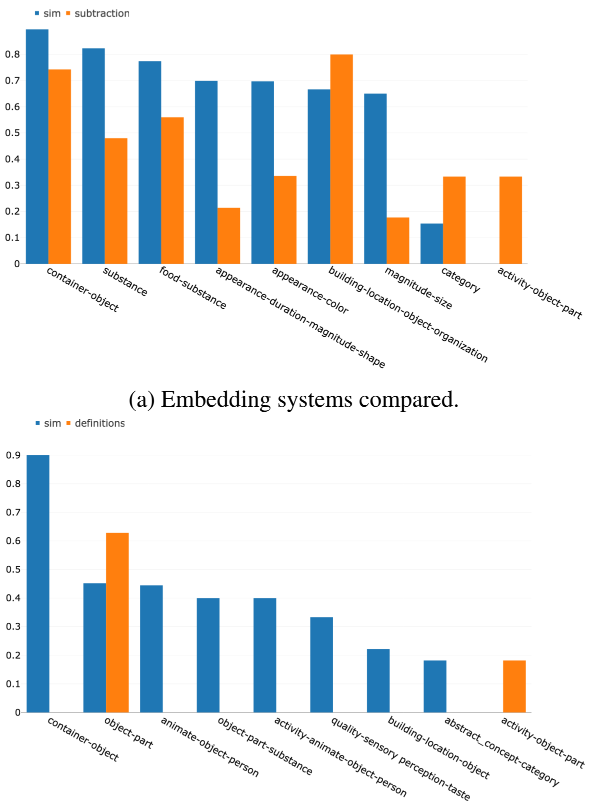
Figure 2: Comparison of f1-scores reached by the three main system components on the 10 most frequent attribute categories.

4.1 Level of distinction in the WordNet hierarchy

We hypothesized that concepts might not always be distinguished on the same level of concreteness, but could also be distinguished on a more abstract level. In order to test this, we counted how many properties are found in glosses of synsets containing the concept and how many are found in glosses of their hypernyms. Out of the total 1,098 attributes found in WordNet glosses, 699 are found on the same level as the concept (i.e. in the definitions of one of the synsets containing the concept) and 366 are found in gloss of one of the hypernyms of the synsets containing the concept.[5] In total, the definition system is able to classify 799 (out of 2,340) concept-concept-attribute triples.

4.2 Comparison of systems across attribute categories.

In this section, we aim at giving some insights into the kinds of attributes systems can detect and identify as discriminative. The attributes in the validation set were categorized by one annotator. The categories are not based on an existing framework, but were chosen intuitively. In most cases, multiple labels are used as the attributes are ambiguous. As there is no overlap between attributes from the validation set and the test set, we present the performance of the systems across the different attribute categories on the validation set. The overall performance on the validation set (presented in Table 2) is similar to the test set with the exception that on the validation set, the system based on vector subtraction performed several points higher than on the test set (0.59) and ranked higher than the WordNet definition systems.



(a) Embedding systems compared.

(b) WordNet definition match system compared to embedding similarity system.

Figure 3: Comparison f1-scores reached by the system components showing the categories with the highest performance differences (frequency >10).

Figure 2 shows the performance of the three individual systems across the 10 most frequent categories. Overall, the vector similarity system outperforms the other system in almost all categories. Of these 10 most frequent categories, there is no category in which the subtraction system outperforms the similarity system.

One of the most striking differences between the embedding-based systems and the WordNet definition system can be seen in the 'activity' category, in which both embedding systems perform almost seven times higher than the WordNet system. This could be explained by the fact that activities associated with concepts can be expected to occur in large corpora, whereas they might not

[5] Note that attributes can be found in glosses of both concepts, meaning that these counts do not add up to the number of triples in the test set.

be specific or relevant enough to be mentioned in a definition. In contrast, WordNet outperforms both embedding systems in the categories 'object-part' and 'animate-object-gender-person' (usually referring to a person of a specific gender), which could be expected to consist of more specific attributes.

We expected rather generic descriptions that are not specific to a concept, but mostly relevant in comparison to the other concept to be the most difficult for any system. These kinds of attributes are not relevant enough to be included in definitions nor do we expect them to frequently co-occur with concepts in texts and thus be apparent from a distributional model. It turned out, however, that these kinds of attributes ('appearance-color', 'magnitude-size') were accurately detected as discriminative by the embedding similarity system. A possible explanation might be that they can co-occur with a wide number of concepts, leading to proximity in the vector space.

When considering the categories with the biggest performance differences and a frequency of at least 10 (presented in Figure 3) the following observations can be made: Whereas the embedding similarity system outperforms the subtraction system in most categories, the subtraction system captures about twice as many attributes that indicate a category (usually meaning that the attribute is a hypernym of one of the concepts) and performs higher on 'building-location-object-organization' attributes (Figure 3a). It could be the case that despite their polysemy, these attributes apply to a more limited range of concepts than general descriptions on which the system performs poorly. The subtraction system also correctly detects attributes that are ambiguous between 'activity', 'object' and 'part' (e.g. attributes such as *sail* and *bark*) category, which is not detected by the similarity system. Finally, we observe that a number of attribute categories that are handled correctly by the embedding-based systems are not captured by WordNet definitions at all (Figure 3b).

Overall, the differences between the approaches seem to indicate that distributional models are stronger in capturing attributes expressing related concepts than attributes expressing similar concepts (e.g. hypernyms). This is in line with the general trend observed in large-scale evaluations (Levy et al., 2015; Baroni et al., 2014) of embedding models using a bag-of-words approach (such as the Google News model). Gamallo (2017) and Levy and Goldberg (2014a) show that embedding models using dependency structures perform better on similarity than relatedness and could thus improve the results for the attributes that are similar rather than related to the concepts.

5 Conclusion

For this SemEval task, we submitted systems consisting of combinations of exploratory approaches. Our best performing systems consisted of a component exploiting knowledge in WordNet definitions and a component extracting knowledge from distributional representations. In our best performing system, the latter component consisted of comparing cosine similarities between concepts and attributes. The vector component in our second full system employs a different strategy of extracting information from vector representations than our highest ranked system. Despite its limitations, its performance is comparable to our best performing system.

As expected, WordNet definitions encode rather specific attributes that are probably most informative for distinguishing one concept from another, while they give less importance to rather general descriptions. In contrast, embedding approaches seem to perform highly on attributes that are related rather than similar to the concepts, also encompassing rather general descriptions.

The main contribution of this paper is an exploration of the different types of attributes that can be recognized by different systems. These strengths and weaknesses of the methods could be further exploited by using the information obtained by vector similarity and subtraction as input of a classifier. We plan to investigate the representation of attribute categories in the semantic space in future work.

6 Acknowledgments

This research is funded by the PhD in the Humanities Grant provided by the Netherlands Organization of Scientific Research (Nederlandse Organisatie voor Wetenschappelijk Onderzoek, NWO) PGW.17.041 awarded to Pia Sommerauer and NWO VENI grant 275-89-029 awarded to Antske Fokkens.

References

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. 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)*, volume 1, pages 238–247.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

Peter Clark, Christiane Fellbaum, and Jerry Hobbs. 2008. Using and extending wordnet to support question-answering. In *Proceedings of the 4th Global WordNet Conference (GWC08)*. Citeseer.

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

Ruiji Fu, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Learning semantic hierarchies via word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1199–1209.

Kata Gábor, Haifa Zargayouna, Isabelle Tellier, Davide Buscaldi, and Thierry Charnois. 2017. Exploring vector spaces for semantic relations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1815–1824, Copenhagen, Denmark. Association for Computational Linguistics.

Pablo Gamallo. 2017. Comparing explicit and predictive distributional semantic models endowed with syntactic contexts. *Language Resources and Evaluation*, 51(3):727–743.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15.

Rebecca Green, Bonnie J Dorr, and Philip Resnik. 2004. Inducing frame semantic verb classes from wordnet and ldoce. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 375. Association for Computational Linguistics.

Alicia Krebs, Alessandro Lenci, and Denis Paperno. 2018. Semeval 2018 task 10: Capturing discriminative attributes. In *Proceedings of the 12th internations workshop on semantic evaluation (SemEval 2018)*.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26. ACM.

Omer Levy, Yaov Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. In *Transactions of the Association for Computational Linguistics*.

Omer Levy and Yoav Goldberg. 2014a. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 302–308.

Omer Levy and Yoav Goldberg. 2014b. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the eighteenth conference on computational natural language learning*, pages 171–180.

Tal Linzen. 2016. Issues in evaluating semantic spaces using word analogies. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 13–18.

Ken McRae, George S Cree, Mark S Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, 37(4):547–559.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. 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.

Joel Pocostales. 2016. Nuig-unlp at semeval-2016 task 13: A simple word embedding-based approach for taxonomy extraction. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1298–1302.

Massimo Poesio and Abdulrahman Almuhareb. 2005. Identifying concept attributes using a classifier. In *Proceedings of the ACL-SIGLEX Workshop on Deep Lexical Acquisition*, pages 18–27. Association for Computational Linguistics.