Synonymy and Antonymy Detection in Distributional Models

Shobhit Chaurasia, Tyler McDonnell
Department of Computer Science
The University of Texas at Austin
shobhit@utexas.edu, tyler@cs.utexas.edu
sc52987, tsm563

Abstract

We present the Distributional Sentiment Hypothesis for detecting synonymy and antonymy in distributional models. The Distributional Sentiment Hypothesis states that synonyms and antonyms tend to occur in similar narrow distributional contexts, but are distinguished by their broader tonal contexts. We show that sentimental polarity features computed using standard sentiment analysis tool-kits outperform pattern-based and narrow-context approaches in the literature and thus affirm the validity of this hypothesis. We also introduce an unsupervised method based on this hypothesis.

Introduction

Distributional models of semantic meaning are flexible representations of word meaning, in which words are represented through the textual contexts in which they appear (Turney, Pantel, and others 2010). Though these models have been widely adopted, one of the central criticisms of distributional models is that they fail to distinguish between semantic relations. For example, the words *good*, *bad*, and *evil* might all be considered semantically similar in a distributional model, even though {good, bad} and {good, evil} are antonym pairs with very different practical meanings. Due to this inability to distinguish between semantic relations of practical importance, Murphy (Murphy 2004) argues that distributional models cannot be valid models of conceptual representation.

Accordingly, much work has been dedicated to detecting specific semantic relationships in distributional models. The consensus seems to suggest that distributional models can distinguish between semantic relations, given suitable similarity measures. However, prior literature offers no convincing solution for resolving the synonymy and antonymy relationships in these models. We choose to focus on these two relationships in this work, as they are crucial to developing a valid model of conceptual representation and may have broader applicability to other applications, such as sentiment analysis.

We present a new guiding principle for resolving synonymy and antonymy in distributional models in the form of the following hypothesis:

Distributional Sentiment Hypothesis: Synonyms and antonyms tend to occur in similar narrow contexts, but are distinguished by their broader tonal contexts.

In its simplest interpretation, this hypothesis observes that many adjectival synonym and antonymy pairs have the same and opposite polarities, respectively.

In this work, we develop a methodology based upon a distributional space constructed according to the distributional sentiment hypothesis. We use features pulled from this distributional space to train a classifier and test our hypothesis by comparing the performance of classification using these features to that using features proposed by other methods in the literature. We also show the performance achievable by integrating the various techniques currently available throughout literature. Finally, we discuss the merit of our approach as a distributional technique and propose an alternative unsupervised formulation.

Distributional Models

Distributional models, also commonly referred to as vector space models of semantics (VSMs), model the semantic meanings of words through the contexts in which they are observed. Distributional models are based on the *Distributional Hypothesis*, which can be stated as "words that occur in the same contexts tend to have similar meanings" (Rubenstein and Goodenough 1965). Traditionally, distributional models model words as vectors in high-dimensional spaces, where dimensions are context items and the coordinates of the vector indicate the word's level of association with a context item.

The most common distributional model is the word-context vector, which typically establishes a window size N and builds the associated vector for each word in a corpus using co-occurrence counts of the N words to the left and right of each occurrence of a word throughout the corpus.

Pattern-Based Approaches

Early work on resolving semantic relationships between similar words in distributional models focused on the use of common textual patterns. (Lin et al. 2003) propose the following hand-crafted patterns for identifying between synonyms and antonyms:

- either X or Y

- neither X nor Y
- from X to Y

A score based on the number of occurrences of the above patterns in a corpus can then be used as a feature for supervised classification. Elsewhere, pattern-based approaches have been applied to many other semantic relations, including meronymy (Berland and Charniak 1999; Girju, Badulescu, and Moldovan 2006; 2003), co-hyponymy (Snow, Jurafsky, and Ng 2004), and hypernymy (Buitelaar, Cimiano, and Magnini 2005) We integrate pattern-based approaches into our study as a point of comparison during evaluation.

The Age of Hypothesis

In evolving from early pattern-based approaches, literature has often developed hypotheses that attempt to capture the distributional nuances of specific semantic relationships. These hypotheses share a common theme: distributional context can, in fact, distinguish between semantic relationships given appropriate similarity measures.

For instance, the *Distributional Inclusion Hypothesis* is a specialization of the distributional hypothesis which states that more specific terms appear in a subset of distributional contexts than their more general counterparts. This hypothesis was integrated into an approach for hypernymy detection proposed by (Baroni et al. 2012), in which a pair of two words are represented as the component-wise difference between their classical distributional vectors. The resulting difference vector may be used as a feature vector for classification. Though this approach was initially not considered a success, (Roller, Erk, and Boleda 2014) found that this difference vector was in fact very effective for classification given three modifications: (1) use of a linear classifier; (2) vectors should be normalized to have a magnitude of 1 prior to taking the difference; and (3) squared difference vectors should be included as features. Our work integrates the idea of a difference vector and the three modifying guidelines given by (Roller, Erk, and Boleda 2014) into an unsupervised approach based on our Distributional Sentiment Hypothesis.

For the case of synonymy and antonymy, (Scheible, Im Walde, and Springorum 2013) suggest that only certain word classes provide useful contextual information for synonyms and antonyms and build separate word vectors using each of the four common *content word* classes - noun, verb, adjective, and adverb. By training and evaluating a supervised classifier which uses as features the cosine similarities in these four distributional spaces, they conclude that verbs are the most important narrow-context word class for determining synonym. Our Distributional Sentiment Hypothesis and related approach consider a wider contextual part-of-speech window which we believe expands upon their claims.

The Distributional Sentiment Hypothesis

The primary goal of this paper is to provide and evaluate a new guiding hypothesis for resolving synonymy and antonymy in distributional models. The *Distributional Sentiment Hypothesis* states that synonyms and antonyms tend to occur in similar narrow distributional contexts, but are distinguished by their broader tonal contexts. We believe that this hypothesis is obviously true for many adjectival synonyms and antonyms with identical and opposite sentimental connotations, respectively, such as {good, great} (synonyms) or {good, bad} (antonyms). These words carry a great deal of tonal information themselves which can be used to usefully distinguish between synonyms and antonyms. However, it is not obvious whether or not this principle extends to pairs of words which are neutral or not clearly related under such opposite polarities.

We contend that this principle does extend to other pairs of synonyms and antonyms, given that we respect the *broader context* specification. Consider the antonym pair {up, down}. In this case, the two words are neutral, but are antonymous. Though it may not seem as if the words are tonally related, consider the following sentences:

1. A *great* day for investors: the market was *up* 50 points.
2. The Wall Street crash of 1929 was the most *devastating* stock market crash in the history of the United States, with the overall market *down* almost 40% in a matter of days.

In these sentences, you can see that the words up and down co-occur with positive and negative words, respectively. We contend that even neutral and non-adjectival synonyms and antonyms tend to co-occur with certain classes of words, most generally captured by tone.

Unfortunately, tone is a very complex catch-all that we use for illustrative convenience and do not attempt to precisely define in this paper. Rather, we contend that tonal context can be approximated using common methods of sentiment analysis. Sentiment scores can be computed to reflect positive or negative connotations attached to a word, sentence, or larger body of text. We will approximate the tonal distributional context of words by computing sentiment-based distributional spaces.

Data

We constructed several distributional spaces based on an array of features drawn from both prior literature and our own distributional sentiment hypothesis. Each of these spaces was built upon a subset 10,000 e-books from Project Gutenberg, a volunteer effort to digitize and archive high quality e-books for public domain. The resulting corpus contains approximately 317,000 unique word tokens and 1.5 billion total tokens. We chose Project Gutenberg over other common corpora in this space, such as Wikipedia dumps, because we believe the latter includes an unnecessary amount of noise per document due to the large number of technical documents across widely varying subject domains.

G-S: This standard distributional space was built by counting co-occurrences of word types within a window of three words on the left and right of the target word, respecting the sentence boundary. Stop words omitted from the co-occurrence vectors, and all words were stemmed and lemmatized.

G-POS (G-NOUN, G-VERB, G-ADV, G-ADJ): These distributional spaces were constructed similarly to G-S, except only co-occurrences of the nearest two nouns, verbs, adverbs, or adjectives, respectively, to the left and right of a word were included in its context vector. These vectors also respect the sentence boundary. Since the e-books included in Project Gutenberg are raw text and not POS-tagged, we first POS-tagged the entire corpus using the NLTK POS Tagger. We then removed stop words and stemmed and lemmatized all tokens prior to building the context vectors. These spaces are collectively referred to as **G-POS**

G-PATTERN (G-NEITHER, G-EITHER, G-FROM): These spaces were built to represent the three patterns proposed by (Lin et al. 2003), *neither x nor y*, *either x or y*, and *from x to y*, where *x* and *y* may be any token. To construct each of these spaces, we identified occurrences of the pattern throughout the corpus and then stemmed and lemmatized the variable tokens x and y. The final vector representations of x and y are then co-occurrence counts of x and y under the respective pattern. These spaces are collectively referred to as **G-PATTERN**

G-SENT: The G-SENT space contains vectors meant to describe the positive or negative distributional polarity of a word. To construct this space, we first computed a sentiment score for every sentence in the corpus using the Valence Aware Dictionary and Sentiment Reasoner (VADER) (Gilbert 2014). VADER is a lexicon and rule-based sentiment analysis tool specifically tuned for social media. We chose VADER despite our literary corpus over other more sophisticated alternatives primarily for its speed. After computing the sentiment score for each sentence, we constructed context vectors along the following 6 dimensions:

1. *Same Sentence Intensity (SSI)*: the cumulative sum of the sentiment intensities of the sentences which contain the token.
2. *Same Sentence Positive Count (SSP)*: a count of the number of sentences containing the token with a positive sentiment.
3. *Same Sentence Negative Count (SSN)*: a count of the number of sentences containing the token with a negative sentiment.
4. *Adjacent Sentence Intensity (ASI)*: the cumulative sum of the sentiment intensities of the sentences before and after the one containing the token.
5. *Adjacent Sentence Positive Count (ASP)*: a count of the number of sentences with positive sentiment before and after those containing the token.
6. *Adjacent Sentence Negative Count (ASN)*: a count of the number of sentences with negative sentiment before and after those containing the token.

Co-occurrence Information

Instead of using raw counts, the co-occurrence information in the above distributional spaces was encoded by Local Mutual Information (LMI). In addition to the observed co-occurrence frequency O, LMI also takes into account the expected co-occurrence frequency E, appropriately weighting the observed raw counts.

$$LMI = O \times \log\left(\frac{O}{E}\right) \quad (1)$$

Using raw counts has the side-effect of highly co-occurring word-pairs, whose high co-occurrence is mainly because of the high individual occurrence of their members, overshadowing other word-pairs whose members have low individual occurrence, but unexpectedly high co-occurrence. LMI mitigates this problem by penalizing high-frequency events that are indeed expected to be highly frequent.

Experimental Evaluation

For evaluation, we use information from the previously described distributional spaces to train and evaluate a supervised classifier. Given an input feature vector V representing a word pair $\{w_1, w_2\}$, we evaluate the model based on its ability to correctly label the pair as *synonyms*, or *antonyms*.

Our dataset consisted of synonym and antonym pairs. We chose to focus on common words to avoid encountering words that were under-specified across our corpus. To build our synonym and antonym pairs, we selected a seed set of 400 common English adjectives across our corpus. For each of these adjectives, we pulled the *common word* synonyms and antonyms using the www.thesaurus.com API, which relates data from Roget's 21^{st} Century Thesaurus. This formed between 1 and 8 pairs of synonyms/antonyms for each word, resulting in a total of 1557 synonym pairs and 1654 antonym pairs.

Feature Representation

For supervised classification, we derived numerous features from the distributional spaces to form V, the vector representation of a word pair, $\{w_1, w_2\}$.

Distributional Features: This set of features were based on the word-context vectors constructed from the G-S, G-NOUN, G-VERB, G-ADV and G-ADJ spaces. For each word pair $\{w_1, w_2\}$ the Cosine Similarity of the word-vectors for w_1 and w_2 across each of these models was used as input features. This results in five features for each word-pair which are related to their distributional spaces.

Sentiment Features: This set of features is derived from the G-SENT space. For each word-pair $\{w_1, w_2\}$, the following three sentiment related features are used:

1. The absolute difference between the word-level sentiment score of w_1 and w_2. The difference should be high for an antonym word-pair, and low for a synonym word-pair. This feature is particularly useful for subjective words, such as good, bad, angry, etc.

2. The absolute difference between the average sentiment intensity of the sentences containing w_1 and w_2 respectively. This feature augments the word-level sentiment feature described above, and accounts for sentiment of the local context in which the words appear.
3. The absolute difference between the average sentiment intensity of the sentences adjoining the ones which contain w_1 and w_2 respectively. This feature captures a wider context, and could be useful for words which do not posses a subjective interpretation of their own, but are often used with a particular connotation.

Pattern Features This set includes the occurrence counts from the G-NEITHER, G-EITHER, G-FROM space. The intuition is that word-pairs $\{w_1, w_2\}$ that frequently occur in one of the above three patterns are likely to be antonyms (except Nouns, such as names of places, that frequently occur in the "*from* w_1 *to* w_2" pattern).

Supervised Classification

We trained numerous classifiers, such as Logistic Regression, and SVM, using different subsets of the features described above on 80% of the our dataset, consisting of 1245 synonym pairs and 1324 antonym pairs. The remaining 20%, consisting of 312 synonym pairs and 331 antonym pairs, was used as a held-out set for evaluation.

Results

The performance of a classifier trained using different subset of features is summarized in Table. 1. These results correspond to the classifier and the corresponding hyper-parameter setting that performed the best on most of the feature subsets - logistic regression. Hyper-parameter tuning was done using 5-fold Cross-Validation on the training set.

Pattern-based Features The most obvious observation corresponds to the performance of the classifier trained only on pattern-based features (G-PATTERN). By construction, only antonym-pairs are highly likely to occur in those patterns. The trained classifier ended up assigning the Antonym class to every word-pair, thereby achieving a perfect Recall for antonyms, and an undefined Precision for synonyms, and an undefined F1 score. Pattern-based features are very brittle, and not useful for generalization to antonym pairs not present in the corpus in the specified patterns. The fact that G-PATTERN has the lowest Precision for antonyms among all feature subsets supports this claim.

Distributional Features The result that stands out the most is the unusually good performance of the classifier with respect to Antonym identification trained only on the distributional features (G-S, G-POS) individually. While the Precision for antonyms and the overall accuracy is not very high, the Recall is over 25 points higher than the rest of the feature subsets. This further corroborates the criticism of distributional models with regards to their inclusion of antonyms in the space of *semantic relatedness* that they induce. However, the unusually low F1 scores with respect to Synonyms are counter-intuitive and inexplicable.

(Scheible, Im Walde, and Springorum 2013) observed that word-context vectors based on specific Part-of-Speech, such as Verb (in our case, G-POS) posses more discriminative power with regards to Antonym identification than word-context vectors built generically (in our case, G-S). In our experiments, the F1 scores obtained by the two sets of features are the same, and the other metrics are also roughly equal. This could potentially be because of the difference in the size of the context window used in the two sets features, 2 for G-POS, and 3 for G-S.

Sentiment Features While the classifier trained only on Sentiment features (G-SENT) does not outperform others on any domain, it achieves balanced scores on both synonym and antonym identification. Its consistent scores on identification of both classes makes it a useful feature for this task.

Pooling of all features The best F1 score is obtained by the classifier trained on the pool of all the features. An interesting point to note is that the performance of classifier trained on {G-SENT, G-PATTERN, G-S} is marginally better than the one trained on {G-SENT, G-PATTERN, G-POS}. This is, again, in conflict with the observations in (Scheible, Im Walde, and Springorum 2013) that Part-of-Speech based word-context vectors (G-POS) are better suited for antonymy identification.

The Ideology of Distributional Models

In many ways, sentiment is an intuitive feature for discerning between synonym and antonym word pairs. As we have discussed, many antonym word pairs (e.g., {good, bad}) obviously have opposite sentimental polarities. Moreover, even antonym word pairs which may be considered neutral (e.g., {up, down} or {truth, lie}) may often be distinguishable because their constituents are used in contexts with different sentimental polarity. Indeed, our results imply that these sentimental features may be more powerful and robust for discerning synonymy and antonymy than previous distributional similarity measures.

However, an interesting question is whether or not our methods clash with the essence of distributional models. In a sense, our sentiment-based approach is similar to other distributional spaces: we are building a representation for each word purely based on characteristics of its context throughout the corpus. At the same time, the tool underlying our sentiment analysis relies on a manually annotated lexicon of word scores. Thus, can we truly claim that our approach offers *flexible* representations of word meaning?

In fact, in some ways, supervised learning in general is antithetical to the the essence of distributional models, since it relies on a "dictionary" of examples that are not induced over the arbitrary corpus of interest and indeed may not be naturally transferable across time or corpora. Thus, one might argue that supervision is at odds with the flexibility and generality of distributional models of semantic meaning. For instance, though (Roller, Erk, and Boleda 2014) achieve

Feature Subset	Precision		Recall		F1		Accuracy
	Syn	Ant	Syn	Ant	Syn	Ant	
G-S	0.59	0.53	0.20	0.87	0.30	**0.66**	0.54
G-POS	0.57	0.53	0.15	**0.89**	0.24	**0.66**	0.53
G-S, G-POS	0.58	0.56	0.40	0.72	0.47	0.63	0.57
G-SENT	0.60	0.66	0.69	0.57	0.64	0.61	0.63
G-PATTERN	*NA*	0.51	0.0	*1.0*	*NA*	*0.68*	0.51
G-SENT, G-PATTERN	0.63	0.70	0.72	0.60	0.67	0.65	0.66
G-SENT, G-PATTERN, G-S	**0.64**	**0.72**	0.75	0.60	0.69	**0.66**	0.68
G-SENT, G-PATTERN, G-POS	0.63	0.70	0.73	0.59	0.67	0.64	0.66
G-SENT, G-PATTERN, G-S, G-POS	**0.64**	**0.72**	**0.76**	0.61	**0.70**	**0.66**	**0.68**

Table 1: Performance of classifier trained on various subset of features

excellent accuracy for hypernymy detection in distributional models using a supervised approach, their model only managed to learn important context features that were specific to the topical domain of their evaluation dataset. It is unclear how their approach would be generalized to an arbitrary topical domain without appropriate training data.

Despite many such criticisms, the literature has shown us that supervised learning is still an invaluable tool for validating distinguishing features of semantic relations in distributional models. In this case, we believe our results lend validity to our distributional sentiment hypothesis. Assuming that our distributional sentiment hypothesis is valid, perhaps we can move beyond a lexicon-based approach to one which more strictly adheres to the ideology of distributional models. Indeed, we know that the information necessary for identifying synonymy and antonymy *must be encoded in the distributional context of a word*, since the sentiment scores used to build our classifier were based purely on this information (i.e., the parent and adjacent sentences of the target word).

Inducing Sentiment: A "Good" Case Study

Here we propose an unsupervised and "devoutly" distributional approach for distinguishing synonymy and antonymy in distributional models, as well as a case study illustrating this methodology. This is approach is based on our distributional sentiment hypothesis, and indirectly captures the notion of tone without the need for a manually graded lexicon.

We begin with the word "good" and isolate ten words which are either synonyms or antonyms of good and are distributionally similar to good as measured by the cosine similarity in the standard distributional space G-S: {greater, wonderful, excellent, marvelous, positive, bad, evil, lousy, awful, poor}. The first five of these words are synonyms of good, whereas the last five are antonyms. We used k-means to cluster these ten words into two clusters using the G-S context vectors, hoping to induce a separation between synonyms and antonyms. The results are shown in Table 2. We interpret those words included in the same cluster as good to be synonyms and those included in the opposing cluster as antonyms. Unsurprisingly, the results are very poor. This echoes the notion that distributional models, at least in their simple narrow-context definition, do no not meaningfully distinguish between all semantic relations.

Intuition and manual investigation of sentiment analysis lexicons suggest that adjectives play a large role in defining the tone or sentimental polarity of a sentence or body of text. We leverage this observation in conjunction with the broader context supposition of our distributional sentiment hypothesis to build our approach. We define a new distributional space, G-ADJ=SENT which, for each occurrence of a target word, counts the co-occurrences of all adjectives in the parent and adjacent sentences. Note that this approach is similar to our method for computing sentiment scores when building G-SENT, except in this case we do not know the scores associated with any of this contextual information. Then, for each word, w_i, in our set of 10 words, we compute the difference between the normalized G-ADJ=SENT context vector of good and w_i. We then cluster the resultant vectors, one for each word pair.

Again, the intuition behind this approach is based on the distributional sentiment hypothesis. Though much of the distributional context is similar between synonyms and antonyms, there are distinguishing characteristics in the form of broader adjective context which are intimately related to the common sentimental polarity of the words. We capture that by first collecting the broader adjective context and isolating the portions that are unique to different word combinations.

The results of this clustering are shown in Table 3. In this case, the clustering correctly classifies all ten similar words by synonymy or antonymy. Additionally, the actual sentiment categories for each of these 10 words as determined by the VADER sentiment analysis tool are shown in Table 4. If we interpret synonyms as words of similar polarity and antonyms as word of opposite polarity, roughly in line with our distributional sentiment hypothesis, we not only see that this hypothesis holds for this particular example, but that we

Synonyms	Antonyms
great	poor
wonderful	
excellent	
marvelous	
positive	
bad	
evil	
lousy	
awful	

Table 2: G-S Clusters

Synonyms	Antonyms
great	poor
wonderful	bad
excellent	evil
marvelous	lousy
positive	awful

Table 3: G-ADJ=SENT Clusters

Positive	Negative
great	poor
wonderful	bad
excellent	evil
marvelous	lousy
positive	awful
good	

Table 4: Actual Sentiment

were able to effectively induce the sentiment labels for each of the words.

Future Work

We outline a few obvious areas of future work. First, an ideal study would have used our entire corpus to build the distributional spaces, but due to the enormous computational resources needed and limited time to complete this study, we were unable to do so. This was notable in some cases, where words used in our evaluation data had relatively small distributional context vectors, suggesting that perhaps a larger corpus might have yielded better results. Due to the same computational constraints, we were unable to thoroughly evaluate our unsupervised model based on clustering. An ideal study would have reproduced the same classifications done using our supervised G-SENT space to offer a direct performance comparison.

Nevertheless, this small-scale case study suggest that our clustering approach might yield new approaches for approximating sentiment relations between words without manual grading. In our small case study, we knew which word was the positive word and simply labeled it and all of the words clustered with it with the same sentiment category. More generally, it would be interesting to deduce positive and negative categories more generally across all words in a corpus. Future work might also consider if distributional models can reproduce not only the sentiment categories, but also the polarities of words.

Conclusion

In this paper, we introduce and test the *Distributional Sentiment Hypothesis*, a guiding principle for detecting synonymy and antonymy in distributional models. The Distributional Sentiment Hypothesis states that though synonyms and antonyms tend to occur in similar narrow contexts, they are distinguished by the tone of their broader contexts. We evaluated this hypothesis by approximating the tonal contexts of words by computing same-sentence and adjacent-sentence sentiment polarity scores using the VADER sentiment analysis tool. We showed that these sentimental polarity features offer higher supervised classification accuracy than previous approaches in the literature, including pattern-based features and narrow-context part-of-speech vectors.

Additionally, we presented a case study and model for unsupervised classification based upon the Distributional Sentiment Hypothesis. This method induces the information of the sentimental polarity scores by first capturing the broad adjectival distributional context, rather than the narrow context, of words in a corpus and then clustering pairs of distributionally similar words based on the difference between their broader adjectival context. In a small case study, we show how this approach can be used to classify synonyms and antonyms, thus further validating our Distributional Sentiment Hypothesis. Furthermore, we believe this unsupervised approach appeals more truly to the flexibility and generality promised by distributional models while offering an avenue for inducing sentiment categories relations without manually graded lexicons.

References

Baroni, M.; Bernardi, R.; Do, N.-Q.; and Shan, C.-c. 2012. Entailment above the word level in distributional semantics. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 23–32. Association for Computational Linguistics.

Berland, M., and Charniak, E. 1999. Finding parts in very large corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, 57–64. Association for Computational Linguistics.

Buitelaar, P.; Cimiano, P.; and Magnini, B. 2005. *Ontology learning from text: methods, evaluation and applications*, volume 123. IOS press.

Gilbert, C. H. E. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text.

Girju, R.; Badulescu, A.; and Moldovan, D. 2003. Learning semantic constraints for the automatic discovery of part-whole relations. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, 1–8. Association for Computational Linguistics.

Girju, R.; Badulescu, A.; and Moldovan, D. 2006. Automatic discovery of part-whole relations. *Computational Linguistics* 32(1):83–135.

Lin, D.; Zhao, S.; Qin, L.; and Zhou, M. 2003. Identifying

synonyms among distributionally similar words. In *IJCAI*, volume 3, 1492–1493.

Murphy, G. 2004. *The big book of concepts*. MIT press.

Roller, S.; Erk, K.; and Boleda, G. 2014. Inclusive yet selective: Supervised distributional hypernymy detection. In *COLING*, 1025–1036.

Rubenstein, H., and Goodenough, J. B. 1965. Contextual correlates of synonymy. *Communications of the ACM* 8(10):627–633.

Scheible, S.; Im Walde, S. S.; and Springorum, S. 2013. Uncovering distributional differences between synonyms and antonyms in a word space model. In *IJCNLP*, 489–497. Citeseer.

Snow, R.; Jurafsky, D.; and Ng, A. Y. 2004. Learning syntactic patterns for automatic hypernym discovery. *Advances in Neural Information Processing Systems 17*.

Turney, P. D.; Pantel, P.; et al. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research* 37(1):141–188.