Uncovering Distributional Differences between Synonyms and Antonyms in a Word Space Model

Silke Scheible, Sabine Schulte im Walde and Sylvia Springorum
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
Pfaffenwaldring 5b, 70569 Stuttgart, Germany
{scheible,schulte,riestesa}@ims.uni-stuttgart.de

Abstract

For many NLP applications such as Information Extraction and Sentiment Detection, it is of vital importance to distinguish between synonyms and antonyms. While the general assumption is that distributional models are not suitable for this task, we demonstrate that using suitable features, differences in the contexts of synonymous and antonymous German adjective pairs can be identified with a simple word space model. Experimenting with two context settings (a simple window-based model and a 'co-disambiguation model' to approximate adjective sense disambiguation), our best model significantly outperforms the 50% baseline and achieves 70.6% accuracy in a synonym/antonym classification task.

1 Introduction

One notorious problem of distributional similarity models is that they tend to not only retrieve words that are strongly alike to each other (such as *synonyms*), but also words that differ in their meaning (i.e. *antonyms*). It has often been argued that this behaviour is due to the distributional similarity of synonyms and antonyms: despite conveying different meanings, antonyms also seem to occur in very similar contexts (Mohammad et al., 2013).

In many applications, such as information retrieval and machine translation, the presence of antonyms can be devastating (Lin et al., 2003). While a number of approaches have addressed the issue of synonym and antonym distinction from a computational point of view, they are usually limited in some way, for example by requiring the antonymous words to co-occur in certain patterns (Lin et al., 2003; Turney, 2008), or by relying on external resources such as thesauri (Mohammad et al., 2013; Yih et al., 2012). Probably due to this strong similarity of their contexts, there have been no successful attempts so far to distinguish the two relations via a standard distributional model such as the word space model (Sahlgren, 2006).

Prominent work in psycholinguistics, however, has shown that humans are able to distinguish the contexts of antonymous words, and that these are by no means interchangeable (Charles and Miller, 1991). The goal of our research is to show that using suitable features these differences can be identified via a simple word space model, relying on contextual clues that govern the ability to distinguish the relations in context. For this purpose, we present a word space model that exploits window-based features for synonymous and antonymous German adjective pairs. Next to investigating the contributions of the various parts-of-speech with regard to the word space model, we experiment with two context settings: one that takes into account all contexts in which the members of the word pairs occur, and one where we approximate context disambiguation by applying 'co-disambiguation': establishing the set of nouns that are modified by both members of the pair, and only including distributional information from contexts in which the adjectives premodify one of the set of shared nouns. Two different scenarios relying on Decision Trees then assess our main hypothesis, that the contexts of adjectival synonyms and antonyms are distinguishable from each other.

Our paper is structured as follows: Section 2 reviews some of the theoretical and psycholinguistic hypotheses and findings concerning synonymy and antonymy, and Section 3 reviews previous approaches to synonym/antonym distinction. Based on these theoretical and practical insights, we introduce our hypotheses and approach in Section 4. Section 5 describes the data and implementation of the word space model used in our experiments, and, finally, in Section 6 we discuss our findings.

International Joint Conference on Natural Language Processing, pages 489–497,
Nagoya, Japan, 14-18 October 2013.

2 Theoretical background

Synonymy and antonymy are without doubt two of the most well-known semantic relations between words, and can be broadly defined as words that are 'similar' in meaning (synonyms), and words that are 'opposite' in meaning (antonyms).[1] The fascinating issue about antonymy is that even though antonymous words are said to be opposites, they are nevertheless semantically very similar. Cruse (1986) observes that there is a notion of simultaneous closeness and distance from one another, and notes that this can be partially explained by the fact that opposites share the same semantic dimension. For example, the antonyms *hot* and *cold* share the dimension 'TEMPERATURE', but unlike synonyms, which are located at identical or close positions on the dimension (such as *hot* and *scorching*), antonyms occupy opposing poles (cf. the schematic representation in Figure 1). Antonymous words are thus similar in all respects but one, in which they are maximally opposed (Willners, 2001).

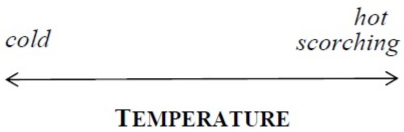

Figure 1: Semantic dimension

There has been extensive work on linguistic and cognitive aspects of synonyms and antonyms (Lehrer and Lehrer, 1982; Cruse, 1986; Charles and Miller, 1989; Justeson and Katz, 1991). Both relations have played a special role in the area of distributional semantics, which investigates how the statistical distribution of words in context can be used to model semantic meaning. Many approaches in this area are based on the *distributional hypothesis*, that words with similar distributions have similar meanings (Harris, 1968).

In a seminal study, Rubenstein and Goodenough (1965) provided support for the distributional hypothesis for synonyms by comparing the collocational overlap of sentences generated for 130 target words (i.e. 65 word pairs ranging from highly synonymous to semantically unrelated) with synonymy judgements for the pairs, showing that there is a positive relationship between the degree of synonymy between a word pair and the degree to which their contexts are similar. The situation for antonyms with respect to the distributional hypothesis has however been less clear. In fact, Charles and Miller (1991) used the contextual distribution of antonyms to argue *against* the reliability of the co-occurrence approach: they measured how often antonyms co-occur within the same sentence (for example, in contrastive constructions such as 'either x or y'), and show that the co-occurrence counts for antonyms such as *big/little*, and *large/small* in the Brown corpus are larger than chance.[2] Charles and Miller claim that the fact that antonyms tend to co-occur in the same contexts constitutes a true counter-example to the co-occurrence approach: they display high contextual similarity, but are of low semantic similarity.

As an alternative to the co-occurrence approach, Charles and Miller (1991) proposed a technique based on *substitutability* (cf. also Deese (1965)). Here, the contextual similarity of synonyms/antonyms is determined by presenting human subjects with sentences in which the occurrences of the two words have been blanked out, and by assessing the amount of confusion between the words when asking the subjects which word belongs in which context. While, as anticipated, the level of confusion was high for synonyms, subjects rarely confused the sentential contexts of antonyms, contrary to Charles and Miller's expectations. They had assumed that direct antonyms[3] such as *strong/weak*, or *powerful/faint*, were interchangeable in most contexts, based on the insight that any noun phrase that can be modified by one member of the pair can also be modified by the other. However, human subjects were very efficient at identifying the correct antonym.

Charles and Miller's findings suggest that in contrast to synonyms, whose distributional properties are similar, there are clear contextual differences that allow humans to distinguish between the members of an antonym pair. In this paper we aim to show that these differences can be detected with a simple distributional word space model, thereby refuting the claim that antonyms are a counter-example to the co-occurrence approach.

3 Previous computational approaches

Due to their special status as both 'similar' and 'different', work in computational linguistics has sometimes included antonymy under the heading

[1] In the following, we work with this simple definition of the two relations. For an account of other, more complex definitions, please refer to Murphy (2003).

[2] Similar results were found by Justeson and Katz (1991).

[3] Commonly associated adjectives (Paradis et al., 2009).

of semantic similarity. Recent research however has called for a strict distinction between *semantic similarity* (where entities are related via likeness) and *semantic relatedness* (where dissimilar entities are related via lexical or functional relationships, or frequent association), cf. Budanitsky and Hirst (2006). Accordingly, antonyms fall into the broader category of 'semantic relatedness', and should not be retrieved by measures of semantic similarity. That this is of crucial importance was highlighted by Lin et al. (2003), who noted that in many NLP applications the presence of antonyms in a list of similar words can be devastating.

A variety of measures have been introduced to measure semantic similarity, for example by drawing on lexical hierarchies such as WordNet (Budanitsky and Hirst, 2006). In addition, there are corpus-based measures that attempt to identify semantic similarities between words by computing their distributional similarity (Hindle, 1990; Lin, 1998). While these are efficient at retrieving synonymous words, they fare less well at identifying antonyms as non-similar words, and routinely include them as semantically similar words. However, despite the problems resulting from this, there have only been few approaches that explicitly tackle the problem of synonym/antonym distinction, rather than focussing on only synonyms (e.g. Edmonds (1997)) or antonyms (e.g. de Marneffe et al. (2008)).

Lin et al. (2003), who implemented a similarity measure to retrieve distributionally similar words for constructing a thesaurus, were one of the first to propose methods for excluding retrieved antonyms. Lin's measure uses dependency triples to extract distributionally similar words. In a post-processing step, they filter out any words that appear with the patterns 'from X to Y' or 'either X or Y' significantly often, as these patterns usually indicate opposition rather than synonymy. They evaluate their technique on a set of 80 synonym and 80 antonym pairs randomly selected from Webster's Collegiate Thesaurus that are also among their top-50 list of distributionally similar words, and achieve an F-score of 90.5% in distinguishing between the two relations.

Turney (2008) also tackles the task of distinguishing synonyms from antonyms as part of his approach to identifying analogies. Like Lin et al. (2003), he relies on a pattern-based approach, but instead of hand-coded patterns, his algorithm uses seed pairs to automatically generate contextual patterns (in which both related words must appear). Using ten-fold cross-validation, his approach achieves an accuracy of 75.0% on a set of 136 'synonyms-or-antonyms' questions, compared to a majority class baseline of 65.4%.

A recent study by Mohammad et al. (2013), whose main focus is on the identification and ranking of opposites, also discusses the task of synonym/antonym distinction. Using Lin (2003) and Turney (2008)'s datasets, they evaluate a thesaurus-based approach,[4] where word pairs that occur in the same thesaurus category are assumed to be close in meaning and marked as synonyms, while word pairs occurring in contrasting thesaurus categories or paragraphs are marked as opposites. To determine contrasting thesaurus categories, Mohammad et al. rely on what they call the 'contrast hypothesis'. Starting with a set of seed opposites across thesaurus categories, they assume that all word pairs across the respective contrasting categories are also contrasting word pairs. The method achieves 88% F-measure on Lin et al. (2003)'s dataset (compared to Lin's 90.5%), and 90% F-measure on Turney (2008)'s set of 'synonyms-or-antonyms' questions, an improvement of 15% compared to Turney's results.

While all three approaches perform fairly well, they all have certain limitations. Mohammad et al. (2013)'s approach requires an external structured resource in form of a thesaurus. Both Lin et al. (2003) and Turney (2008)'s methods require antonyms to co-occur in fixed patterns, which may be less successful for lower-frequency antonyms. Incidentally, Lin et al. (2003)'s antonyms were chosen from a list of high-frequency terms to increase the chances of finding them in one of their patterns, while Turney (1998)'s data was drawn from websites for Learner English, and is therefore also likely to consist of higher-frequency words.[5] Our proposed model is not subject to such limitations: it does not require external structured resources or co-occurrences in fixed patterns.

4 Approach

Our hypotheses So far, there have been no successful attempts to distinguish synonymy and antonymy via standard distributional models such as the word space model (Sahlgren, 2006). This

[4] Yih et al. (2012) is another thesaurus-based approach.

[5] Mohammad et al. (2013) show that Lin et al. (2003)'s patterns have a low coverage for their antonym set.

is likely to be due to the assumed similarity of their contexts: Mohammad et al. (2013), for example, state that measures of distributional similarity typically fail to distinguish synonyms from semantically contrasting word pairs. They back up this claim with their own findings: Applying Lin (1998)'s similarity measure to a set of highly-contrasting antonyms, synonyms, and random pairs they show that both the high-contrast set and the synonyms set have a higher average distributional similarity than the random pairs. Interestingly, they also found that, on average, the set of opposites had a higher distributional similarity than the synonyms.

From an intuitive viewpoint such results are surprising: according to Charles and Miller (1991)'s substitutability experiments, there must be contextual clues that allow humans to distinguish between synonyms and antonyms. It appears, however, that these contextual differences are not captured by current measures of semantic similarity, leading to the assumption that synonyms and antonyms are distributionally similar and the claim that antonyms are counter-examples to the distributional hypothesis (cf. Section 2). The goal of our research is to show that this assumption is incorrect, and that contextual differences can be identified via standard distributional approaches using suitable features. In particular, we aim to provide support for the following hypotheses:

- **Hypothesis A.** The contexts of adjectival synonyms and antonyms are ***not*** distributionally similar.
- **Hypothesis B.** Not all word classes are useful for modelling the contextual differences between adjectival synonyms and antonyms.

We claim that the assumption that synonyms and antonyms are distributionally similar is incorrect. Their distributions may well be similar with respect to certain features (namely the ones commonly used in similarity measures), but our goal is to show that it is possible to identify distributional features that allow an automatic distinction between synonyms and antonyms (Hypothesis A). In particular, we expect synonyms to have a higher level of distributional similarity than antonyms (contrary to Mohammad et al. (2013)'s findings).

We further hypothesise that only some word classes are useful for modelling the contextual differences between adjectival synonyms and antonyms (Hypothesis B). For this purpose, we plan to investigate the influence of the following parts of speech in our distributional model: adjectives (ADJ), adverbs (ADV), verbs (VV), and nouns (NN). Our prediction is that the class of nouns will not be a useful indicator for distributional differences. This is motivated by Charles and Miller (1991)'s substitutability experiment, in which they claim that a noun phrase that can incorporate one adjective can also incorporate its antonym. As nouns relate to the semantic dimension denoted by the adjectives (cf. Section 2), they are in fact likely to co-occur with both the synonym (SYN) and the antonym (ANT) of a given target (T), resulting in high mutual information values for both, but not necessarily expressing the potential semantic differences between them:

- T: *unhappy* {**man**, **woman**, **child**, ...}
- SYN: *sad* {**man**, **woman**, **child**, ...}
- ANT: *happy* {**man**, **woman**, **child**, ...}

We expect to find more meaningful distributional differences in the *contexts* of such adjective-noun pairs, as illustrated in a simplified example:

- T: *unhappy* **man** – {*cry*, *moan*, *lament*, ...}
- SYN: *sad* **man** – {*cry*, *frown*, *moan*, ...}
- ANT: *happy* **man** – {*smile*, *laugh*, *sing*, ...}

We would for example assume that the set of verbs co-occurring with the target *unhappy* is more similar to the set of verbs co-occurring with its synonym *sad* than to the sets of verbs co-occurring with the antonym *happy*, resulting in higher similarity values for the pair of synonyms than for the pair of antonyms.

Addressing polysemy Addressing polysemy is an important task in distributional semantics, both with regard to type-based and token-based word senses: a distributional vector for a word type comprises features and associated feature strengths across all word senses, and a distributional vector for a word token does not indicate a sense if no disambiguation is performed. In recent years there have been a number of proposals that explicitly address the representation and identification of multiple senses in vector models, such as (Erk, 2009; Erk and Padó, 2010; Reisinger and Mooney, 2010; Boleda et al., 2012), with some focussing on identifying predominant word senses, such as (McCarthy et al., 2007; Mohammad and Hirst, 2006). In our experiments, we also aim to incorporate methods for dealing with multiple word senses.

In the task of synonym/antonym distinction, polysemy plays a central role as semantic relations tend to hold between specific *senses* of words rather than between *word forms* (cf. Mohammad et al. (2013)). For adjectives, polysemy directly relates to the semantic dimension they express. For example, depending on the dimension denoted by *hot* (cf. Section 2) we may expect different synonyms and antonyms. If we position *hot* on the dimension of TEMPERATURE, we might expect *scorching* as a synonym, and *cold* as an antonym. However, when *hot* is used to describe a person, we might instead use *attractive* as synonym, and *unattractive* as antonym. In their experiments on adjective synonym and antonym generation, Murphy and Andrew (1993) found that there was indeed considerable context sensitivity depending on the nouns that were modified by the target adjectives, with different synonyms and antonyms being generated.

Based on these insights we experiment with two different context settings: one that takes into account *all* contexts in which the target word and its synonym/antonym occur ('All-Contexts'), and one where we aim to resolve polysemy by applying the method of 'co-disambiguation' ('Codis-Contexts'). The co-disambiguation method attempts to exclude contexts of unrelated senses from consideration by establishing the set of nouns that are modified by both members of the synonym/antonym pair, and only including distributional information from contexts in which the adjectives co-occur (premodify) one of the set of shared nouns. This approach is motivated by the way in which humans might identify the semantic dimension of a pair of synonyms or antonyms out of context: using one member to disambiguate the other by figuring out which common property they express. For example, we intuitively realise that the synonyms *sweet* and *cute* are not related via the dimension of TASTE (as *sweet* might otherwise imply), but are used to describe a pleasing disposition. The co-disambiguation approach attempts to model this strategy by first identifying the nouns shared by the two adjectives across the corpus (such as *sweet/cute* {kid, dog, cottage, ...}, and then only collecting distributional information from such contexts. In the experiments described in the next sections we investigate if this smaller, but more focussed set of contexts can improve the results of our standard 'All-Contexts' model.

5 Experimental setup

This section provides an overview of the experimental setup and the distributional model we implemented to test our hypotheses. We work with German data in these experiments, but expect that the findings extend to other languages.

5.1 Training and test data

Our dataset is part of a collection of semantically related word pairs compiled via two separate experiments hosted on Amazon Mechanical Turk (AMT)[6]. The experiments were based on a set of 99 target adjectives which were selected from the lexical database GermaNet[7] using a stratified sampling technique accounting for 16 semantic categories, three polysemy classes, and three frequency classes. The first experiment asked AMT workers to propose synonyms, antonyms, and hypernyms for each of the targets. In the second experiment, workers were asked to rate the resulting pairs for the strength of antonymy, synonymy, and hypernymy between them, on a scale between 1 (minimum) and 6 (maximum). Both experiments resulted in 10 solutions per task.

To validate the generated synonym and antonym pairs, we carried out an assessment of their rating means (calculated over 10 ratings per word pair). The results show that there is a highly negative correlation between them with a Pearson r value of -0.895. This means that the higher a pair's rating as antonym, the lower its rating as synonym, and vice versa, which corresponds to our intuition that synonymy and antonymy are mutually exclusive relations. Figure 2 illustrates the relationship by plotting the average antonym and synonym ratings of all pairs in the dataset against each other.

For the current study we selected 97 synonym and 97 antonym pairs from this data as follows:

- The pairs have a rating means of ≥ 5, representing strong examples of the respective relation types. This narrowed the set of 99 adjective targets to 91 targets, participating in 116 antonym pairs and 145 synonym pairs.
- To decrease sparse data problems we excluded pairs where at least one of its members had a token frequency of < 20 in the sDeWaC-v3 corpus (Faaß et al., 2010), removing 6 antonym and 4 synonym pairs.

[6] https://www.mturk.com

[7] http://www.sfs.uni-tuebingen.de/lsd

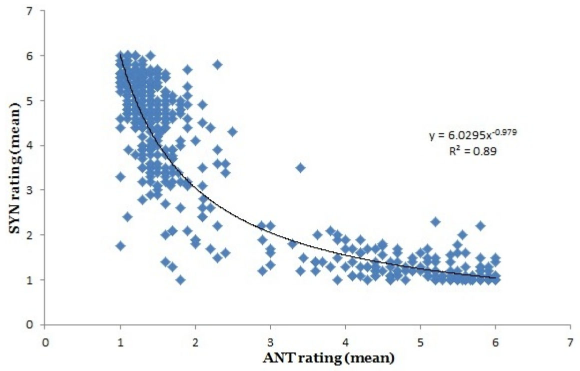

Figure 2: Scatter plot of rating means

- To allow a target-based assessment (cf. Section 5.2), our dataset was reduced to those targets which participate in at least one synonymy and one antonymy relation: 63 targets in total; examples are shown in Table 1. Note that the synonym and antonym pairs of a given target are not necessarily located on the same semantic dimension, as illustrated by the target *süß* (`sweet`).
- Based on these targets, we sampled an equal number of synonym and antonym pairs from the set, including at least one synonym and one antonym relation for each target, and giving preference to pairs with higher rating means. The resulting set includes 97 synonym and 97 antonym pairs altogether.

Target	**Synonym**	**Antonym**
fett ('fat')	*dick* ('thick')	*dünn* ('thin')
süß ('sweet')	*niedlich* ('cute')	*sauer* ('sour')
dunkel ('dark')	*düster* ('gloomy')	*hell* ('light')

Table 1: Dataset examples

5.2 Distributional model

Overview The main goal of this research is to show that there are distributional differences between synonym and antonym pairs that allow an automatic distinction between them (cf. Hypothesis A). The automatic method we use to address this task is an implementation of the word space model (Sahlgren, 2006; Turney and Pantel, 2010; Erk, 2012) where the members of the word pairs are represented as vectors in space, using contextual co-occurrence counts as vector dimension elements. The distributional similarity of two words is then calculated by means of the cosine function (a standard way of measuring vector similarity in word space models), which quantifies similarity by measuring the angle between two vectors v_T and v_{SYN} (or v_{ANT}) in vector space:

$$sim_{COS}(v_T, v_{SYN}) = \frac{v_T \cdot v_{SYN}}{|v_T| \cdot |v_{SYN}|}$$

Following from the discussion in Section 4, we expect higher cosine similarity values for synonyms, and lower values for antonyms. We establish the effectiveness of our proposed model for synonym/antonym distinction by means of an automatic classifier on the set of relation pairs introduced in Section 5.1.

Co-occurrence information The co-occurrence information included in the model is drawn from the sDeWaC-v3 corpus (Faaß et al., 2010), a cleaned version of the German web corpus deWaC[8], which contains around 880 million tokens and has been parsed with Bohnet's MATE dependency parser (Bohnet, 2010). The corpus further provides lemma and part-of-speech annotations (STTS tagset). We varied the window sizes we took into account as co-occurrence information; here we report our findings for the best window size of 5 tokens to the left and right of the adjectives (but not crossing sentence boundaries).

Instead of simple co-occurrence frequencies, our model uses *local mutual information (LMI)* scores as vector values. LMI is a measure from information theory that compares observed frequencies O with expected frequencies E, taking marginal frequencies into account: $LMI = O \times log \frac{O}{E}$, with E representing the product of the marginal frequencies over the sample size.[9] In comparison to (pointwise) mutual information (Church and Hanks, 1990), LMI improves the well-known problem of propagating low-frequent events through multiplying mutual information by the observed frequency.

Experimental settings To address our hypothesis that only some word classes are useful for modelling the contextual differences between adjectival synonyms and antonyms (Hypothesis B), we build separate word spaces for the following collocate types: adjectives (ADJ), adverbs (ADV), verbs (VV), and nouns (NN). In addition, we also consider a combination of all four word classes (COMB). For this purpose, we compiled co-occurrence vectors for each word class by counting the frequencies of all adjective–collocate tuples that appeared in the sdeWaC corpus within

[8]`http://wacky.sslmit.unibo.it/doku.php?id=corpora`

[9]See `http://www.collocations.de/AM/` for a detailed illustration of association measures (incl. LMI).

the specified window (here, size 5). For example, the model 'VV in window w5' includes all verbs that appear in a context window of five words from the adjectives, such as [*süß* – *verspeisen* – 3 – 12.4448] ('sweet' – 'devour' – frequency – LMI).

As discussed in Section 4, we consider two context settings: one that collects co-occurrence information from *all* contexts of the adjectives ('All-Contexts'), and one that applies co-disambiguation to address polysemy ('Codis-Contexts'). For the latter, word vectors only include co-occurrence information from contexts in which the members of a synonym/antonym pair modify a shared noun.

Classifier To establish whether there are significant distributional differences between synonyms and antonyms, and to assess the discriminative power of the different word class models, we experimented with several WEKA[10] classifiers and measures (e.g. Jaccard) and assessed their performance at synonym/antonym distinction using 10-fold cross-validation. Here we describe the results of the best-performing combination of classifier and measure: a Decision Tree classifier ('J48') with one single feature (*standard-cosine*, or *cosine-difference* values). Thus, for each of the experimental settings described above we run the classifier twice. In the first scenario, we use the plain cosine values (i.e. the distributional similarity values of the synonym/antonym pairs) as features in the classification. This default scenario is somewhat unrealistic, as it assumes a specific cosine cut-off value that distinguishes synonyms from antonyms. The second scenario addresses this issue and refers to a target-based point of view: It may be the case that for the majority of targets, the cosine values of their synonyms are significantly higher than those of their antonyms, indicating clear distributional differences. However, such information is lost when training the classifier on *all* cosine values in cases where the cosine value of the antonym of a target T_1 is greater than the synonym value of another target T_2, as illustrated in Figure 3, making it difficult to find an appropriate cut-off value to split the data in classes. We take this into consideration as follows: for each synonym and antonym pair involving target T (cf. Section 5.1), we calculate the difference between their cosine values and use these difference values as input to the classifier. For example, the cosine values for the synonym pair *süß* - *niedlich* and antonym pair *süß* - *sauer* (cf. Table 1) are 0.94 (T:SYN) and 0.18 (T:ANT), respectively, and the difference value is calculated as (T:SYN - T:ANT). The resulting value (which may be positive or negative) is used as input for the synonym pair (here, 0.76), while the negated value is used as input for the antonym pair (-0.76). For cases where several synonym or antonym pairs are available, an average difference value is calculated.

[10] www.cs.waikato.ac.nz/ml/weka

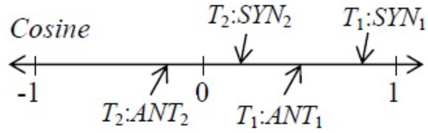

Figure 3: Relative cosine values

6 Results

This section presents the results of the Decision Tree classification of synonyms vs. antonyms, using *standard-cosine* values as features (Figure 4) and using *cosine-difference* values (Figure 5). The graphs show the performance of the classifiers in % accuracy for the five part-of-speech-based word space models (ADJ, ADV, VV, NN, and COMB), while at the same time comparing the performances of the two context settings 'Codis-Contexts' (dark bars) and 'All-Contexts' (light bars). The results are compared against a 50% baseline (dotted line), and significant improvements are marked with a star.

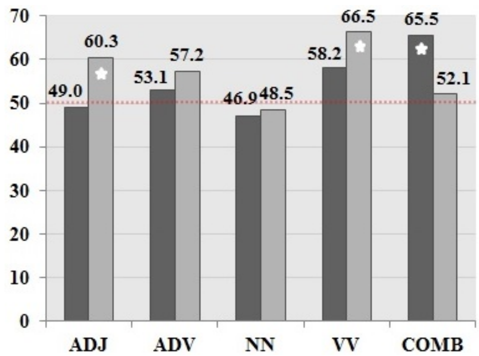

Figure 4: Classification results (*standard-cosine*)

7 Discussion

Hypothesis A The graphs in Figures 4 and 5 clearly show that it is possible to automatically distinguish between synonymy and antonymy by means of a word space model, with significant improvements over the 50% baseline. These results support our hypothesis that synonyms and antonyms are *not* distributionally similar, and refute the claim that antonyms constitute a counter-example to the distributional hypothesis. An in-

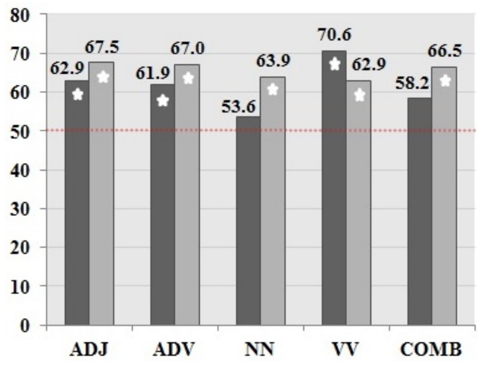

Figure 5: Classification results (*cosine-difference*)

vestigation of the decision trees underlying the best-performing classifiers in Figure 4 further shows surprisingly clearly that there *is* a cut-off point over the cosine values that separates synonyms from antonyms, with antonyms in the lower-value and synonyms in the higher-value partition. For example, the cut-off value for the 'All-Contexts' model for verbs (light bar in Figure 4) is 0.1186, and any instances with lower cosine values are labelled as antonyms, and with higher values as synonyms, achieving 66.5% accuracy. This is in line with our prediction that synonyms are more distributionally similar than antonyms.

Hypothesis B Our second hypothesis, that not all word classes are useful for modelling the contextual differences between adjectival synonyms and antonyms, is also supported by the findings: the word space models built on the class of collocate verbs (VV) appear to be the best discriminators of the relations overall, outperforming the baseline in all four scenarios shown in Figures 4 and 5. All except one of these improvements are statistically significant.[11] The second-best class according to our statistical analysis is the class of adjectives (ADJ), which outperforms the baseline in three of four scenarios (all three being statistically significant). The class of adverbs occupies middle ground, significantly outperforming the baseline only in the *cosine-difference* scenario. As predicted, the noun class (NN) fares worst in the experiments, only (significantly) beating the baseline in one scenario (*cosine-difference*, 'All-Contexts').

Polysemy The graphs in Figures 4 and 5 show that in most experiment conditions the 'All-Contexts' setting (which incorporates co-occurrence information from all contexts) achieves better results than the 'Codis-Contexts' setting (which aims to address polysemy by means of 'co-disambiguation'). However, in the *cosine-difference* scenario, which aims to provide a more accurate representation of distributional differences, the 'Codis-Contexts' setting provides a much clearer picture of the differences between the word classes than the 'All-Contexts' setting (with accuracy values ranging from 53.6% for nouns to 70.6% for verbs for the former, and 62.9% for verbs to 67.5% for adjectives for the latter). Furthermore, the overall best result (i.e. relying on verbs in the *cosine-difference* scenario) is achieved in the 'Codis-Contexts' setting.

A closer analysis of the vector sizes shows that the performance of the 'co-disambiguation' approach might be affected by sparse data. Given a larger source of co-occurrence data, the approach may achieve better results than shown in Figures 4 and 5. Overall, our findings suggest that the 'co-disambiguation' approach to dealing with polysemy represents a worthwhile avenue for future research, especially on consideration of its other advantages such as ease of implementation and reduced space requirements.

8 Conclusion

Our experiments demonstrated that synonyms and antonyms can be distinguished by means of a distributional word space model, refuting the general assumption that synonyms and antonyms are distributionally similar. With 66.5% and 70.6% accuracy in two different classification settings, our model achieves significant improvements over a 50% baseline, and compares favourably to previous approaches by Turney (2008), who achieved an improvement of 9.6% over his baseline, and Lin et al. (2003), whose method is assumed to only work for high-frequency antonyms.

What are the implications of our findings for distributional semantics? First of all, we have shown that the distributional hypothesis holds true even for antonyms. Secondly, our finding that not all word classes are equally useful for modelling the contextual differences between synonyms and antonyms suggests that the performance of distributional measures may be improved by excluding certain word classes from consideration, depending on the task. Finally, we introduced a simple 'co-disambiguation' approach to dealing with polysemy in distributional word space models.

[11] *standard-cosine*, 'All-Contexts': χ^2 = 10.85, $p < .001$; *cosine-difference*, 'All-Contexts': χ^2 = 6.55, $p < .05$; *cosine-difference*, 'Codis-Contexts': χ^2 = 8.18, $p < .005$.

References

Bernd Bohnet. 2010. Top Accuracy and Fast Dependency Parsing is not a Contradiction. In *Proceedings of COLING*, Beijing, China.

Gemma Boleda, Sabine Schulte im Walde, and Toni Badia. 2012. Modelling Regular Polysemy: A Study on the Semantic Classification of Catalan Adjectives. *Computational Linguistics*, 38(3):575–616.

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*, 32(1):13–47.

Walter Charles and George Miller. 1989. Contexts of Antonymous Adjectives. *Applied Psycholinguistics*, 10:357–375.

Walter Charles and George Miller. 1991. Contextual Correlates of Semantic Similarity. *Language and Cognitive Processes*, 6(1):1–28.

Alan Cruse. 1986. *Lexical Semantics*. CUP, Cambridge, UK.

Marie-Catherine de Marneffe, Anna N. Rafferty, and Christopher D. Manning. 2008. Finding Contradictions in Text. In *Proceedings of ACL-HLT*, pages 1039–1047, Columbus, OH.

James Deese. 1965. *The Structure of Associations in Language and Thought*. The John Hopkins Press, Baltimore, MD.

Philip Edmonds. 1997. Choosing the Word most typical in Context using a Lexical Co-occurrence Network. In *Proceedings of ACL*, pages 507–509, Madrid, Spain.

Katrin Erk and Sebastian Padó. 2010. Exemplar-based Models for Word Meaning in Context. In *Proceedings of ACL*, Uppsala, Sweden.

Katrin Erk. 2009. Representing Words in Regions in Vector Space. In *Proceedings of CoNLL*, pages 57–65, Boulder, Colorado.

Katrin Erk. 2012. Vector Space Models of Word Meaning and Phrase Meaning: A Survey. *Language and Linguistics Compass*, 6(10):635–653.

Gertrud Faaß, Ulrich Heid, and Helmut Schmid. 2010. Design and Application of a Gold Standard for Morphological Analysis: SMOR in Validation. In *Proceedings of LREC*, pages 803–810, Valletta, Malta.

Zellig Harris. 1968. Distributional Structure. In *The Philosophy of Linguistics*, pages 26–47. OUP.

Donald Hindle. 1990. Noun Classification from Predicate-Argument Structures. In *Proceedings of ACL*, pages 268–275.

John S. Justeson and Slava M. Katz. 1991. Co-Occurrence of Antonymous Adjectives and their Contexts. *Computational Linguistics*, 17:1–19.

Adrienne Lehrer and Keith Lehrer. 1982. Antonymy. *Linguistics and Philosophy*, 5:483–501.

Dekang Lin, Shaojun Zhao, Lijuan Qin, and Ming Zhou. 2003. Identifying Synonyms among Distributionally Similar Words. In *Proceedings of the IJCAI*, pages 1492–1493, Acapulco, Mexico.

Dekang Lin. 1998. Automatic Retrieval and Clustering of Similar Words. In *Proceedings of COLING*, Montreal, Canada.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2007. Unsupervised Acquisition of Predominant Word Senses. *Computational Linguistics*, 33(4):553–590.

Saif Mohammad and Graeme Hirst. 2006. Determining Word Sense Dominance Using a Thesaurus. In *Proceedings of EACL*, Trento, Italy.

Saif M. Mohammad, Bonnie J. Dorr, Graeme Hirst, and Peter D. Turney. 2013. Computing Lexical Contrast. *Computational Linguistics*, 39(3).

Gregory L. Murphy and Jane M. Andrew. 1993. The Conceptual Basis of Antonymy and Synonymy in Adjectives. *Memory and Language*, 32(3):1–19.

M. Lynne Murphy. 2003. *Semantic Relations and the Lexicon*. Cambridge University Press.

Carita Paradis, Caroline Willners, and Steven Jones. 2009. Good and Bad Opposites: Using Textual and Experimental Techniques to Measure Antonym Canonicity. *The Mental Lexicon*, 4(3):380–429.

Joseph Reisinger and Raymond J. Mooney. 2010. Multi-Prototype Vector-Space Models of Word Meaning. In *Proceedings NAACL*, pages 109–117.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual Correlates of Synonymy. *Communications of the ACM*, 8(10):627–633.

Magnus Sahlgren. 2006. *The Word-Space Model*. Ph.D. thesis, Stockholm University.

Peter D. Turney and Patrick Pantel. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *Artificial Intelligence Research*, 37:141–188.

Peter D. Turney. 2008. A Uniform Approach to Analogies, Synonyms, Antonyms, and Associations. In *Proceedings COLING*, pages 905–912, Manchester, UK.

Caroline Willners. 2001. Antonyms in Context. In *Travaux de Institut de Linguistique de Lund 40*, Lund, Sweden.

Wen-Tau Yih, Geoffrey Zweig, and John C. Platt. 2012. Polarity Inducing Latent Semantic Analysis. In *Proceedings of the EMNLP and CoNLL*, pages 1212–1222, Jeju Island, Korea.