Learning to identify antonyms

Natalia Silveira
CS 229 Final Project, Fall 2013

1. Introduction

Antonymy is a common lexical relation that is intuitively clear (if not always easy to define) for humans, but challenging for machines. In Natural Language Processing, antonymy detection has applications in tasks of understanding language, such as paraphrase detection, question answering, and textual inference. For that reason, WordNet (Fellbaum, 2005) includes some antonymy annotation; but the relation is relatively rare, and a quick manual inspection reveals that there are many more antonym pairs (including very common ones) than those shown in WordNet. It also becomes clear that the relation is somewhat vague; *masculine* and *neuter*, for example, are listed as antonyms in WordNet, but many native speakers of English would not intuitively consider these words antonyms.

The presence of these antonyms is WordNet makes an interesting resource for supervised learning, however, which opens the possibility of trying to automatically extend the annotation. Identifying pairs of antonyms in corpora is the task I propose in this paper. The approach is to train a classifier to distinguish between sets of sentences that contain pairs of antonyms from sets of sentences that do not. The intuition behind it is that antonyms are often used contrastively in the same sentence. The highest-performing classifier obtains 84% accuracy. Because the literature on this task is limited, it is hard to rigorously compare the performance of the classifier with existing published results; however, it does seem to in line with, and perhaps above, results reported from other research.

An important aspect of this approach is that knowledge-rich feature engineering was deliberately avoided. The reason for this is that an approach to detecting antonymy is much more productive if it fits into a general framework for learning other lexical relations, such as synonymy, hypernymy, entailment etc.. Therefore, whereas linguistic knowledge about the antonymy relation can be useful for the task (for example, features indicating the presence of morphemes such as *un-* or *dis-* would clearly be informative), I instead opted for a distributional approach, where the features are

2. Literature Review

There has been some previous work in learning lexical relations in general, and antonymy in particular. Mohammad et al. (2008) present the task of detecting antonymy degree, and bring in insights from the study of antonyms in corpora. Although their work is of limited methodological interest for me (because the goal of the authors was different than mine), the theoretical insights inform my approach. They point out that antonym pairs are formed in part by collocation; speakers think of words as antonyms not by reasoning exclusively about the semantics, but also by observing that they occur in similar contexts, and are used contrastively. Furthermore, antonyms are "similar, but different;" antonyms are not words with meanings that are as different as possible, but words with meanings that are very similar, but different in some respect, such as referring to opposite ends of the same scale for measuring the same property. Mohammad et al.

(2008) also present evidence about the behavior of antonyms in corpora that will be important for my approach. Antonym adjectives occur in the same sentence more often than expected by chance. In fact, Mohammad and Hirst (2006) show that they tend to co-occur in a five-word window. This seems to happen because antonyms are often used contrastively (for example, in a phrase like: *not just not cold, but quite hot*). They also occur in similar syntactic contexts; that is, the syntactic structures that allow the occurrence of one word will also allow for its antonym. The main insight here is that cooccurrence data can be informative

Distributional methods, based on bag-of-words vectors, have been criticized for not being able to distinguish between different types of similarity. Word vectors of antonyms will look similar, because the meanings of antonyms are closely related, and they can usually modify the same types of nouns, and be modified by the same types of adverbs. Therefore, a naive approach might not be able to distinguish pairs of synonyms from pairs of antonyms. However, being able to use distributional information in refined ways might be a path for a general framework for detecting lexical relations.

Turney (2008) presents a unified framework for detecting lexical relations with a distributional approach, by introducing features that refer to the syntactic and lexical patterns that connect words when they occur together, rather than simply looking at the contexts of each of the word individually. An SVM classifier was then trained on these features. This seems particularly promising for antonymy, because, as mentioned above, it is clear from corpus research that antonyms occur together in contrastive patterns. In a task of classifying pairs of words as antonyms or synonyms, Turney's approach had 75% accuracy.

Baroni et al. (2012) presents an interesting distributional take on detecting lexical entailment, another relation. This means identifying that, for instance, *all dogs* entails *some dogs*, or that *bright student* entails *student*. They create a number of pairs of entailing and non-entailing phrases, and train a polynomial-kernel SVM to work on the concatenation of the word vectors of each phrase in the pair, for a large training corpus. This method obtained 70% accuracy. The evaluation is less rigorous than that of Turney (2008), because Baroni et al. they report results only for distinguishing antonyms from random phrase pairs, but not from, for example, synonyms; still, this is an interesting result for an innovative task, and it raises a question of how far distributional methods can be taken in the discovery of fine lexical relations.

3. Data

Several different parts of speech can enter antonymy relations, and WordNet has the annotation for nouns, verbs, adverbs and adjectives. In this paper, I focus on adjectives. The reason is for this choice is that the annotations in WordNet are more numerous and higher-quality for adjectives; also intuitions about adjective antonymy seem crisper to me, which is relevant because the approach relies on antonym pairs being seen as such by speakers, which would be reflected in their use of antonyms in the corpus. There are 3048 adjectives in WordNet that are annotated for antonymy.

The text comes from the LDC Annotated Gigaword (LDC release LDC2012T21). This choice of corpus is essentially motivated by the size of this resource, and the availability of an annotated version. I worked with the New York Times section, which contains thousands of stories published in the New York Times from July 1994 to

December 2010. The total number of tokens in this section of the corpus is approximately 900m words.

My data consists of a pair-to-word matrix, indicating the frequency with which word *w* cooccurs in a sentence with *both* adjectives in the pair of adjectives (*x*, *y*). Essentially, the vector for the pair (*x*, *y*)

The pairs were formed only adjectives occurring at least 500 times total (if this threshold seems high, note that the corpus has almost 1 billion words). Since many of the adjectives that are annotated with the antonymy relation on Wordnet are somewhat rare, this threshold was meant to guarantee that we have enough information about the data points to meaningfully evaluate the classifier's performance. A total of 653 pairs of antonyms were found with this methodology. To create the negative examples, I randomly drew 653 pairs of synonym adjectives (each occurring at least 500 times) from WordNet, and then created another 653 by randomly pairing up the adjectives already harvested for the synonym and antonym pairs, and ensuring that no new pairs of synonyms or antonyms were created. The pairs were split 80-20 for training and testing.

Note that there is no word sense disambiguation in the corpus, so there is some noise in the data. For example, *hot* is the antonym of *cold* only when it refers to temperature sense, not when it refers to spiciness. But in this methodology, sentences such as *The curry wasn't good, it was too hot for my taste, and cold by the time it reached the table* would contribute to the vector for the pair (*hot*, *cold*).

4. Experiments and Discussion

The type of classifier chosen for this task was SVM, which has often been shown high performance in NLP tasks.

Preprocessing. I experimented with different transformations commonly used in the NLP literature on bag-of-word vectors: pointwise mutual information, length normalization, and TFIDF (term frequency-inverse document frequency). These transformations were not helpful in preliminary experiments with a linear kernel; my intuition is that in this case the length of the documents, which is roughly the number of sentence in which the adjectives cooccur, is particularly important, because For this reason, transformations designed to "factor out" document length are not appropriate. The only transformation on the design matrix that improved results was standardizing the feature values to have unit feature-variance. This is widely reported to be helpful for classification with SVMs.

I also experimented with chi-square-based feature selection. It is often reported for text classification tasks that using only a subset of features does not improve performance. In this case, intuitively it seems like feature selection may be more helpful; in classification with richer labels (based on genre or topic), it seems that having a richer representation of the text is intuitively important; but since in the current task the key features are simply signals that a contrast is being expressed, feature selection seemed like a good idea. The results for subsets of features of various sizes were slightly worse in preliminary experiments with the training set; since the literature points in the same direction, I opted to use the full set of features. This being a very large corpus, there were around 200,000 features total.

Choice of kernel. I chose a linear kernel for the SVM, based on performance on a held-out set. For this, I ran a very coarse grid search to optimize an RBF kernel and a

degree-2 polynomial kernel. Both performed consistently worse than the linear kernel: for all parameter settings I tried, the results were below the lowest results obtained with the linear kernel. Additionally, the accuracy of the best linear kernel classifier on the training set was 94%, another signal that it was appropriate for this data. For these reasons, I chose the linear kernel for further experiments.

Parameter estimation. With this choice of kernel, I performed a grid search on the parameters C and γ, optimizing for cross-validation accuracy on the training set within the values [1, 10, 100, 1000] for C and [0.001, 0.0001, 0.00001] for γ. The best results were obtained with C=1 and γ=0.001.

Results. The results obtained are described in the table below. As a reminder, the test set contains 130 pairs of antonyms, 130 pairs of synonyms, and 130 random pairs.

	Antonyms	Non-antonyms	Total
Precision	0.78	0.72	0.84
Recall	0.87	0.90	0.84
F-Score	0.75	0.88	0.84

The total accuracy is 84%; these are better results than I was able to find in the literature.

5. Conclusion and Future Work

The results obtained here are encouraging. Although it is hard to make an exact comparison, because of distinctions in how different authors define the task, these results are clearly not worse than previously achieved results for learning antonymy (and other lexical relations) with distributional methods; they may in fact be better, as we see almost 40% error reduction with relation to the accuracy reported in Turney (2008) for identifying antonym pairs. Excluding the insight that antonyms tend to occur in the same sentence, no knowledge about how antonymy behaves specifically was used, which opens an avenue for extending the approach to other relations.

A clear next step in this work would be to incorporate syntactic features that might help characterize the constructions in which the pairs of adjectives occur together.