Computing Word-Pair Antonymy

Saif Mohammad[†] **Bonnie Dorr**[*] **Graeme Hirst**[ϕ]

[†*]Laboratory for Computational Linguistics and Information Processing
[†*]Institute for Advanced Computer Studies and [*]Computer Science
[†*]University of Maryland and [*]Human Language Technology Center of Excellence
{saif,bonnie}@umiacs.umd.edu

[ϕ]Department of Computer Science
University of Toronto
gh@cs.toronto.edu

Abstract

Knowing the degree of antonymy between words has widespread applications in natural language processing. Manually-created lexicons have limited coverage and do not include most semantically contrasting word pairs. We present a new automatic and empirical measure of antonymy that combines corpus statistics with the structure of a published thesaurus. The approach is evaluated on a set of closest-opposite questions, obtaining a precision of over 80%. Along the way, we discuss what humans consider antonymous and how antonymy manifests itself in utterances.

1 Introduction

Native speakers of a language intuitively recognize different **degrees of antonymy**—whether two words are strongly antonymous (*hot–cold, good–bad, friend–enemy*), just semantically contrasting (*enemy–fan, cold–lukewarm, ascend–slip*) or not antonymous at all (*penguin–clown, cold–chilly, boat–rudder*). Over the years, many definitions of antonymy have been proposed by linguists (Cruse, 1986; Lehrer and Lehrer, 1982), cognitive scientists (Kagan, 1984), psycholinguists (Deese, 1965), and lexicographers (Egan, 1984), which differ from each other in small and large respects. In its strictest sense, antonymy applies to gradable adjectives, such as *hot–cold* and *tall–short*, where the two words represent the two ends of a semantic dimension. In a broader sense, it includes other adjectives, nouns, and verbs as well (*life–death, ascend–descend, shout–whisper*). In its broadest sense, it applies to any two words that represent contrasting meanings. We will use the term **degree of antonymy** to encompass the complete semantic range—a combined measure of the contrast in meaning conveyed by two words and the tendency of native speakers to call them opposites. The higher the degree of antonymy between a target word pair, the greater the semantic contrast between them and the greater their tendency to be considered antonym pairs by native speakers.

Automatically determining the degree of antonymy between words has many uses including detecting and generating paraphrases (*The dementors **caught** Sirius Black* / *Black could **not escape** the dementors*) and detecting contradictions (Marneffe et al., 2008; Voorhees, 2008) (*Kyoto has a predominantly **wet** climate* / *It is mostly **dry** in Kyoto*). Of course, such "contradictions" may be a result of differing sentiment, new information, non-coreferent mentions, or genuinely contradictory statements. Antonyms often indicate the discourse relation of contrast (Marcu and Echihabi, 2002). They are also useful for detecting humor (Mihalcea and Strapparava, 2005), as satire and jokes tend to have contradictions and oxymorons. Lastly, it is useful to know which words are semantically contrasting to a target word, even if simply to filter them out. For example, in the automatic creation of a thesaurus it is necessary to distinguish near-synonyms from word pairs that are semantically contrasting. Measures of distributional similarity fail to do so. Detecting antonymous words is not sufficient to solve most of these problems, but it remains a crucial, and largely unsolved, component.

Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 982–991,
Honolulu, October 2008.

Lexicons of pairs of words that native speakers consider antonyms have been created for certain languages, but their coverage has been limited. Further, as each term of an antonymous pair can have many semantically close terms, the contrasting word pairs far outnumber those that are commonly considered antonym pairs, and they remain unrecorded. Even though a number of computational approaches have been proposed for semantic closeness, and some for hypernymy–hyponymy (Hearst, 1992), measures of antonymy have been less successful. To some extent, this is because antonymy is not as well understood as other classical lexical-semantic relations.

We first very briefly summarize insights and intuitions about this phenomenon, as proposed by linguists and lexicographers (Section 2). We discuss related work (Section 3). We describe the resources we use (Section 4) and present experiments that examine the manifestation of antonymy in text (Sections 5 and 6). We then propose a new empirical approach to determine the degree of antonymy between two words (Section 7). We compiled a dataset of 950 closest-opposite questions, which we used for evaluation (Section 8). We conclude with a discussion of the merits and limitations of this approach and outline future work.

2 The paradoxes of antonymy

Antonymy, like synonymy and hyponymy, is a lexical-semantic relation that, strictly speaking, applies to two **lexical units**—combinations of surface form and word sense. (That said, for simplicity and where appropriate we will use the term "antonymous words" as a proxy for "antonymous lexical units".) However, accepting this leads to two interesting and seemingly paradoxical questions (described below in the two subsections).

2.1 Why are some pairs better antonyms?

Native speakers of a language consider certain contrasting word pairs to be antonymous (for example, *large–small*), and certain other seemingly equivalent word pairs as less so (for example, *large–little*). A number of reasons have been suggested: (1) Cruse (1986) observes that if the meaning of the target words is completely defined by one semantic dimension and the words represent the two ends of this semantic dimension, then they tend to be considered antonyms. We will refer to this semantic dimension as the **dimension of opposition**. (2) If on the other hand, as Lehrer and Lehrer (1982) point out, there is more to the meaning of the antonymous words than the dimension of opposition—for example, more semantic dimensions or added connotations—then the two words are not so strongly antonymous. Most people do not think of *chubby* as a direct antonym of *thin* because it has the additional connotation of being cute and informal. (3) Cruse (1986) also postulates that word pairs are not considered strictly antonymous if it is difficult to identify the dimension of opposition (for example, *city–farm*). (4) Charles and Miller (1989) claim that two contrasting words are identified as antonyms if they occur together in a sentence more often than chance. However, Murphy and Andrew (1993) claim that the greater-than-chance co-occurrence of antonyms in sentences is because together they convey contrast well, which is rhetorically useful, and not really the reason why they are considered antonyms in the first place.

2.2 Are semantic closeness and antonymy opposites?

Two words (more precisely, two lexical units) are considered to be close in meaning if there is a lexical-semantic relation between them. Lexical-semantic relations are of two kinds: **classical** and **non-classical**. Examples of classical relations include synonymy, hyponymy, troponymy, and meronymy. Non-classical relations, as pointed out by Morris and Hirst (2004), are much more common and include concepts pertaining to another concept (*kind, chivalrous, formal* pertaining to *gentlemanly*), and commonly co-occurring words (for example, problem–solution pairs such as *homeless, shelter*). Semantic distance (or closeness) in this broad sense is known as semantic relatedness. Two words are considered to be **semantically similar** if they are associated via the synonymy, hyponymy–hypernymy, or the troponymy relation. So terms that are semantically similar (*plane–glider, doctor–surgeon*) are also semantically related, but terms that are semantically related may not always be semantically similar (*plane–sky, surgeon–scalpel*).

Antonymy is unique among these relations because it simultaneously conveys both a sense of

closeness and of distance (Cruse, 1986). Antonymous concepts are semantically related but not semantically similar.

3 Related work

Charles and Miller (1989) proposed that antonyms occur together in a sentence more often than chance. This is known as the **co-occurrence hypothesis**. They also showed that this was empirically true for four adjective antonym pairs. Justeson and Katz (1991) demonstrated the co-occurrence hypothesis for 35 prototypical antonym pairs (from an original set of 39 antonym pairs compiled by Deese (1965)) and also for an additional 22 frequent antonym pairs. All of these pairs were adjectives. Fellbaum (1995) conducted similar experiments on 47 noun, verb, adjective, and adverb pairs (noun–noun, noun–verb, noun–adjective, verb–adverb and so on) pertaining to 18 concepts (for example, *lose(v)–gain(n)* and *loss(n)–gain(n)*, where *lose(v)* and *loss(n)* pertain to the concept of "failing to have/maintain"). However, non-antonymous semantically related words such as hypernyms, holonyms, meronyms, and near-synonyms also tend to occur together more often than chance. Thus, separating antonyms from them has proven to be difficult.

Lin et al. (2003) used patterns such as "from *X* to *Y*" and "either *X* or *Y*" to separate antonym word pairs from distributionally similar pairs. They evaluated their method on 80 pairs of antonyms and 80 pairs of synonyms taken from the *Webster's Collegiate Thesaurus* (Kay, 1988). In this paper, we propose a method to determine the degree of antonymy between any word pair and not just those that are distributionally similar. Turney (2008) proposed a uniform method to solve word analogy problems that require identifying synonyms, antonyms, hypernyms, and other lexical-semantic relations between word pairs. However, the Turney method is supervised whereas the method proposed in this paper is completely unsupervised.

Harabagiu et al. (2006) detected antonyms for the purpose of identifying contradictions by using WordNet chains—synsets connected by the hypernymy–hyponymy links and exactly one antonymy link. Lucerto et al. (2002) proposed detecting antonym pairs using the number of words between two words in text and also cue words such as *but*, *from*, and *and*. Unfortunately, they evaluated their method on only 18 word pairs. Neither of these methods determines the degree of antonymy between words and they have not been shown to have substantial coverage. Schwab et al. (2002) create "antonymous vector" for a target word. The closer this vector is to the context vectors of the other target word, the more antonymous the two target words are. However, the antonymous vectors are manually created. Further, the approach is not evaluated beyond a handful of word pairs.

Work in sentiment detection and opinion mining aims at determining the polarity of words. For example, Pang, Lee and Vaithyanathan (2002) detect that adjectives such as *dazzling*, *brilliant*, and *gripping* cast their qualifying nouns positively whereas adjectives such as *bad*, *cliched*, and *boring* portray the noun negatively. Many of these gradable adjectives have antonyms. but these approaches do not attempt to determine pairs of positive and negative polarity words that are antonyms.

4 Resources

4.1 Published thesauri

Published thesauri, such as the *Roget's* and *Macquarie*, divide the vocabulary into about a thousand **categories**. Words within a category tend to be near-synonymous or semantically similar. One may also find antonymous and semantically related words in the same category, but this is rare. The intuition is that words within a category represent a coarse concept. Words with more than one meaning may be found in more than one category; these represent its coarse senses. Within a category, the words are grouped into **paragraphs**. Words in the same paragraph tend to be closer in meaning than those in different paragraphs. We will take advantage of the structure of the thesaurus in our approach.

4.2 WordNet

Unlike the traditional approach to antonymy, WordNet encodes antonymy as a lexical relationship—a relation between two words (not concepts) (Gross et al., 1989). Even though a synset (a WordNet concept) may be represented by more than one word, individual words across synsets are marked as (di-

rect) antonyms. Gross et al. argue that other words in the synsets form "indirect antonyms".

Even after including the indirect antonyms, WordNet's coverage is limited. As Marcu and Echihabi (2002) point out, WordNet does not encode antonymy across part-of-speech (for example, *legally–embargo*). Further, the noun–noun, verb–verb, and adjective–adjective antonym pairs of WordNet largely ignore near-opposites as revealed by our experiments (Section 8 below). Also, WordNet (or any other manually-created repository of antonyms for that matter) does not encode the *degree* of antonymy between words. Nevertheless, we investigate the usefulness of WordNet as a source of seed antonym pairs for our approach.

4.3 Co-occurrence statistics

The **distributional hypothesis of closeness** states that words that occur in similar contexts tend to be semantically close (Firth, 1957). Distributional measures of distance, such as those proposed by Lin (1998), quantify how similar the two sets of contexts of a target word pair are. Equation 1 is a modified form of Lin's measure that ignores syntactic dependencies and hence it estimates semantic relatedness rather than semantic similarity:

$$Lin(w_1, w_2) = \frac{\sum_{w \in T(w_1) \cap T(w_2)} (I(w_1, w) + I(w_2, w))}{\sum_{w' \in T(w_1)} I(w_1, w') + \sum_{w'' \in T(w_2)} I(w_2, w'')} \quad (1)$$

Here w_1 and w_2 are the target words; $I(x,y)$ is the pointwise mutual information between x and y; and $T(x)$ is the set of all words y that have positive pointwise mutual information with the word x ($I(x,y) > 0$).

Mohammad and Hirst (2006) showed that these distributional word-distance measures perform poorly when compared with WordNet-based concept-distance measures. They argued that this is because the word-distance measures clump together the contexts of the different senses of the target words. They proposed a way to obtain distributional distance between word *senses*, using any of the distributional measures such as *cosine* or that proposed by Lin, and showed that this approach performed markedly better than the traditional *word-distance* approach. They used thesaurus categories as very coarse word senses. Equation 2 shows how Lin's formula is used to determine distributional distance between two thesaurus categories c_1 and c_2:

$$Lin(c_1, c_2) = \frac{\sum_{w \in T(c_1) \cap T(c_2)} (I(c_1, w) + I(c_2, w))}{\sum_{w' \in T(c_1)} I(c_1, w') + \sum_{w'' \in T(c_2)} I(c_2, w'')} \quad (2)$$

Here $T(c)$ is the set of all words w that have positive pointwise mutual information with the thesaurus category c ($I(c,w) > 0$). We adopt this method for use in our approach to determine word-pair antonymy.

5 The co-occurrence hypothesis of antonyms

As a first step towards formulating our approach, we investigated the co-occurrence hypothesis on a significantly larger set of antonym pairs than those studied before. We randomly selected a thousand antonym pairs (nouns, verbs, and adjectives) from WordNet and counted the number of times (1) they occurred individually and (2) they co-occurred in the same sentence within a window of five words, in the *British National Corpus (BNC)* (Burnard, 2000). We then calculated the mutual information for each of these word pairs and averaged it. We randomly generated another set of a thousand word pairs, without regard to whether they were antonymous or not, and used it as a control set. The average mutual information between the words in the antonym set was 0.94 with a standard deviation of 2.27. The average mutual information between the words in the control set was 0.01 with a standard deviation of 0.37. Thus antonymous word pairs occur together much more often than chance irrespective of their intended senses ($p < 0.01$). Of course, a number of non-antonymous words also tend to co-occur more often than chance—commonly known as collocations. Thus, strong co-occurrence is not a sufficient condition for detecting antonyms, but these results show that it can be a useful cue.

6 The substitutional and distributional hypotheses of antonyms

Charles and Miller (1989) also proposed that in most contexts, antonyms may be interchanged. The

meaning of the utterance will be inverted, of course, but the sentence will remain grammatical and linguistically plausible. This came to be known as the **substitutability hypothesis**. However, their experiments did not support this claim. They found that given a sentence with the target adjective removed, most people did not confound the missing word with its antonym. Justeson and Katz (1991) later showed that in sentences that contain both members of an antonymous adjective pair, the target adjectives do indeed occur in similar syntactic structures at the phrasal level. From this (and to some extent from the co-occurrence hypothesis), we can derive the **distributional hypothesis of antonyms**: antonyms occur in similar contexts more often than non-antonymous words.

We used the same set of one thousand antonym pairs and one thousand control pairs as in the previous experiment to gather empirical proof of the distributional hypothesis. For each word pair from the antonym set, we calculated the distributional distance between each of their senses using Mohammad and Hirst's (2006) method of concept distance along with the modified form of Lin's (1998) distributional measure (equation 2). The distance between the closest senses of the word pairs was averaged for all thousand antonyms. The process was then repeated for the control set.

The control set had an average semantic closeness of 0.23 with a standard deviation of 0.11 on a scale from 0 (unrelated) to 1 (identical). On the other hand, antonymous word pairs had an average semantic closeness of 0.30 with a standard deviation of 0.23.[1] This demonstrates that relative to other word pairs, antonymous words tend to occur in similar contexts ($p < 0.01$). However, near-synonymous and similar word pairs also occur in similar contexts. (the distributional hypothesis of closeness). Thus, just like the co-occurrence hypothesis, occurrence in similar contexts is not sufficient, but rather yet another useful cue towards detecting antonyms.

[1]It should be noted that absolute values in the range between 0 and 1 are meaningless by themselves. However, if a set of word pairs is shown to consistently have higher values than another set, then we can conclude that the members of the former set tend to be semantically closer than those of the latter.

7 Our approach

We now present an empirical approach to determine the degree of antonymy between words. In order to maximize applicability and usefulness in natural language applications, we model the broad sense of antonymy. Given a target word pair, the approach determines whether they are antonymous or not, and if they are antonymous whether they have a high, medium, or low degree of antonymy. More precisely, the approach presents a way to determine whether one word pair is more antonymous than another.

The approach relies on the structure of the published thesaurus as well as the co-occurrence and distributional hypotheses. As mentioned earlier, a thesaurus organizes words in sets representing concepts or categories. We first determine pairs of thesaurus categories that are contrasting in meaning (Section 7.1). We then use the co-occurrence and distributional hypotheses to determine the degree of antonymy (Section 7.2).

7.1 Detecting contrasting categories

We propose two ways of detecting thesaurus category pairs that represent contrasting concepts (we will call these pairs **contrasting categories**): (1) using a seed set of antonyms and (2) using a simple heuristic that exploits how thesaurus categories are ordered.

7.1.1 Seed sets

Affix-generated seed set Antonym pairs such as *hot–cold* and *dark–light* occur frequently in text, but in terms of type-pairs they are outnumbered by those created using affixes, such as *un-* (*clear–unclear*) and *dis-* (*honest–dishonest*). Further, this phenomenon is observed in most languages (Lyons, 1977).

Table 1 lists sixteen morphological rules that tend to generate antonyms in English. These rules were applied to each of the words in the *Macquarie Thesaurus* and if the resulting term was also a valid word in the thesaurus, then the word-pair was added to the **affix-generated seed set**. These sixteen rules generated 2,734 word pairs. Of course, not all of them are antonymous, for example *sect–insect* and *coy–decoy*. However, these are relatively few in

w_1	w_2	example pair	w_1	w_2	example pair	w_1	w_2	example pair
X	*ab*X	*normal–abnormal*	X	*mis*X	*fortune–misfortune*	*im*X	*ex*X	*implicit–explicit*
X	*anti*X	*clockwise–anticlockwise*	X	*non*X	*aligned–nonaligned*	*in*X	*ex*X	*introvert–extrovert*
X	*dis*X	*interest–disinterest*	X	*un*X	*biased–unbiased*	*up*X	*down*X	*uphill–downhill*
X	*im*X	*possible–impossible*	*l*X	*ill*X	*legal–illegal*	*over*X	*under*X	*overdone–underdone*
X	*in*X	*consistent–inconsistent*	*r*X	*ir*X	*regular–irregular*	X*less*	X*ful*	*harmless–harmful*
X	*mal*X	*adroit–maladroit*						

Table 1: Sixteen affix rules to generate antonym pairs. Here 'X' stands for any sequence of letters common to both words w_1 and w_2.

number and were found to have only a small impact on the results.

WordNet seed set We compiled a list of 20,611 semantically contrasting word pairs from WordNet. If two words from two synsets in WordNet are connected by an antonymy link, then every possible word pair across the two synsets was considered to be semantically contrasting. A large number of them include multiword expressions. For only 10,807 of the 20,611 pairs were both words found in the *Macquarie Thesaurus*—the vocabulary used for our experiments. We will refer to them as the **WordNet seed set**.

Then, given these two seed sets, if any word in thesaurus category C_1 is antonymous to any word in category C_2 as per a seed antonym pair, then the two categories are marked as contrasting. It should be noted, however, that the seed antonym pair may be antonymous only in certain senses. For example, consider the antonym pair *work–play*. Here, *play* is antonymous to *work* only in its ACTIVITY FOR FUN sense and not its DRAMA sense. In such cases, we employ the distributional hypothesis of closeness: two words are antonymous to each other in those senses which are closest in meaning to each other. Since the thesaurus category pertaining to WORK is relatively closer in meaning to the ACTIVITY FOR FUN sense than the DRAMA sense, those two categories will be considered contrasting and not the categories pertaining to WORK and DRAMA.

If no word in C_1 is antonymous to any word in C_2, then the categories are considered not contrasting. As the seed sets, both automatically generated and manually created, are relatively large in comparison to the total number of categories in the *Macquarie Thesaurus* (812), this simple approach has reasonable coverage and accuracy.

7.1.2 Order of thesaurus categories

Most published thesauri are ordered such that contrasting categories tend to be adjacent. This is not a hard-and-fast rule, and often a category may be contrasting in meaning to several other categories. Further, often adjacent categories are not semantically contrasting. However, since this was an easy-enough heuristic to implement, we investigated the usefulness of considering adjacent categories as contrasting. We will refer to this as the **adjacency heuristic**.

7.2 Determining the degree of antonymy

Once we know which category pairs are contrasting (using the methods from the previous subsection), we determine the *degree* of antonymy between the two categories (Section 7.2.1). The aim is to assign contrasting category pairs a non-zero value signifying the degree of contrast. In turn, we will use that information to determine the degree of antonymy between any word pair whose members belong to two contrasting categories (Sections 7.2.2 and 7.2.3).

7.2.1 Category level

Using the distributional hypothesis of antonyms, we claim that the degree of antonymy between two *contrasting* concepts (thesaurus categories) is directly proportional to the distributional closeness of the two concepts. In other words, the more the words representing two contrasting concepts occur in similar contexts, the more the two concepts are considered to be antonymous.

Again we used Mohammad and Hirst's (2006) method along with Lin's (1998) distributional measure to determine the distributional closeness of two thesaurus concepts. Co-occurrence statistics required for the approach were computed from the

BNC. Words that occurred within a window of 5 words were considered to co-occur.

7.2.2 Lexical unit level

Recall that strictly speaking, antonymy (like other lexical-semantic relations) applies to lexical units (a combination of surface form and word sense). If two words are used in senses pertaining to contrasting categories (as per the methods described in Section 7.1), then we will consider them to be antonymous (degree of antonymy is greater than zero). If two words are used in senses pertaining to non-contrasting senses, then we will consider them to be not antonymous (degree of antonymy is equal to 0).

If the target words belong to the same thesaurus paragraphs as any of the seed antonyms linking the two contrasting categories, then the words are considered to have a high degree of antonymy. This is because words that occur in the same thesaurus paragraph tend to be semantically very close in meaning. Relying on the co-occurrence hypothesis, we claim that for word pairs listed in contrasting categories, the greater their tendency to co-occur in text, the higher their degree of antonymy. We use mutual information to capture the tendency of word–word co-occurrence.

If the target words do not both belong to the same paragraphs as a seed antonym pair, but occur in contrasting categories, then the target words are considered to have a low or medium degree of antonymy (less antonymous than the word pairs discussed above). Such word pairs that have a higher tendency to co-occur are considered to have a medium degree of antonymy, whereas those that have a lower tendency to co-occur are considered to have a low degree of antonymy.

Co-occurrence statistics for this purpose were collected from the *Google n-gram corpus* (Brants and Franz, 2006).[2] Words that occurred within a window of 5 words were considered to be co-occurring.

7.2.3 Word level

Even though antonymy applies to pairs of word and sense combinations, most available texts are not sense-annotated. If antonymous occurrences are to be exploited for any of the purposes listed in the beginning of this paper, then the text must be sense disambiguated. However, word sense disambiguation is a hard problem. Yet, and to some extent because unsupervised word sense disambiguation systems perform poorly, much can be gained by using simple heuristics. For example, it has been shown that cohesive text tends to have words that are close in meaning rather than unrelated words. This, along with the distributional hypothesis of antonyms, and the findings by Justeson and Katz (1991) (antonymous concepts tend to occur more often than chance in the same sentence), suggests that if we find a word pair in a sentence such that two of its senses are strongly contrasting (as per the algorithm described in Section 7.2.2), then it is probable that the two words are used in those contrasting senses.

8 Evaluation

8.1 Task and data

In order to best evaluate a computational measure of antonymy, we need a task that not only requires knowing whether two words are antonymous but also whether one word pair is more antonymous than another pair. Therefore, we evaluated our system on a set of closest-opposite questions. Each question has one target word and five alternatives. The objective is to identify that alternative which is the closest opposite of the target. For example, consider:

adulterate: a. *renounce* b. *forbid*
c. *purify* d. *criticize* e. *correct*

Here the target word is *adulterate*. One of the alternatives provided is *correct*, which as a verb has a meaning that contrasts with that of *adulterate*; however, *purify* has a greater degree of antonymy with *adulterate* than *correct* does and must be chosen in order for the instance to be marked as correctly answered. This evaluation is similar to how others have evaluated semantic distance algorithms on TOEFL synonym questions (Turney, 2001), except that in those cases the system had to choose the alternative which is *closest* in meaning to the target.

We looked on the World Wide Web for large sets of closest antonym questions. We found two independent sets of questions designed to prepare stu-

[2] We used the *Google n-gram corpus* is created from a text collection of over 1 trillion words. We intend to use the same corpus (and not the *BNC*) to determine semantic distance as well, in the near future.

	development data			test data		
	P	R	F	P	R	F
a. random baseline	0.20	0.20	0.20	0.20	0.20	0.20
b. affix-generated seeds only	0.72	0.53	0.61	0.71	0.51	0.60
c. WordNet seeds only	0.79	0.52	0.63	0.75	0.50	0.60
d. both seed sets	0.77	0.65	0.70	0.73	0.60	0.65
e. adjacency heuristic only	**0.81**	0.43	0.56	**0.83**	0.46	0.59
f. affix seed set + heuristic	0.75	0.60	0.67	0.76	0.61	0.68
g. both seed sets + heuristic	0.76	**0.66**	**0.70**	0.76	**0.64**	**0.70**

Table 2: Results obtained on closest-opposite questions.

dents for the Graduate Record Examination.[3] The first set consists of 162 questions. We used this set to develop our approach and will refer to it as the *development set*. Even though the algorithm does not have any tuned parameters per se, the development set helped determine which cues of antonymy were useful and which were not. The second set has 1208 closest-opposite questions. We discarded questions that had a multiword target or alternative. After removing duplicates we were left with 950 questions, which we used as the unseen *test set*.

Interestingly, the data contains many instances that have the same target word used in different senses. For example:

(1) ***obdurate:*** a. *meager* b. *unsusceptible* c. *right* d. *tender* e. *intelligent*
(2) ***obdurate:*** a. *yielding* b. *motivated* c. *moribund* d. *azure* e. *hard*
(3) ***obdurate:*** a. *transitory* b. *commensurate* c. *complaisant* d. *similar* e. *uncommunicative*

In (1), *obdurate* is used in the HARDENED IN FEELINGS sense and the closest opposite is *tender*. In (2), it is used in the RESISTANT TO PERSUASION sense and the closest opposite is *yielding*. In (3), it is used in the PERSISTENT sense and the closest opposite is *transitory*.

The datasets also contain questions in which one or more of the alternatives is a near-synonym of the target word. For example:

astute: a. *shrewd* b. *foolish* c. *callow* d. *winning* e. *debating*

Observe that *shrewd* is a near-synonym of *astute*. The closest-opposite of *astute* is *foolish*. A manual check of a randomly selected set of 100 test-set questions revealed that, on overage, one in four had a near-synonym as one of the alternative.

[3]Both datasets are apparently in the public domain and will be made available on request.

8.2 Experiments

We used the algorithm proposed in Section 7 to automatically solve the closest-opposite questions. Since individual words may have more than one meaning, we relied on the hypothesis that the intended sense of the alternatives are those which are most antonymous to one of the senses of the target word. (This follows from the discussion earlier in Section 7.2.3.) So for each of the alternatives we used the target word as context (but not the other alternatives). We think that using a larger context to determine antonymy will be especially useful when the target words are found in sentences and natural text—something we intend to explore in the future.

Table 2 presents results obtained on the development and test data using different combinations of the seed sets and the adjacency heuristic. If the system did not find any evidence of antonymy between the target and any of its alternatives, then it refrained from attempting that question. We therefore report precision (number of questions answered correctly / number of questions attempted), recall (number of questions answered correctly / total number of questions), and F-score values ($2 \times P \times R/(P+R)$).

Observe that all results are well above the random baseline of 0.20 (obtained when a system randomly guesses one of the five alternatives to be the answer). Also, using only the small set of sixteen affix rules, the system performs almost as well as when it uses 10,807 WordNet antonym pairs. Using both the affix-generated and the WordNet seed sets, the system obtains markedly improved precision and coverage. Using only the adjacency heuristic gave best precision values (upwards of 0.8) with substan-

tial coverage (attempting close to half the questions). However, best overall performance was obtained using both seed sets and the adjacency heuristic (F-score of 0.7).

8.3 Discussion

These results show that, to some degree, the automatic approach does indeed mimic human intuitions of antonymy. In tasks that require higher precision, using only the adjacency heuristic is best, whereas in tasks that require both precision and coverage, the seed sets may be included. Even when both seed sets were included, only four instances in the development set and twenty in the test set had target–answer pairs that matched a seed antonym pair. For all remaining instances, the approach had to generalize to determine the closest opposite. This also shows that even the seemingly large number of direct and indirect antonyms from WordNet (more than 10,000) are by themselves insufficient.

The comparable performance obtained using the affix rules alone suggests that even in languages without a wordnet, substantial accuracies may be achieved. Of course, improved results when using WordNet antonyms as well suggests that the information they provide is complementary.

Error analysis revealed that at times the system failed to identify that a category pertaining to the target word contrasted with a category pertaining to the answer. Additional methods to identify seed antonym pairs will help in such cases. Certain other errors occurred because one or more alternatives other than the official answer were also antonymous to the target. For example, the system chose *accept* as the opposite of *chasten* instead of *reward*.

9 Conclusion

We have proposed an empirical approach to antonymy that combines corpus co-occurrence statistics with the structure of a published thesaurus. The method can determine the degree of antonymy or contrast between any two thesaurus categories (sets of words representing a coarse concept) and between any two word pairs. We evaluated the approach on a large set of closest-opposite questions wherein the system not only identified whether two words are antonymous but also distinguished between pairs of antonymous words of different degrees. It achieved an F-score of 0.7 in this task where the random baseline was only 0.2. When aiming for high precision it scores over 0.8, but there is some drop in the number of questions attempted. In the process of developing this approach we validated the co-occurrence hypothesis proposed by Charles and Miller (1989) on a large set of 1000 noun, verb, and adjective pairs. We also gave empirical proof that antonym pairs tend to be used in similar contexts—the distributional hypothesis for antonyms.

Our future goals include porting this approach to a cross-lingual framework in order to determine antonymy in a resource-poor language by combining its text with a thesaurus from a resource-rich language. We will use antonym pairs to identify contrast relations between sentences to in turn improve automatic summarization. We also intend to use the approach proposed here in tasks where keyword matching is especially problematic, for example, separating paraphrases from contradictions.

Acknowledgments

We thank Smaranda Muresan, Siddharth Patwardhan, members of the CLIP lab at the University of Maryland, College Park, and the anonymous reviewers for their valuable feedback. This work was supported, in part, by the National Science Foundation under Grant No. IIS-0705832, in part, by the Human Language Technology Center of Excellence, and in part, by the Natural Sciences and Engineering Research Council of Canada. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsor.

References

Thorsten Brants and Alex Franz. 2006. Web 1t 5-gram version 1. *Linguistic Data Consortium*.

Lou Burnard. 2000. *Reference Guide for the British National Corpus (World Edition)*. Oxford University Computing Services.

Walter G. Charles and George A. Miller. 1989. Contexts of antonymous adjectives. *Applied Psychology*, 10:357–375.

David A. Cruse. 1986. *Lexical semantics*. Cambridge University Press.

James Deese. 1965. *The structure of associations in language and thought*. The Johns Hopkins Press.

Rose F. Egan. 1984. Survey of the history of English synonymy. *Webster's New Dictionary of Synonyms*, pages 5a–25a.

Christiane Fellbaum. 1995. Co-occurrence and antonymy. *International Journal of Lexicography*, 8:281–303.

John R. Firth. 1957. A synopsis of linguistic theory 1930–55. In *Studies in Linguistic Analysis*, pages 1–32, Oxford: The Philological Society. (Reprinted in F.R. Palmer (ed.), Selected Papers of J.R. Firth 1952-1959, Longman).

Derek Gross, Ute Fischer, and George A. Miller. 1989. Antonymy and the representation of adjectival meanings. *Memory and Language*, 28(1):92–106.

Sanda M. Harabagiu, Andrew Hickl, and Finley Lacatusu. 2006. Lacatusu: Negation, contrast and contradiction in text processing. In *Proceedings of the 23rd National Conference on Artificial Intelligence (AAAI-06)*, Boston, MA.

Marti Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*, pages 539–546, Nantes, France.

John S. Justeson and Slava M. Katz. 1991. Co-occurrences of antonymous adjectives and their contexts. *Computational Linguistics*, 17:1–19.

Jerome Kagan. 1984. *The Nature of the Child*. Basic Books.

Maire Weir Kay, editor. 1988. *Webster's Collegiate Thesaurus*. Merrian-Webster.

Adrienne Lehrer and K. 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 18th International Joint Conference on Artificial Intelligence (IJCAI-03)*, pages 1492–1493, Acapulco, Mexico.

Dekang Lin. 1998. Automatic retreival and clustering of similar words. In *Proceedings of the 17th International Conference on Computational Linguistics (COLING-98)*, pages 768–773, Montreal, Canada.

Cupertino Lucerto, David Pinto, and Héctor Jimiénez-Salazar. 2002. An automatic method to identify antonymy. In *Workshop on Lexical Resources and the Web for Word Sense Disambiguation*, pages 105–111, Puebla, Mexico.

John Lyons. 1977. *Semantics*, volume 1. Cambridge University Press.

Daniel Marcu and Abdesammad Echihabi. 2002. An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*, Philadelphia, PA.

Marie-Catherine de Marneffe, Anna Rafferty, and Christopher D. Manning. 2008. Finding contradictions in text. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL-08)*, Columbus, OH.

Rada Mihalcea and Carlo Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 531–538, Vancouver, Canada.

Saif Mohammad and Graeme Hirst. 2006. Distributional measures of concept-distance: A task-oriented evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia.

Jane Morris and Graeme Hirst. 2004. Non-classical lexical semantic relations. In *Proceedings of the Workshop on Computational Lexical Semantics, HLT*, Boston, MA.

Gregory L. Murphy and Jane M. Andrew. 1993. The conceptual basis of antonymy and synonymy in adjectives. *Journal of Memory and Language*, 32(3):1–19.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 79–86, Philadelphia, PA.

Didier Schwab, Mathieu Lafourcade, and Violaine Prince. 2002. Antonymy and conceptual vectors. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING-02)*, pages 904–910.

Peter Turney. 2001. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the Twelfth European Conference on Machine Learning*, pages 491–502, Freiburg, Germany.

Peter Turney. 2008. A uniform approach to analogies, synonyms, antonyms, and associations. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING-08)*, pages 905–912, Manchester, UK.

Ellen M Voorhees. 2008. Contradictions and justifications: Extensions to the textual entailment task. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL-08)*, Columbus, OH.