When Similarity Becomes Opposition: Synonyms and Antonyms Discrimination in DSMs

Enrico Santus[*]
Hong Kong Polytechnic University

Alessandro Lenci[§]
Università di Pisa

Qin Lu[†]
Hong Kong Polytechnic University

Chu-Ren Huang[‡]
Hong Kong Polytechnic University

This paper analyzes the concept of opposition and describes a fully unsupervised method for its automatic discrimination from near-synonymy in Distributional Semantic Models (DSMs). The discriminating method is based on the hypothesis that, even though both near-synonyms and opposites are mostly distributionally similar, opposites are different from each other in at least one dimension of meaning, which can be assumed to be salient. Such hypothesis has been implemented in APAnt*, a distributional measure that evaluates the extent of the intersection among the most relevant contexts of two words (where relevance is measured as mutual dependency), and its saliency (i.e. their average rank in the mutual dependency sorted list of contexts). The measure – previously introduced in some pilot studies – is presented here with two variants. Evaluation shows that it outperforms three baselines in an antonym retrieval task: the vector cosine, a baseline implementing the co-occurrence hypothesis, and a random rank. This paper describes the algorithm in details and analyzes its current limitations, suggesting that extensions may be developed for discriminating antonyms not only from near-synonyms but also from other semantic relations. During the evaluation, we have noticed that* APAnt *also has a particular preference for hypernyms.*

1. Introduction

Similarity is one of the fundamental principles organizing the semantic lexicon (Lenci, 2008; Landauer and Dumais, 1997). Distributional Semantic Models (DSMs) encoding the frequency of co-occurrences between words in large corpora are proven to be successful in representing word meanings in terms of distributional similarity (Turney and Pantel, 2010; Padó and Lapata, 2007; Sahlgren, 2006).

These models allow a geometric representation of the *Distributional Hypothesis* (Harris, 1954), that is, words occurring in similar contexts also have similar meanings. They represent words as vectors in a continuous vector space, where distributional similarity can be measured as vector proximity. This, in turn, can be calculated through the *vector cosine* (Turney and Pantel, 2010). This representation is so effective that DSMs are known to be able to replicate human judgments with a reasonable accuracy (Lenci, 2008).

[*] E-mail: e.santus@connect.polyu.hk
[§] E-mail: alessandro.lenci@ling.unipi.it
[†] E-mail: qin.lu@polyu.edu.hk
[‡] E-mail: churen.huang@polyu.edu.hk

However, the *Distributional Hypothesis* shapes the concept of similarity in a very loose way, including among the distributionally similar words not only those that refer to similar referents (e.g. co-hyponyms and near-synonyms), but – more in general – all those words that share many contexts (Harris, 1954). As a consequence of such definition, words like *dog* may be considered similar not only to the co-hyponym lexeme *cat*, but also to the hypernym *animal*, the meronym *tail* (Morlane-Hondère, 2015), and so on. This loose definition, therefore, poses a big challenge in Natural Language Processing (NLP), and in particular in Computational Lexical Semantics, where the meaning of a word and the type of relations it holds with others need to be univocally identified. For instance, in a task such as *Textual Entailment* it is crucial not only to identify whether two words are semantically similar, but also whether they entail each other, like hyponym-hypernym pairs do. Similarly, in *Sentiment Analysis* the correct discrimination of antonyms (e.g. *good* from *bad*) is extremely important to identify the positive or negative polarity of a text.

Among the relations that fall under the large umbrella of distributional similarity, there is indeed opposition, also known as *antonymy*. According to Cruse (1986), antonymy is characterized by the *paradox of simultaneous similarity and difference*: Opposites are identical in every dimension of meaning except for one. A typical example of such paradox is the relation between *dwarf* and *giant*. These words are semantically similar in many aspects (i.e. they may refer to similar entities, such as humans, trees, galaxies), differing only for what concerns the size, which is assumed to be a salient semantic dimension for them. Distributionally speaking, *dwarfs* and *giants* share many contexts (e.g., both *giant* and *dwarf* may be used to refer to *galaxies, stars, planets, companies, people*[1]), differing for those related to the semantic dimension of size. For example, *giant* is likely to occur in contexts related to big sizes, such as *global, corporate, dominate* and so on[2], while *dwarf* is likely to occur in contexts related to small sizes, such as *virus, elf, shrub* and so on[3].

Starting from this observation, we describe and analyze a method aiming to identify opposites in DSMs. The method, which is directly inspired to Cruse's paradox, is named *APAnt* (from *Average Precision for Antonyms*) and lies on the hypothesis that antonyms share less salient contexts than synonyms. The method was first presented in two previous pilot studies of Santus et al. (2014b, 2014c). In those papers, *APant* was shown to outperform the *vector cosine* and a baseline implementing the *co-occurrence hypothesis* (Charles and Miller, 1989) in an *antonym retrieval* task (AR), using a standard window-based DSM, built by collecting the co-occurrences between the two content words on the left and the right of the target word, in a combination of ukWaC and WaCkypedia (Santus et al., 2014a)[4]. The task was performed using the Lenci/Benotto dataset (Santus et al., 2014b) and evaluated through *Average Precision* (AP; Kotlerman et al., 2010).

In this paper, we first give a more detailed description of *APAnt* presenting also two variants. All the measures are evaluated in two *antonym retrieval* tasks, performed on an extended dataset, which includes antonyms, synonyms, hypernyms and co-hyponyms (henceforth, also referred as coordinates, according to Baroni and Lenci, 2011) from the Lenci/Benotto (Santus et al., 2014b), *BLESS* (Baroni and Lenci, 2011) and *EVALution 1.0* (Santus et al., 2015). Again, *APAnt*

[1] These examples were found by searching in *Sketch Engine* (https://www.sketchengine.co.uk), using the *word sketch* function.

[2] Ibid.

[3] Ibid.

[4] Similar experiments on a standard five content words window DSM have confirmed that *APAnt* outperforms the *vector cosine* and the *co-occurrence* baseline. The actual impact of the window size still needs to be properly analyzed.

outperforms the above-mentioned baselines plus another one based on random ranking.

The paper is organized as follows. In the next section, we define opposites and their properties (Section 2), moving then to the state of the art for their discrimination (Section 3). We introduce our method and its variations (Section 4) and describe their evaluation (Section 5). A detailed discussion of the results (Sections 6 and 7) and the conclusions are reported at the end of the paper (Conclusions).

2. Opposites

People do not always perfectly agree on classifying word pairs as opposites (Mohammad et al., 2013), confirming that their identification is indeed a hard task, even for native speakers. The major problems in such task are that (1) opposites are rarely in a truly binary contrast (e.g. *warm/hot*); (2) the contrast can be of different kinds (e.g. semantic, as in *hot/cold*, or referential, as in *shark/dolphin*); and (3) opposition is often context-dependent (e.g. consider the near-synonyms *very good* and *excellent* in the following sentence: "not simply *very good*, but *excellent*"; Cruse, 1986; Murphy, 2003). All these issues make opposites difficult to define, so that linguists often need to rely on diagnostic tests to make the opposition clear (Murphy, 2003).

Over the years, many scholars from different disciplines have tried to provide a precise definition of this semantic relation. They are yet to reach any conclusive agreement. Kempson (1977) defines opposites as word pairs with a "binary incompatible relation", such that the presence of one meaning entails the absence of the other. In this sense, *giant* and *dwarf* are good opposites, while *giant* and *person* are not. Mohammad et al. (2013), noticing that the terms *opposites, contrasting* and *antonyms* have often been used interchangeably, have proposed the following distinction: (1) *opposites* are word pairs that are strongly incompatible with each other and/or are saliently different across a dimension of meaning; (2) *contrasting word pairs* have some non-zero degree of binary incompatibility and/or some non-zero difference across a dimension of meaning; (3) *antonyms* are opposites that are also gradable adjectives. They have also provided a simple but comprehensive classification of opposites based on Cruse (1986), including (1) *antipodals* (e.g. *top-bottom*), pairs whose terms are at the opposite extremes of a specific meaning dimension; (2) *complementaries* (e.g. *open-shut*), pairs whose terms divide the domain in two mutual exclusive compartments; (3) *disjoints* (e.g. *hot-cold*), pairs whose words occupy non-overlapping regions in a specific semantic dimension, generally representing a state; (4) *gradable opposites* (e.g. *long-short*), adjective- or adverb-pairs that gradually describe some semantic dimensions, such as length, speed, etc.; (5) *reversibles* (e.g. *rise-fall*), verb-pairs whose words respectively describe the change from state A to state B and the inverse, from state B to state A.

In this paper, we will not account for all these differences, but we will use the terms *opposites* and *antonyms* as synonyms, meaning all pairs of words in which a certain level of contrast is perceived. Under such category we include also the *paranyms*, which are a specific type of coordinates (Huang et al., 2007) that partition a conceptual field into complementary subfields. For instance, although *dry season*, *spring*, *summer*, *autumn* and *winter* are all co-hyponyms, only the latter four are paranyms, as they split the conceptual field of *seasons*.

3. Related Works

Opposites identification is very challenging for computational models (Mohammad et al., 2008; Deese, 1965; Deese, 1964). Yet, this relation is essential for many NLP applications, such as *Information Retrieval* (IR), *Ontology Learning* (OL), *Machine Translation* (MT), *Sentiment Analysis* (SA) and *Dialogue Systems* (Roth and Schulte im Walde, 2014; Mohammad et al., 2013). In particular, the automatic identification of semantic opposition is crucial for the detection and generation of paraphrases (i.e. during the generation, similar but contrasting candidates should be filtered out, as described in Marton et al., 2011), the understanding of contradictions (de Marneffe et al., 2008) and the identification of irony (Xu et al., 2015; Tungthamthiti et al., 2015) and humor (Mihalcea and Strapparava, 2005).

Several existing hand-crafted computational lexicons and thesauri explicitly encoding opposition are often used to support the above mentioned NLP tasks, even though many scholars have shown their limits. Mohammad et al. (2013), for example, point out that "more than 90% of the contrasting pairs in GRE closest-to-opposite questions[5] are not listed as opposites in WordNet". Moreover, the relations encoded in such resources are mostly context independent.

Given the already mentioned reliability of Distributional Semantic Models (DSMs) in the detection of distributional similarity between lexemes, several studies have tried to exploit these models for the identification of semantic relations (Santus et al., 2014a; Baroni and Lenci, 2010; Turney and Pantel, 2010; Padó and Lapata, 2007; Sahlgren, 2006). As mentioned before, however, DSMs are characterized by a major shortcoming. That is, they are not able to discriminate among different kinds of semantic relations linking distributionally similar lexemes (Santus et al., 2014a). This is the reason why supervised and pattern-based approaches have often been preferred (Pantel and Pennacchiotti, 2006; Hearst, 1992). However, these latter methods have also various problems, most notably the difficulty of finding patterns that are highly reliable and univocally associated with specific relations, without incurring at the same time in data-sparsity problems. The experience of pattern-based approaches has shown that these two criteria can rarely be satisfied simultaneously.

The foundation of most corpus-based research on opposition is the *co-occurrence hypothesis* (Lobanova, 2012), formulated by Charles and Miller (1989) after observing that opposites co-occur in the same sentence more often than expected by chance. Such claim has then found many empirical confirmations (Justeson and Katz, 1991; Fellbaum, 1995) and it is used in the present work as a baseline. Ding and Huang (2014; 2013) also pointed out that, unlike co-hyponyms, opposites generally have a strongly preferred word order when they co-occur in a coordinate context (i.e. A and/or B). Another part of related research has been focused on the study of lexical-syntactic constructions that can work as linguistic tests for opposition definition and classification (Cruse, 1986).

Starting from all these observations, several computational methods for opposition identification were implemented. Most of them rely on patterns (Schulte im Walde and Köper, 2013; Lobanova et al., 2010; Turney, 2008; Pantel and Pennacchiotti, 2006; Lin et al., 2003), which unfortunately suffer from low recall, because they can be applied only to frequent words. Others, like Lucerto et al. (2002), use the number of tokens between the target words and other clues (e.g. the presence/absence of conjunctions like but, from, and, etc.) to identify contrasting words.

[5] GRE stands for *Graduate Record Examination*, which is a standardized test, often used as an admissions requirement for graduate schools in the United States.

Turney (2008) proposed a supervised algorithm for the identification of several semantic relations, including synonyms and opposites. The algorithm relied on a training set of word pairs with class labels to assign the labels also to a testing set of word pairs. All word pairs were represented as vectors encoding the frequencies of co-occurrence in textual patterns extracted from a large corpus of web pages. He used the sequential minimal optimization (SMO) support vector machine (SVM) with a radial basis function (RBF) kernel (Platt, 1998) implemented in Weka (Waikato Environment for Knowledge Analysis) (Witten and Frank, 1999). In the discrimination between synonyms and opposites, the system achieved an accuracy of 75% against a majority class baseline of 65.4%.

Mohammad et al. (2008) proposed a method for determining the degree of semantic contrast (i.e. how much two contrasting words are semantically close) based on the use of thesauri categories and corpus statistics. For each target word pair, they used the co-occurrence and the distributional hypothesis to establish the degree of opposition. Their algorithm achieved an F-score of 0.7, against a random baseline of 0.2.

Mohammad et al. (2013) used an analogical method based on a given set of contrasting words to identify and classify different kinds of opposites by hypothesizing that for every opposing pair of words, A and B, there is at least another opposing pair, C and D, such that A is similar to C and B is similar to D. For example, for the pair *night-day*, there is the pair *darkness-daylight*, such that *night* is similar to *darkness* and *day* to *daylight*. Given the existence of contrast, they calculated its degree relying on the *co-occurrence* hypothesis. Their approach outperformed other state-of-the-art measures.

Schulte im Walde and Köper (2013) proposed a vector space model relying on lexico-syntactic patterns to distinguish between synonymy, antonymy and hypernymy. Their approach was tested on German nouns, verbs and adjectives, achieving a precision of 59.80%, which was above the majority baseline.

More recently, Roth and Schulte im Walde (2014) proposed that statistics over discourse relations can be used as indicators for paradigmatic relations, including opposition.

4. Our Method: *APAnt*

Starting from the already mentioned *paradox of simultaneous similarity and difference between antonyms* (Cruse, 1986), in Santus et al. (2014b, 2014c) we have proposed a distributional measure that modifies the *Average Precision* formula (Kotlerman et al., 2010) to discriminate antonyms from near-synonyms. *APAnt*, from *Average Precision for Antonymy*, takes into account two main factors: i) the extent of the intersection among the *N* most relevant contexts of two words (where relevance is measured as mutual dependency); and ii) the salience of such intersection (i.e. the average rank in the mutual dependency sorted list of contexts). These factors are considered under the hypothesis that near-synonyms are likely to share a larger part of the salient contexts compared to antonyms.

In this section, we describe in details the *APAnt* algorithm, proposing also two variants aimed to improve *APAnt* stability and extend its scope. They will be named with an increasing number, *APAnt2* (which consists in a simple normalization of *APAnt*) and *APAnt3* (which introduces a new factor to *APAnt2*, that is, the distributional similarity among the word pairs).

APAnt should be seen as the inverse of *APSyn* (*Average Precision for Synonymy*). While *APSyn* assigns higher scores to near-synonyms, *APAnt* assigns higher scores to antonyms. Such scores can then be used for semantic relations discrimination tasks. Given a target pair w_1 *and* w_2, *APSyn* first selects the *N* most relevant contexts for each of the two terms. *N* should be large enough to sufficiently describe the distributional semantics of a term for a given purpose. Relevance is calculated in terms of Local Mutual Information (LMI; Evert, 2005), which is a measure that describes the mutual dependence between two variables, like pointwise mutual information, while avoiding the bias of the latter towards low frequency items. In our experiments we have chosen some values of *N* (*N*=50, 100, 150, 200 and 250), and we leave the optimization of this parameter for future experiments.

Once the *N* most relevant contexts of $w_1$and w_2 have been selected, *APSyn* calculates the extent of their intersection, by summing up for each intersected context a function of its salience score. The idea behind such operation is that synonyms are likely to share more salient contexts than antonyms. For example, *dress* and *clothe* are very likely to have among their most relevant contexts words like *wear*, *thick*, *light* and so on. On the other hand, *dwarf* and *giant* will probably share contexts like *eat* and *sleep*, but they will differ on other very salient contexts such as *big* and *small*. To exemplify such idea, in Table 1 we report the first 16 most relevant contexts for the pairs of verbs *fall-lower* and *fall-raise*, respectively near-synonyms and antonyms.

Table 1
Top 16 contexts for the verbs to *fall*, to *lower* and to *raise*. These terms are present in our dataset. At this cutoff, the antonyms do not yet share any context.

TARGET	SYNONYM	ANTONYM
fall-v	**lower-v** (2 shared)	**raise-v** (0 shared)
1. love-n	1. cholesterol-n	1. awareness-n
2. category-n	2. raise-v	2. fund-n
3. short-j	3. level-n	3. money-n
4. disrepair-n	4. blood-n	4. issue-n
5. rain-n	5. cost-n	5. question-n
6. victim-n	6. pressure-n	6. concern-n
7. price-n (rank=7)	**7. rate-n** (rank=7)	7. profile-n
8. disuse-n	**8. price-n** (rank=8)	8. bear-v
9. cent-n	9. risk-n	9. standard-n
10. rise-v	10. temperature-n	10. charity-n
11. foul-j	11. water-n	11. help-v
12. hand-n	12. threshold-n	12. eyebrow-n
13. trap-n	13. standard-n	13. level-n
14. snow-n	14. flag-n	14. aim-v
15. ground-n	15. age-n	15. point-n
16. rate-n (rank=16)	16. lipid-n	16. objection-n
17. …	17. …	17. …

APSyn weights the saliency of the contexts with the minimum rank among the two LMI ranked lists, containing the *N* most relevant contexts for w_1 *and* w_2. Mathematically, *APSyn* can be defined as follows:

$$APSyn(w_1, w_2) = \sum_{f \in N(F_1) \cap N(F_2)} \frac{1}{\min(rank_1(f_1), rank_2(f_2))} \quad (1)$$

where $N(F_x)$ is the list of the N most relevant contexts f of a term w_x, and $rank_x(F_x)$ is the rank of the feature f_x in such salience ranked feature list. It is important to note here that a small *N* would inevitably reduce the intersection, forcing most of the scores to the same values (and eventually to zero), independently on the relation the pair under examination holds. On the other hand, a very large value of *N* will inevitably include also contexts with very low values of LMI and, therefore, much less relevant for the target noun. Finally, it can be seen that *APSyn* assigns the highest scores to the identity pairs (e.g. *dog-dog*).

If *APSyn* assigns high scores to the near-synonyms, its inverse – *APAnt* – is intended to assign high scores to the antonyms:

$$APAnt(w_1, w_2) = \frac{1}{APSyn(w_1, w_2)} \quad (2)$$

Two cases need to be considered here:

- if *APSyn* has not found any intersection among the *N* most relevant contexts, it will be set to zero, and consequently *APAnt* will be infinite;
- if *APSyn* has found a large and salient intersection, it will get a high value, and consequently *APAnt* will have a very low one.

The first case happens when the two terms in the pair are distributionally unrelated or when *N* is not sufficiently high. Therefore, *APant* is set to the maximum attested value. The second case, instead, can occur when two terms are distributionally very similar, sharing therefore many salient contexts. Ideally, this should only be the case for near-synonyms.

As we will see in Section 7, most of the scores given by *APSyn* and *APAnt* are either very high or very low. In order to scale them between 0 and 1, we use the *Min-Max function* (our infinite values will be set – together with the maximum ones – to 1):

$$MinMax(x_i) = \frac{x_i - \min(X)}{\max(X) - \min(X)} \quad (3)$$

Two variants of *APSyn* (and consequently of *APAnt*) have been also tested: *APSyn2* and *APSyn3*. Below we define them with the same notation as in the equation (1), while *APAnt2* and *APAnt3* can be defined as their respective reciprocal:

$$APSyn2(w_1, w_2) = \sum_{f \in N(F_1) \cap N(F_2)} \frac{1}{(rank_1(f_1) + rank_2(f_2))/2} \quad (4)$$

$$APSyn3(w_1, w_2) = \sum_{f \in N(F_1) \cap N(F_2)} \frac{\cos(w_1, w_2)}{(rank_1(f_1) + rank_2(f_2))/2} \quad (5)$$

The first variant simply uses the average rank rather than the minimum one, as a saliency index. The second variant introduces the use of the cosine as numerator instead of simply using the constant 1. While *APSyn2* is mainly meant to normalize *APSyn*'s denominator, *APSyn3* introduces a new criterion for measuring the distributional similarity between the pairs. In fact, both strongly and weakly related pairs may share some relevant contexts. If the extent of such sharing is not enough discriminative, the use of the *vector cosine* adds a discriminative criterion, which should assign higher scores to strongly related pairs.

5. Performance Evaluation

In order to evaluate *APAnt* and its variants, we set up two *antonym retrieval* tasks (AR). These two tasks consist of scoring pairs of words belonging to known semantic relations with *APAnt*, its variants and three baselines (i.e. *vector cosine, frequency of co-occurrence, random rank*), and then evaluate the resulting ranks with the *Average Precision* (AP; Kotlerman et al., 2010). In task 1, we only evaluate ranks consisting of pairs related by antonymy and synonymy, whereas in task 2 we also introduce hypernymy and co-hyponymy (henceforth, coordination).

DSM. In our experiments, we use a standard window-based DSM recording word co-occurrences within the two nearest content words to the left and right of each target. Co-occurrences are extracted from a combination of the freely available ukWaC and WaCkypedia corpora (with 1.915 billion and 820 million words, respectively) and weighted with LMI (Santus et al., 2014a).

DATASETS. To assess APAnt, we rely on a joint dataset consisting of subsets of English word pairs extracted from the Lenci/Benotto dataset (Santus et al., 2014b), BLESS (Baroni and Lenci, 2011) and EVALution 1.0 (Santus et al., 2015). Our final dataset for task 1 contains 4,735 word pairs, including 2,545 antonyms and 2,190 synonyms. The class of antonyms consists of 1,427 noun pairs (e.g. *parody-reality*), 420 adjective pairs (e.g. *unknown-famous*) and 698 verb pairs (e.g. *try-procrastinate*). The class of synonyms consists of 1,243 noun pairs (e.g. *completeness-entirety*), 397 adjective pairs (e.g. *determined-focused*) and 550 verb pairs (e.g. *picture-illustrate*).

For task 2, we aimed at discriminating antonyms also from relations other than synonyms. Thus, we also include 4,261 hypernyms from the Lenci/Benotto dataset, BLESS and EVALution, and 3,231 coordinates from BLESS. The class of hypernyms consists of 3,251 noun pairs (e.g. *violin-instrument*), 364 adjective pairs (e.g. *able-capable*) and 646 verb pairs (e.g. *journey-move*). The coordinates only include noun pairs (e.g. *violin-piano*).

EVALUATION MEASURE and BASELINES. The ranks obtained by sorting the scores in a decreasing way were then evaluated with *Average Precision* (Kotlerman et al., 2010), a measure used in *Information Retrieval* (IR) to combine precision, relevance ranking and overall recall. Since *APAnt* has been designed to identify antonyms, we would expect AP=1 if all antonyms are on top of our rank, AP=0 if they are all placed in the bottom.

Finally, for both tasks we have used three baselines for performance comparison: *vector cosine*, *co-occurrence frequency* and *random rank*. While the *vector cosine* is motivated by the fact that antonyms have a high degree of distributional similarity, the *random rank* should keep information about the different sizes of the classes. The frequency of co-occurrence, then, is motivated by the *co-occurrence hypothesis* (Charles and Miller, 1989). Our implementation of such baseline is supported by several examples in Justeson and Katz (1991), where the co-occurrence is mostly found within the window adopted in our DSM (e.g. coordination, etc.).

6. Experimental Results

In Table 2, we report the AP values for all the variants of *APAnt* and the baselines. Since the Average Precision values may be biased by pairs obtaining the same scores – in these cases, in fact, the rank cannot be univocally determined, except by assigning it randomly or adding a new criterion (we have adopted the alphabetic one) –, for every measure, we provide information about how many pairs have identical scores. As it can be seen in the table, when *N* is big enough (in our case *N>=200*), *APAnt* has less identical scores than the *vector cosine*.

Table 2
AP scores for APAnt, its variants and the baselines on the dataset containing 4,735 word pairs, including 2,545 antonyms and 2,190 synonyms. The second column contains the values of *N* (only for APAnt) and – between brackets – the quantity of pairs having identical scores. Note: three values are provided for APAnt (i.e. one for each variant), while for the other measures only one.

MEASURE	***N* (Pairs with identical score: APAnt, APAnt2, APAnt3)**	**Antonyms (APAnt2, APAnt3)**	**Synonyms (APAnt2, APAnt3)**
APAnt	50 (1672, 1374, 703)	0.60 (0.60, 0.60)	0.41 (0.41, 0.41)
APAnt	100 (339, 274, 180)	0.60 (0.60, 0.60)	0.41 (0.41, 0.41)
APAnt	150 (118, 96, 86)	0.60 (0.61, 0.60)	0.41 (0.40, 0.41)
APAnt	200 (75, 67, 64)	0.61 (0.61, 0.60)	0.40 (0.40, 0.41)
APAnt	**250 (75, 67, 64)**	**0.61 (0.61, 0.60)**	**0.40 (0.40, 0.41)**
Co-occurrence	(3591)	0.54	0.46
Cosine	(85)	0.50	0.50
Random	(3)	0.55	0.45

APAnt and its variants obtain almost the same AP scores, outperforming all the baselines. *APAnt3* seems to perform slightly worse than the other variants. Given that our dataset contains few more antonyms than synonyms, we expect the *random rank* to have a certain preference for antonyms. This is, in fact, what happens, making the random baseline outperforming the *co-occurrence baseline*. The *vector cosine*, instead, has a preference for synonyms, balancing the AP independently of the different sizes of the two classes. Finally, we can notice that while the values of *N* seem to have a small impact on the performance, they have a high impact in reducing the number of identical scores. That is, the larger the value of *N*, the less pairs have identical scores. Co-occurrence frequency is the worst measure in this sense, since almost 76% of the pairs obtained identical scores. Such a high number has to be attributed to the sparseness of the data and may be eventually reduced by choosing a larger window in the construction of the DSM. However, this also shows that use of co-occurrence data alone may be of little help in discriminating antonyms from other semantic relations.

In Table 3 we report the AP scores for the second AR task, which is performed on a dataset including also hypernyms and coordinates. Again, *APAnt* and its variants outperform the baselines. *APAnt3* is confirmed to work slightly worse than the other variants. An interesting and unexpected result is obtained for the hypernyms. Even though their class is almost twice the size of antonyms and synonyms (this can be seen also in the AP scores obtained by the baselines), this result is important and it will be discussed in Section 7. Once more, the AP value for the *random rank* is proportional to the sizes of the classes. Co-occurrence frequency seems to have a slight preference for antonyms and hypernyms (which may be due to the size of these classes), while the *vector cosine* seems to prefer synonyms and coordinates.

Table 3
AP scores for the APAnt, its variants and the baselines on the dataset containing 12,227 word pairs, including 4,261 hypernyms and 3,231 coordinates. The second column contains the values of *N* (only for APAnt) and – between brackets – the quantity of pairs having identical scores. Note: three values are provided for APAnt (i.e. one for each variant), while for the other measures only one.

MEASURE	*N* (Pairs with identical score: APAnt, APAnt2, APAnt3)	Antonyms (APAnt2, APAnt3)	Synonyms (APAnt2, APAnt3)	Hypernyms (APAnt2, APAnt3)	Coordinates (APAnt2, APAnt3)
APAnt	50 (5543, 4756, 3233)	0.26 (0.27, 0.26)	0.18 (0.18, 0.18)	0.42 (0.43, 0.42)	0.18 (0.18, 0.18)
APAnt	100 (2600, 2449, 2147)	0.27 (0.27, 0.26)	0.18 (0.18, 0.18)	0.43 (0.44, 0.43)	0.18 (0.17, 0.18)
APAnt	150 (2042, 1987, 1939)	0.27 (0.28, 0.26)	0.18 (0.18, 0.18)	0.43 (0.44, 0.42)	0.18 (0.17, 0.18)
APAnt	200 (1951, 1939, 1907)	0.28 (0.28, 0.26)	0.18 (0.18, 0.18)	0.43 (0.44, 0.42)	0.17 (0.17, 0.18)
APAnt	**250 (1939, 1901, 1892)**	**0.28 (0.28, 0.26)**	**0.18 (0.18, 0.18)**	**0.43 (0.44, 0.42)**	**0.17 (0.17, 0.18)**
Co-occ.	(10760)	0.23	0.19	0.36	0.23
Cosine	(2096)	0.20	0.20	0.31	0.29
Random	(15)	0.21	0.18	0.35	0.26

Once more, the values of *N* do not significantly affect the AP scores, but they influence the number of identical scores (*N>=150* is necessary to have less identical scores than those obtained with the *vector cosine*). Co-occurrence frequency is again the worst measure in this sense, since it has as many as 10,760 pairs with the same score on 12,227 (88%).

7. Discussion and Distribution of Scores

The AP scores shown and discussed in the previous section confirm that *APAnt* assigns higher scores to antonyms compared to both synonyms and coordinates. Such results is coherent with our hypothesis that antonyms share less relevant contexts than both synonyms and coordinates. Figure 1 shows boxplots[6] describing the distribution of scores for *APAnt* (on the left) and *vector cosine* (on the right). As it can be seen, *APAnt* scores are – on average – higher for antonymy, while the *vector cosine* scores are similarly distributed for both relations.

A surprising result instead occurs for the class of hypernyms, as shown in Table 3, to which *APAnt* assigns high scores. Although such class is almost twice the size of both antonyms and synonyms, the *APAnt* AP score for such class is much higher than the AP scores assigned to the baselines, even overcoming the

[6] Boxplots display the median of a distribution as a horizontal line within a box extending from the first to the third quartile, with whiskers covering 1.5 times the interquartile range in each direction from the box, and outliers plotted as circles.

value reached with antonyms. The reason may be that hypernymy related pairs – even though they are known to be characterized by high distributional similarity – do not share many salient contexts. In other words, even though hypernyms are expected to share several contexts, they do not seem to share a large amount of their most mutually dependent ones. That is, contexts that are salient for one of the two terms (e.g. *wild* for the hypernym *animal*) are not necessarily salient for the other one (e.g. the hyponym *dog*), and viceversa (e.g. *bark* is not salient for the hypernym *animal*, while it is for the hyponym *dog*). This result is coherent with what we have found in Santus et al. (2014a), where we have shown how hypernyms tend to co-occur with more general contexts compared to hyponyms, which are instead likely to occur with less general ones. More investigation is required in this respect, but it is possible that *APAnt* (or its variants) can be used in combination with other measures (e.g. *SLQS* or entropy) for discriminating also hypernymy.

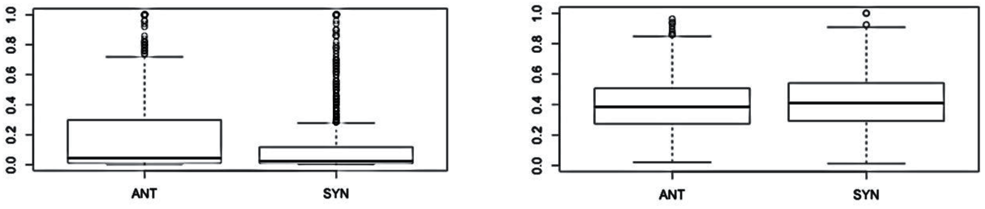

Figure 1
APAnt scores (on the left) for N=50 and *vector cosine* ones (on the right).

Another relevant point is the role of *N*. As it can be seen from the results, it has a low impact on the AP values, meaning that the rank is not strongly affected by its change (at least for what concerns the values we have tested, which are 50, 100, 150, 200 and 250). However, the best results are generally obtained with *N>150*. The value of *N* is instead inversely proportional to the number of identical scores (the same can be said also for the two variants, *APAnt2* and *APAnt3*, which generates slightly fewer identical scores than *APAnt*).

For what concerns the variants, *APAnt2* and *APAnt3* have been shown to perform in a very similar way to *APAnt*. *APAnt3*, in particular, achieves slightly worse results than the other two measures in the second task. We believe that this measure should be tested against other semantic relations in the future.

Finally, during our experiments, we have found that AP may be subjected to a bias that is concerned with how to rank pairs that have obtained the same score. In this case, we have used the alphabetical order as the secondary criterion for ranking. Such criterion does not affect the evaluation of APAnt (including its variants) and *vector cosine*, as these measures assign a fairly small amount of identical scores (around 15% of 12,227 pairs). It instead certainly affects the reliability of the co-occurrence frequency, where the amount of pairs obtaining identical scores amount up to 88%. Even though such result is certainly imputable to the sparseness of the data, we should certainly consider whether the co-occurrence frequency can properly account for antonymy.

8. Conclusions

In this paper, we have further described and analyzed *APAnt*, a distributional measure firstly introduced in Santus et al. (2014b, 2014c). Two more variants have been proposed for the normalization of *APAnt* and for the extension of its scope to the discrimination of antonymy from semantic relations other than synonymy. *APAnt* and its variants have been shown to outperform several baselines in our experiments. Surprisingly, they seem to assign high scores to hypernyms, which do probably share few salient contexts too. This fact suggests the need for further refinement of the APant.

APAnt should not be considered as the final result of this research, but much more as a work in progress. It should be further explored and improved to put light on some distributional properties of antonymy and other semantic relations, which can be exploited to develop a unified method that may account for issues that are currently treated as separate tasks, such as *sense disambiguation* and *semantic relations identification*. In this sense, we believe that there are many properties that need to be further explored by looking at the most relevant contexts of each term, rather than at their full set. Such exploration and investigation should be linguistically grounded and should aim not only to the improvement of algorithms' performance, but also to a better understanding of the linguistic properties of semantic relations.

Acknowledgements

This work is partially supported by HK PhD Fellowship Scheme under PF12-13656.

References

Baroni, Marco and Alessandro Lenci. 2010. Distributional Memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Baroni, Marco and Alessandro Lenci. 2011. How we BLESSed distributional semantic evaluation. In *Proceedings of the EMNLP 2011, Geometrical Models for Natural Language Semantics Workshop* (GEMS 2011), 1-10, Edinburg, UK.

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

Cruse, David A. 1986. *Lexical Semantics*. Cambridge University Press, Cambridge.

Evert, Stefan. 2005. *The Statistics of Word Cooccurrences*. Dissertation, Stuttgart University.

Deese, J. 1964. The Associative Structure of Some Common English Adjectives. *Journal of Verbal Learning and Verbal Behavior*, 3:347–57.

Deese J. 1965. *The Structure of Associations in Language and Thought*. Johns Hopkins University Press, Baltimore.

Ding, Jing and Chu-Ren Huang. 2014. Word Ordering in Chinese Opposite Compounds. In Xinchun Xu and Tingting He (Eds.), *Chinese Lexical Semantics: 15th Workshop, CLSW 2014, Macao, China, July 9-12, 2012, Revised Selected Papers* (pp. 12-20). Berlin Heidelberg: Springer-Verlag. DOI: 10.1007/978-3-319-14331-6_2

Ding, Jing, and Chu-Ren Huang. 2013. Markedness of opposite. In Pengyuan Liu and Qi Su (Eds.), *Chinese Lexical Semantics: 14th Workshop, CLSW 2013, Zhengzhou, China, May 10-12, 2013. Revised Selected Papers* (pp. 191-195). Berlin Heidelberg: Springer-Verlag. DOI: 10.1007/978-3-642-45185-0_21

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

Harris, Zellig. 1954. Distributional structure. *Word*, 10(23):146–162.

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

Huang, Chu-Ren, I-Li Su, Pei-Yi Hsiao, and Xiu-Ling Ke. 2007. Paranyms, Co-Hyponyms and Antonyms: Representing Semantic Fields with Lexical Semantic Relations. In *Proceedings of*

Chinese Lexical Semantics Workshop 2007, pages 66-72, Hong Kong Polytechnic University, May 20-23.

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

Kempson, Ruth M. 1977. *Semantic Theory*. Cambridge University Press, Cambridge.

Kotlerman, Lili, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. 2010. Directional Distributional Similarity for Lexical Inference. *Natural Language Engineering*, 16(4):359–389.

Landauer, Thomas K. and Susan T. Dumais. 1997. A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211-240.

Lenci, Alessandro. 2008. Distributional semantics in linguistic and cognitive research. In A. Lenci (ed.), *From context to meaning: distributional models of the lexicon in linguistics and cognitive science*, *Italian Journal of Linguistics*, 20(1):1–31.

Lin, Dekang, 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 1,492–1,493, Acapulco.

Lobanova, Anna. 2012. *The Anatomy of Antonymy: a Corpus-driven Approach*. Dissertation. University of Groningen.

Lobanova, Anna, Tom van der Kleij, and Jennifer Spenader. 2010. Defining antonymy: A corpus-based study of opposites by lexico-syntactic patterns. *International Journal of Lexicography*, 23(1):19–53.

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

de Marneffe, Marie-Catherine, 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)*, pages 1,039–1,047, Columbus, OH.

Marton, Yuval, Ahmed El Kholy, and Nizar Habash. 2011. Filtering antonymous, trend-contrasting, and polarity-dissimilar distributional paraphrases for improving statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 237–249, Edinburgh.

Mihalcea, Rada 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.

Mohammad, Saif, Bonnie Dorr, Graeme Hirst, and Peter D. Turney. 2013. Computing lexical contrast. *Computational Linguistics*, 39(3):555–590.

Mohammad, Saif, Bonnie Dorr, and Graeme Hirst. 2008. Computing word-pair antonymy. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2008)*, pages 982–991, Waikiki, HI.

Morlane-Hondère, François. 2015. What can distributional semantic models tell us about part-of relations? In *Proceedings of the NetWordS Final Conference on Word Knowledge and Word Usage: Representations and Processes in the Mental Lexicon*, vol. 1347, pages 46-50, CEUR-WS.org , Aachen (DEU).

Murphy, M. Lynne. 2003. *Semantic relations and the lexicon: antonymy, synonymy, and other paradigms*. Cambridge University Press, Cambridge, UK. ISBN 9780521780674

Padó, Sebastian and Mirella Lapata. 2007. Dependency-based Construction of Semantic Space Models. *Computational Linguistics*, 33(2):161–199.

Pantel, Patrick and Marco Pennacchiotti. 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 113-120, Sydney, Australia.

Platt, John C. 1998. Fast training of support vector machines using sequential minimal optimization. In *Advances in Kernel Methods: Support Vector Learning*, pages 185–208. MIT Press Cambridge, MA, USA.

Roth, Michael and Sabine Schulte im Walde. 2014. Combining word patterns and discourse markers for paradigmatic relation classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2:524–530, Baltimore, Maryland, USA.

Sahlgren, Magnus. 2006. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. dissertation, Department of Linguistics, Stockholm University.

Santus, Enrico, Alessandro Lenci, Qin Lu, and Sabine Schulte im Walde. 2014a. Chasing Hypernyms in Vector Spaces with Entropy. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 2:38–42, Gothenburg, Sweden.

Santus, Enrico, Qin Lu, Alessandro Lenci, and Chu-Ren Huang. 2014b. Unsupervised Antonym-Synonym Discrimination in Vector Space. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014*, 9-10 December 2014, Pisa, volume 1, pages 328-333, Pisa University Press.

Santus, Enrico, Qin Lu, Alessandro Lenci and Chu-Ren Huang. 2014c. Taking Antonymy Mask off in Vector Space. In *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computation* (PACLIC), pages 135-144, Phuket, Thailand.

Santus, Enrico, Frances Yung, Alessandro Lenci and Chu-Ren Huang. 2015. EVALution 1.0: An Evolving Semantic Dataset for Training and Evaluation of Distributional Semantic Models. In *Proceedings of the 4th Workshop on Linked Data in Linguistics* (LDL-2015), 64–69, Beijing, China.

Schulte im Walde, Sabine and Maximilian Köper. 2013. Pattern-based distinction of paradigmatic relations for German nouns, verbs, adjectives. In *Language Processing and Knowledge in the Web*, 184-198. Springer.

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

Turney, Peter D. 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.

Tungthamthiti, Piyoros, Enrico Santus, Hongzhi Xu, Chu-Ren Huang and Shirai Kiyoaki. 2015. Sentiment Analyzer with Rich Features for Ironic and Sarcastic Tweets. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation* (PACLIC), Shanghai, China.

Xu, Hongzhi, Enrico Santus, Anna Laszlo and Chu-Ren Huang. 2015. LLT-PolyU: Identifying Sentiment Intensity in Ironic Tweets. In *Proceedings of the 9th Workshop on Semantic Evaluation* (SemEval 2015), pages 673-678, Denver, Colorado, USA.

Witten, Ian H. and Eibe Frank. 1999. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco.