10.12871/CLICIT2014163

Unsupervised Antonym-Synonym Discrimination in Vector Space

Enrico Santus[*], Qin Lu[*], Alessandro Lenci[§], Chu-Ren Huang[*]

[*] The Hong Kong Polytechnic University, Hong Kong
e.santus@connect.polyu.hk, {qin.lu, churen.huang}@polyu.edu.hk

[§] University of Pisa, Italy
alessandro.lenci@ling.unipi.it

Abstract

English. Automatic detection of antonymy is an important task in Natural Language Processing (NLP). However, currently, there is no effective measure to discriminate antonyms from synonyms because they share many common features. In this paper, we introduce *APAnt*, a new Average-Precision-based measure for the unsupervised identification of antonymy using Distributional Semantic Models (DSMs). *APAnt* makes use of Average Precision to estimate the extent and salience of the intersection among the most descriptive contexts of two target words. Evaluation shows that the proposed method is able to distinguish antonyms and synonyms with high accuracy, outperforming a baseline model implementing the *co-occurrence hypothesis*.

Italiano. *Sebbene l'identificazione automatica di antonimi sia un compito fondamentale del Natural Language Processing (NLP), ad oggi non esistono sistemi soddisfacenti per risolvere questo problema. Gli antonimi, infatti, condividono molte caratteristiche con i sinonimi, e vengono spesso confusi con essi. In questo articolo introduciamo APAnt, una misura basata sull'Average Precision (AP) per l'identificazione automatica degli antonimi nei Modelli Distribuzionali (DSMs). APAnt fa uso dell'AP per stimare il grado e la rilevanza dell'intersezione tra i contesti più descrittivi di due parole target. I risultati dimostrano che APAnt è in grado di distinguere gli antonimi dai sinonimi con elevata precisione, superando la baseline basata sull'ipotesi della co-occorrenza.*

1 Introduction

Antonymy is one of the fundamental relations shaping the organization of the semantic lexicon and its identification is very challenging for computational models (Mohammad et al., 2008). Yet, antonymy is essential for many Natural Language Processing (NLP) applications, such as Machine Translation (MT), Sentiment Analysis (SA) and Information Retrieval (IR) (Roth and Schulte im Walde, 2014; Mohammad et al., 2013).

As well as for other semantic relations, computational lexicons and thesauri explicitly encoding antonymy already exist. Although such resources are often used to support the above mentioned NLP tasks, they have low coverage and many scholars have shown their limits: Mohammad et al. (2013), for example, have noticed that "more than 90% of the contrasting pairs in GRE closest-to-opposite questions are not listed as opposites in WordNet".

The automatic identification of semantic relations is a core task in computational semantics. Distributional Semantic Models (DSMs) have often been used for their well known ability to identify semantically similar lexemes using corpus-derived co-occurrences encoded as distributional vectors (Santus et al., 2014a; Baroni and Lenci, 2010; Turney and Pantel, 2010; Padó and Lapata, 2007; Sahlgren, 2006). These models are based on the *Distributional Hypothesis* (Harris, 1954) and represent lexical semantic similarity in function of distributional similarity, which can be measured by *vector cosine* (Turney and Pantel, 2010). However, these models are characterized by a major shortcoming. That is, they are not able to discriminate among different kinds of semantic relations linking distributionally similar lexemes. For instance, the nearest neighbors of *castle* in the vector space typically include hypernyms like *building*, co-hyponyms like *house*, meronyms like *brick*, antonyms like *shack*, together with other semantically related words. While impressive results have been achieved in the automatic

identification of synonymy (Baroni and Lenci, 2010; Padó and Lapata, 2007), methods for the identification of hypernymy (Santus et al., 2014a; Lenci and Benotto, 2012) and antonymy (Roth and Schulte im Walde, 2014; Mohammad et al., 2013) still need much work to achieve satisfying precision and coverage (Turney, 2008; Mohammad et al., 2008). This is the reason why semi-supervised pattern-based approaches have often been preferred to purely unsupervised DSMs (Pantel and Pennacchiotti, 2006; Hearst, 1992)

In this paper, we introduce a new Average-Precision-based distributional measures that is able to successfully discriminate antonyms from synonyms, outperforming a baseline implementing the *co-occurrence hypothesis,* formulated by Charles and Miller in 1989 and confirmed in other studies, such as those of Justeson and Katz (1991) and Fellbaum (1995).

2 Defining Semantic Opposition

People do not always agree on classifying word-pairs as antonyms (Mohammad et al., 2013), confirming that antonymy classification is indeed a difficult task, even for native speakers of a language. Antonymy is in fact a complex relation and opposites can be of different types, making this class hard to define (Cruse, 1986).

Over the years, many scholars from different disciplines have tried to contribute to its definition. Though, 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. Cruse (1986) points out the paradox of simultaneous similarity and difference between the antonyms, claiming that opposites are indeed similar in every dimension of meaning except in a specific one (e.g. both *giant* and *dwarf* refer to a person, with a head, two legs and two feet, but their size is different).

In our work, we aim to distinguish antonyms from synonyms. Therefore we will adopt the word "antonym" in its broader sense.

3 Related Works

Most of the work about the automatic antonymy identification is based on the *co-occurrence hypothesis*, proposed by Charles and Miller (1989), who have noticed that antonyms co-occur in the same sentence more often than expected by chance (Justeson and Katz, 1991; Fellbaum, 1995).

Other automatic methods include pattern based approaches (Schulte im Walde and Köper, 2013; Lobanova et al., 2010; Turney, 2008; Pantel and Pennacchiotti, 2006; Lin et al., 2003), which rely on specific patterns to distinguish antonymy-related pairs from others. Pattern based methods, however, are mostly semi-supervised. Moreover they require a large amount of data and suffer from low recall, because they can be applied only to frequently occurring words, which are the only ones likely to fit into the given patterns.

Mohammad et al. (2013) have 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. Their approach outperforms other measures, but still is not completely unsupervised and it relies on thesauri, which are manually created resources.

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

4 *APAnt*: an Average-Precision-based measure

Antonyms are often similar in every dimension of meaning except one (e.g. *giant* and *dwarf* are very similar and they differ only in respect to the size).

This peculiarity of antonymy – called by Cruse (1986) the *paradox of simultaneous similarity and difference* – has an important distributional correlate. Antonyms occur in similar contexts exactly as much as synonyms do, making the DSMs models unable to discriminate them. However, according to Cruse's definition, we can expect there to be a dimension of meaning in which antonyms have a different distributional behaviour. We can also hypothesize that this dimension of meaning is a salient one and that it can be used to discriminate antonyms from synonyms. For example, *size* is the salient dimension of meaning for the words *giant* and *dwarf* and we can expect that while *giant* occurs

more often with words such as *big*, *huge*, etc., *dwarf* is more likely to occur in contexts such as *small*, *hide*, and so on.

To verify this hypothesis, we select the *N* most salient contexts of the two target words (*N*=100[1]). We define the salience of a context for a specific target word by ranking the contexts through *Local Mutual Information* (LMI, Evert, 2005) and collecting the first *N*, as already done by Santus et al. (2014a). Once the *N* most salient contexts for the two target words have been identified, we verify the extent and the salience of the contexts shared by both the words. We predict that synonyms share a number of salient contexts that is significantly higher than the one shared by antonyms. To estimate the extent and the salience of the shared contexts, we adapt the Average Precision measure (AP; Voorhees and Harman, 1999), a common Information Retrieval (IR) evaluation metric already used by Kotlerman et al. (2010) to identify lexical entailment. In IR systems, this measure is used to evaluate the ranked documents returned for a specific query. It assigns high values to the rankings in which most or all the relevant documents are on the top (recall), while irrelevant documents are either removed or in the bottom (precision). For our purposes, we modify this measure in order to increase the scores as a function of (1) the size of the intersection and (2) the salience of the common features for the target words. To do so, we consider the common contexts as relevant documents and the maximum salience among the two target words as their rank. In this way, the score will be promoted when the context is highly salient for at least one of the two target words in the pair. For instance, in the pair *dog* – *cat*, if *home* is a common context, and it has salience=1 for *dog* and salience=*N*-1 for *cat*, we will consider *home* as a relevant document with rank=1. Formula (1) below provides the formal definition of *APAnt* measure:

$$APAnt = 1/\sum_{f \epsilon F_1 \cap F_2} \frac{1}{min(rank_1(f_1), rank_2(f_2))} \quad (1)$$

where *Fx* is the set of the *N* most salient features of a term *x* and $rank_x(f_x)$ is the position of the feature f_x in the salience ranked feature list for the term *x*. It is important to note that *APAnt* is defined as a reciprocal measure, so that the higher scores are assigned to antonyms.

5 Experiments and Evaluation

The evaluation includes two parts. The first part is a box-plot visualization to summarize the distributions of scores per relation. In the second part, the Average Precision (AP; Kotlerman et al., 2010) is used to compute the ability of our proposed measure to discriminate antonyms from synonyms. For comparison, we take as the baseline a model using the co-occurrence frequency of the target pairs.

5.1 The DSM and the Dataset

For the evaluation, we use a standard window-based DSM recording co-occurrences with context window of the nearest 2 content words both to the left and right of each target word. 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.

To assess *APAnt*, we rely on a subset of English word-pairs collected by Lenci and Benotto in 2012/13 using Amazon Mechanical Turk, following the method described by Schulte im Walde and Köper (2013). Among the criteria used for the collection, Lenci and Benotto balanced target items across word categories and took in consideration the frequency, the degree of ambiguity and the semantic classes.

Our subset contains 2.232 word-pairs[2], including 1.070 antonymy-related pairs and 1.162 synonymy-related pairs. Among the antonymy-related pairs, we have 434 noun-pairs (e.g. *parody-reality*), 262 adjective-pairs (e.g. *unknown-famous*) and 374 verb-pairs (e.g. *try-procrastinate*); among the synonymy-related pairs, we have 409 noun-pairs (e.g. *completeness-entirety*), 364 adjective-pairs (e.g. *determined-focused*) and 389 verb-pairs (e.g. *picture-illustrate*).

[1] *N=100* is the result of an optimization of the model against the dataset. Also the following suboptimal values have been tried: 50 and 150. In all the cases, the model outperformed the baseline.

[2] The sub-set include all the pairs for which both the target words exist in the DSM.

5.2 Results

5.2.1 *APAnt* Values Distribution

Figures 1 and 2 show the box-plots summarizing respectively the logarithmic distributions of *APAnt* and baseline scores for antonyms and synonyms. The logarithmic distribution is used to normalize the range of data, which would be otherwise too large and sparse for the box-plot representation.

Box-plots 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 of the interquartile range in each direction from the box, and outliers plotted as circles.

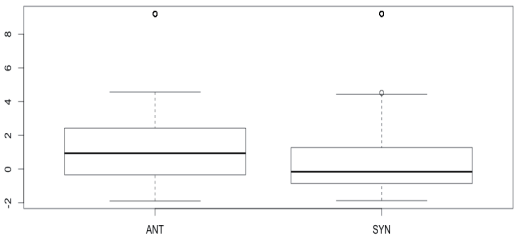

Figure 1: Logarithmic distribution of *APAnt* scores (*N=100*)

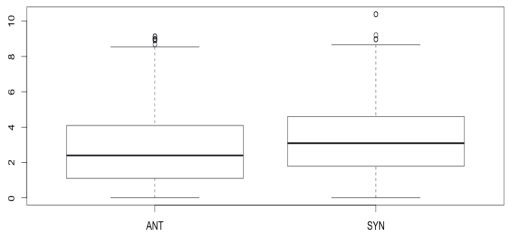

Figure 2: Logarithmic distribution of the baseline scores[3].

In Figure 2, we can observe that the baseline promotes synonyms over antonyms and also that there is a large range of overlap among synonyms and antonyms distributions, showing the weakness of the co-occurrence hypothesis on our data. On the other hand, in Figure 1 we can observe that, on average, *APAnt* scores are much higher for antonymy-related pairs and that the overlap is much smaller. In terms of distribution of values, in fact, synonyms have much lower values for *APAnt*.

[3] 410 pairs with co-occurrence equal to zero on a total of 2.232 have been removed to make the box-plot readable (i.e. *log(0)=-inf*).

5.2.2 Average Precision

Table 1 shows the second performance measure we used in our evaluation, the Average Precision (Lenci and Benotto, 2012; Kotlerman et al., 2010) per relation for both *APAnt* and baseline scores. As already mentioned above, AP is a method used in Information Retrieval to combine precision, relevance ranking and overall recall. The best possible score would be 1 for antonymy and 0 for synonymy.

	ANT	SYN
APAnt	**0.73**	0.55
Baseline	0.56	0.74

Table 1: Average Precision (AP).

Table 1 shows that *APAnt* is a much more effective measure for antonymy identification as it achieves +0.17 compared to the baseline. This value results in a 30% improvement for antonymy identification. This improvement comes together with a higher ability in discriminating antonyms from synonyms. The results confirm the trend shown in the box-plots of Figure 1 and Figure 2. *APAnt* clearly outperforms the baseline, confirming the robustness of our hypothesis.

6 Conclusions and Ongoing Work

This paper introduces *APAnt*, a new distributional measure for the identification of antonymy (an extended version of this paper will appear in Santus et al., 2014b).

APAnt is evaluated in a discrimination task in which both antonymy- and synonymy-related pairs are present. In the task, *APAnt* has outperformed the baseline implementing the *co-occurrence hypothesis* (Fellbaum, 1995; Justeson and Katz, 1991; Charles and Miller, 1989) by 17%. *APAnt* performance supports our hypothesis, according to which synonyms share a number of salient contexts that is significantly higher than the one shared by antonyms.

Ongoing research includes the application of *APAnt* to discriminate antonymy also from other semantic relations and to automatically extract antonymy-related pairs for the population of ontologies and lexical resources. Further work can be conducted to apply *APAnt* to other languages.

Acknowledgments

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.

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.

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.

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.

Lenci, Alessandro and Giulia Benotto. 2012. Identifying hypernyms in distributional semantic spaces. In *SEM 2012 – The First Joint Conference on Lexical and Computational Semantics*, 2:75–79, Montréal, Canada.

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, 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.

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.

Padó, Sebastian and Lapata, Mirella. 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.

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. Taking Antonymy Mask off in Vector Space. To Appear in *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computation (PACLIC),* Phuket, Thailand.

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