Scope International Journal of Science, Humanities, Management and Technology. ISSN : 2455-068X

Vol.3 Issue 4 (2017) 179 - 183. Submitted 25/11/2017. Published 30/12/2017

A SYNONYM FOR AVERAGE

[1]T. Indhu, [2]R. Kosalai Pushpam,
Assistant Professor, Department of Statistics, Shri Sakthi Kailassh Women's College, Salem.

Abstract:

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.

Keywords: DSM, APAnt, NLP.

1. INTRODUCTION

Antonymy is one of the fundamental relations shaping the organization of the semantic lexicon nd 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. 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 . 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. RELATED WORK

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. 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.

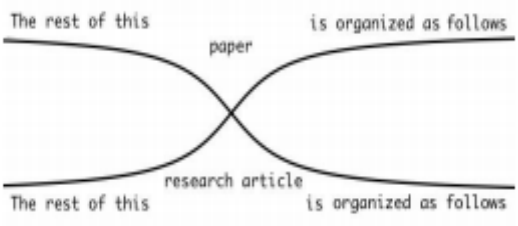

Fig.1.Phara Cast

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.

Here we differ from the pattern-based approach that use local general environment. We propose to align paraphrases from domain corpora and discover words that are possibly substitutable for one another in a given context (paraphrase casts), and as such are syn- onymsornear-synonyms. Comparativelytoexist- ing approaches, we propose an unsupervised and language-independent methodology which does not depend on linguistic processing2, nor manual definition of patterns or training sets and leads to higher precision when compared to distributional similarity-based approaches. The main goal of our research is to build knowl- edge resources in different domains that can ef- fectively be used in different NLP applications. As such, precision takes an important part in the overall process of our methodology. For that pur- pose, we first identify the phrasal terms (or multi- word units) present in the corpora. Indeed, it has beenshowninmanyworksthatphrasaltermscon- vey most of the specific contents of a given do- main. Our

approach to term extraction is based on linguistic pattern matching and Inverse Doc- ument Frequency (IDF) measurements for term.

3. METHODOLOGY

A few unsupervised metrics have been applied to automatic paraphrase identification and extraction (Barzilay and McKeown, 2001) and (Dolan et al., 2004). However, these unsupervised methodolo- gies show a major drawback by extracting quasi- exact or even exact match pairs of sentences as they rely on classical string similarity measures. Such pairs are useless for our research purpose. More recently, (Cordeiro et al., 2007a) proposed the sumo metric specially designed for asymmet- rical entailed pair identification in corpora which obtained better performance than previously es- tablished metrics, even in corpora with exclu- sively symmetrical en We can attribute different levels of confidence to different paraphrase casts. Indeed, the larger the contexts and the smaller the misaligned se- quences are, the more likely it is for single or phrasal terms to be synonyms or near-synonyms. Note that in the cast shown in figure 3, each con- text has a significant size, with four words on each side, and themisaligned segments arein fact equivalent expressions i.e. ”paper” is a synonym of ”research article”.

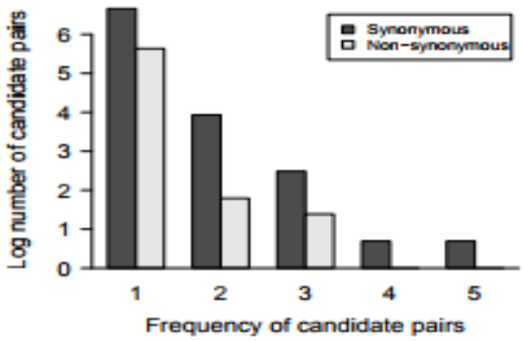

Fig.2.Frequency pair

In the analyzed domain these expressions are equivalent and interchange mains. For the purpose of this paper, we only take into account the casts where the misaligned sequences of words contain only one word or one multi-. Thatis,wehave a one-to-one match. However, no constraints - sults are provided in the next section. To evaluate our methodology we have used two different corpora, both from scientific do- mains built from abstracts of publications (see Table 1). The corpus of computer secu- rity (COCS) is a collection of 4854 abstracts on computer security extracted from the IEEE (http://ieee.rkbexplorer.com/) repository7. The corpus of cancer research (COCR) contains 3334 domain specific abstracts of scientific publica- tions extracted from the PubMed8 on three types of cancer: (1) the mammary carcinoma register (COCR1) consisting of 1500 abstracts, (2) the nephroblastoma register (COCR2) consisting of 1500 abstracts, and (3) the rhabdoid tumor regis- ter (COCR3) consisting of 334 abstracts. In order to assess the quality of our results, we calculated the similarity between all extracted pairs of synonyms following the distributional analysis paradigm as in (Moraliyski and Dias, 2007)whobuildcontext9 featurevectorsfornoun synonyms. In particular, we used the cosine sim- ilarity measure and the Loglike association mea- sure (Dunning, 1993) as the weighting scheme of the context features. The distribution of the simi- larity measure for all noun synonyms (62 pairs).

4. ANALYSIS

Nearly half of the cases have similarities higher than 0.5. It is important to notice that a spe- cific corpus10 was built to calculate as sharply as possible the similarity measures as it is done in (Moraliyski and Dias, 2007). Indeed, based on the COCS and the COCR most statistics were in- significant leading to zero-valued features. This situation is well-known as it is one of the major drawbacks of the distributional

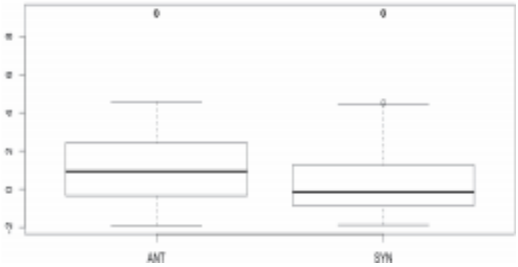

Fig.3.Analysis Graph

analysis approach which needs huge quantities of texts to make. So we note that applying the distributional analysis approach to such small corpora would have led to rather poor results. Even with an adapted corpus, figure 5 (left-most bar) shows thattherearenosufficientstatisticsfor30pairsof synonyms. for the computer security do- main and 66.06% for the cancer research domain. However, further improvements of the method should be considered. A measure o fquality of the paraphrase casts is necessary to provide a mea- sure of confidence to the kind of extracted word semantic relationship. Indeed, the larger the con- texts and the smaller the misaligned sequences are,the more likely it is for single or phras alterms to be synonyms or near-synonyms. Further work should also be carried out to differentiate the ac- quired types of semantically related pairs.

CONCLUSION

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.

REFERENCES

1. Baroni, Marco and Alessandro Lenci. 2010. Distributional Memory: A general framework for corpus-based semantics. Computational Linguistics, 36(4):673–721.
2. Charles, Walter G. and George A. Miller. 1989. Contexts of antonymous adjectives. Applied Psychology, 10:357–375.
3. Cruse, David A. 1986. Lexical Semantics. Cambridge University Press, Cambridge.

4. Evert, Stefan. 2005. The Statistics of Word Cooccurrences. Dissertation, Stuttgart University.
5. Fellbaum, Christiane. 1995. Co-occurrence and antonymy. International Journal of Lexicography, 8:281–303.
6. Harris, Zellig. 1954. Distributional structure. Word, 10(23):146–162.
7. 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.