On the use of antonyms and synonyms from a domain perspective

Debela Tesfaye
IT PhD Program
Addis Ababa University
Addis Ababa, Ethiopia

dabookoo@gmail.com

Carita Paradis
Centre for Languages and Literature
Lund University
Lund, Sweden

carita.paradis@englund.lu.se

Abstract

This corpus study addresses the question of the nature and the structure of antonymy and synonymy in language use, following automatic methods to identify their behavioral patterns in texts. We examine the conceptual closeness/distance of synonyms and antonyms through the lens of their DOMAIN instantiations.

1 Introduction

Using data from Wikipedia, this corpus study addresses the question of the nature and the structure of antonym and synonymy in language use. While quite a lot of empirical research using different observational techniques has been carried on antonymy (e.g. Roehm et al. 2007, Lobanova 2013, Paradis et al. 2009, Jones et al. 2012), not as much has been devoted to synonymy (e.g. Divjak 2010) and very little has been carried out on both of them using the same methodologies (Gries & Otani 2010). The goal of this study is to bring antonyms and synonyms together, using the same automatic methods to identify their behavioral patterns in texts. We examine the conceptual closeness/distance of synonyms and antonyms through the lens of their domain instantiations. For instance, *strong* used in the context of wind or taste (of tea) as compared to *light* and *weak* respectively, and *light* as compared to *heavy* when talking about rain or weight.

The basic assumption underlying this study is that the strength of co-occurrence of antonyms and synonyms is dependent on the domain in which they are instantiated and co-occur. In order to test the hypothesis we mine the co-occurrence information of the antonyms and the synonyms relative to the domains using a dependency grammar method. [1]

---

[1] http://nlp.stanford.edu/software/lexparser.shtml

The rationale is that the dependency parsing produces the relational information among the constituent words of a given sentence, which allows us to (i) extract co-occurrences specific to a given domain/context, and (ii) capture long distance co-occurrences between the word pairs. Consider (1).

1. Winters are cold and dry, summers are cool in the hills and quite hot in the plains.

In (1), the antonyms *cold*: *hot* modify winters and summers respectively. Those forms express the lexical concepts winter and summer in the domain temperature. The antonyms *cold*: *hot* co-occur but at a distance in the sentence. Thanks to the dependency information, it is possible to extract such long distance co-occurrences together with the concepts modified.

The article is organized as follows. In section 2, we describe the procedure and the two methods used: co-occurrence extraction of lexical items in the same sentence and a variant domain dependent co-occurrence extraction method. The latter method extracts patterns of co-occurrence information of the synonyms and antonyms in different sentences. In section 3 we present the results and discussions followed by a discussion of our results in comparison with related previous works in section 4. The conclusions are presented in section 5.

2 Procedure

Using an algorithm similar to the one proposed by Tesfaye & Zock (2012) and Zock & Tesfaye (2012), we extracted the co-occurrence information of the pairs in different domains separately, measuring the strength of their relation in the different domains with the aim of (i) making principled comparisons between antonyms and synonyms from a domain perspective, and (ii) determining the structure of antonymy and synonymy as categories in language and cognition.

Our algorithm is similar to the standard n-gram co-occurrences extraction algorithms, but

instead of using the linear ordering of the words in the text, it generates co-occurrences frequencies along paths in the dependency tree of the sentence as presented in the sections 2.2–2.5.

2.1 Training and testing data

The antonyms and synonyms employed for training and testing were extracted from the data used by Paradis et al. (2009) where the antonyms are presented according to their underlying dimensions and synonyms were provided for all the individual antonyms (for a description of the principles see Paradis et al. 2009). That set of antonyms and synonyms were used to extract their co-occurrence patterns from the Wikipedia texts in this study.

Dimensions	Antonyms	The associated synonyms of the antonyms
Size	Large	huge, vast, massive ,big ,bulky, giant ,gross, heavy, significant ,wide
	Small	little, low, minor, minute, petite, slim, tiny
Speed	Fast	quick, hurried, prompt, accelerating, rapid
	Slow	sudden, dull, gradual, lazy
Strength	Strong	forceful, hard, heavy, muscular, powerful, substantial, tough
	Weak	light, soft, thin, wimpy
Merit	Bad	crappy, defective, evil ,harmful, poor ,shitty ,spoiled ,unhappy
	Good	awful ,genuine ,great, honorable ,hot, neat, nice, reputable, right ,safe ,well

Table 1. The antonym pairs in their meaning dimensions and the associated synonyms.

2.2 Extracting the co-occurrences of the antonyms and synonyms in the respective domains

In order to extract the co-occurrences of the antonyms/synonyms in the respective domains we produced the relational information among the constituent words of a given sentence. To this end, we extracted the patterns linking the synonyms/antonyms and the concepts they modify and used this same pattern to extract more lexical concepts. The procedure was as follows.

- Start with the selected set of synonym/antonym pairs
- Extract sentences containing the pairs
- Identify the dependency information of the sentences
- Mine the dependency patterns linking the pairs with the concepts they modify
- Use these learned patters to extract further relations (synonym/antonym pairs and the associated concepts)

2.3 Extracting the domains

We created a matrix of antonym and synonym pairs matching every antonym and synonym from the list in Table 1. Using the patterns learned in section 2.2 we identified as many domains as possible for the pairs of synonyms and antonyms and calculated their frequency of co-occurrence in the respective domains.

When the lexical concepts were considered too specific, we referred them to more inclusive, superordinate domains. Frequency of occurrence was used as a criterion for conflation of concepts into superordinate ones as follows.

- Extract term co-occurrence frequencies within a window of sentences constituting both the antonyms/synonyms and the potential domain concepts. For instance:
  - Antonyms: *cold*: *hot*, domain concepts: winter, summer
  - Synonyms: *strong*: *heavy*, domain concepts: wind, rain
- Create a matrix of the potential domain concepts and the co-occurring terms with their frequencies
- Cluster them using the k-means algorithm
- Take the term with the maximal frequency (centroid) in each cluster and consider it the domain term
- Test the result using expert judgment running the algorithm on the test set.

Antonym/Synonym	Potential Domain concept	Words co-occurring with possible domain concepts	Frequency
hot cold	summer winter	temperature	50
		climate	43
		Wind	30

strong heavy	wind rain	wind rain	86
	winds snowfall	winds snowfall	3
	winds rainfall	winds rainfall	34
	waves rainfall	waves rainfall	4

Table 2. The matrix of the frequencies of terms co-occurring with sample antonyms and the associated potential domain concepts

2.4 Extracting co-occurrences frequency specific to a given Domain/Context

The algorithm calculated the co-occurrence frequency of the antonyms/synonyms with the different concepts they refer to (or modify) as presented in table 3 by combining the information obtained in section 2.3 and section 2.4.

Antonyms	Concept 1	Concept 2	Frequency	Domain
hot cold	summer	winter	10 5	temperature
strong heavy	wind	rain	11 2	winds rain
	winds	snowfall		
	winds	rainfall		
	waves	rainfall		

Table 3. The frequency of sample antonym specific to the underlying domains

2.5 Variant Domain Dependent Co-occurrence Extraction

In the previous algorithm, the co-occurrence information was extracted from the same sentence. However, unlike the antonyms, synonyms rarely occurred together in the same context (the same sentence and domain). It is natural to assume that in most cases synonyms are used in different contexts since they evoke similar but not identical meanings. This is however not the case for antonyms, which were always used to evoke properties of the same meanings when these antonymic words were used to express opposition (Paradis & Willners 2011), and in fact also when they are not used to express opposition (Paradis,et al., 2015). Because of this we decided to extract a variant domain dependent co-occurrence algorithm for the synonyms and antonyms, which instead extracts patterns of co-occurrence information of the synonyms and an-

tonyms in different sentences, because we expected synonyms to be applicable to different, rather than the same contexts, since complete overlap of meanings of words are rare or even non-existent. This way we were able to gain information indirectly about their use by extracting their co-occurrence when they appear separately in different sentences while still being instantiated in the same domain. We mined the co-occurrence information of the synonym/antonym pairs separately in all possible domains and check if they co-occurred in the same sorts of domains:

- X(y, f)
- Z(y, f)

Where,

> X and Z are a pair of a given antonym/synonym, Y is the domain within which the pairs of the antonym/synonym co-occur and f the frequency of the x-y or z-y co-occurrence.

The frequency of a pair of the antonyms/synonyms in the Y domain was counted and the same applies to the other pair. This made it possible to measure the degree of co-occurrence of the antonym/synonym pairs from the domain perspective indirectly.

3. Results and discussion

3.1 Co-occurrences in the same sentence

Based on the results of the experiment the strength of the antonyms/synonyms varies in relation to the domains of instantiation. Hence, the strength of the co-occurrence of antonyms and synonyms is a function of the domains. For instance, the antonyms: *slow: fast*, *slow: quick* and *slow: rapid* were used in completely different domains with little or no overlap. *Slow: fast* is used in the domains of motion, movement, speed; *slow: quick* is used for time, march, steps domains. The synonyms *powerful: strong* are used in the domains of voices, links, meaning; *strong: muscular* in the domains of legs, neck; *strong: heavy* are used in the domains of wind rain, waves rainfall, winds snow respectively; *intense: strong* in the domains of battle resistance, radiation gravity, updrafts clouds respectively.

We observed some unique patterns among the antonyms and synonyms as described below:

The antonyms:

- Co-occurred frequently in the same domain in the same sentence.

- The strength of the co-occurrence depends on the domain: *slow: fast* in the domains of growth, lines , motion, movement, speed ,trains, music, pitch; *slow: quick* in the domains of time, march, steps; *slow: gradual* in the domains of process, change, transition; *small: big* in the domains of screen, band; *small: large* in the domains of intestine, companies, businesses; *week: strong* in the domains of force, interaction, team, ties, points, sides, wind.

The Synonyms:
- Co-occurred in the same sentence but mainly in different domains. For instance, *fast: quick, strong: heavy.* Few co-occurrences in the same sentences in the same domains as exhibited by the pairs *gradual: slow* in the domains of process, change, development.
- The strength of the synonym co-occurrence depends on the domains. For instance, the synonyms *strong: heavy* in wind *and* rain domains respectively to express intensity; the synonyms *large: wide* in the domains of population and distribution domains respectively; *gradual: slow* in the domains of process, change, development; *small: low* in the domain of size cost, range, size weight, area, size price, amount density; *micro: small* in the domains of enterprises, businesses, entrepreneurs..

3.2 The variant domain dependent co-occurrence method

As mentioned before, the variant domain dependent co-occurrence extraction algorithm mines the patterns of co-occurrence information of the synonyms and antonyms in different sentences. The result from the variant co-occurrence experiment showed hardly any differences in the domains with which the synonyms and antonyms are associated. *Strong* in the domains of influence, force, wind, interactions, evidence, ties; *Heavy* in the domains of loss, rain, industry, traffic; *gradual: slow* in the domains of process, change, transition. However, we observed that the frequency of co-occurrence differed significantly. For instance, the frequency of the pair *gradual: slow* was 76 in same sentences experiment but 1436 in the variant co-occurrence experiment.

4 Comparison with related works

Previous research has shown that there are antonyms that are strongly opposing (canonical antonyms) (Paradis et al. 2009, Jones et al. 2012). Such antonyms are very frequent in terms of co-occurrence as compared to other antonyms: *small: large* as compared with *small: big*. In this experiment we found that the canonical antonyms are the set of antonyms the domains in which they function were numerous and productive. For instance the number of domains for *small*: *large* (11704) is by far greater than for *small*: *big* (120). However this doesn't make the antonym *small: large* more felicitous in all the domains. *Small: big* are the most felicitous antonyms for the domains such as screen, band as compared to *small: large*.

Measuring the strength of antonyms without taking domains into account provided higher values for the canonicals as they tended to be used in several domains. If domains were taken in to account, as we did in this experiment, all the antonyms were strong in their specific domains. The antonym pair *small*: l*arge* had higher value without considering domain in to account yet had 0.29 value in the domain of screen where *small*: *big* has much higher value (0.71). The values were calculated taking the frequency of co-occurrence of the domain term (*screen* in this case*)* with each antonyms and dividing it by the summation of the frequency of co-occurrence of the domain term (again *screen* in this case*)* with both antonyms (small big and small large).

5 Conclusion

The strength of the antonyms/synonyms varied in relation to the domains of instantiation. The use of antonyms and synonyms was very consistent with few overlaps across the domains. Similar results were observed in both experiments from the domain perspective although with significant differences in frequency. Antonyms frequently co-occurred in the same domains in the same sentences and synonyms co-occurred in different domains in the same sentences (with less frequency) and more frequently in different sentences in the same domains.

Acknowledgments

References

Dagmar Divjak. 2010. Structuring the lexicon: a clustered model for near-synonymy. Berlin: de Gruyter.

Gries Stefan Th. & N. Otani. 2010. Behavioral profiles: a corpus-based perspective on synonymy and antonymy. *ICAME Journal, 34:121–150.*

Jones Steven, M.L. Murphy, Carita Paradis & Caroline Willners. 2012. Antonyms in English: Construals, constructions and canonicity. *Cambridge University Press*, Cambridge, UK.

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

Carita Paradis. 2005. Ontologies and construals in lexical semantics. *Axiomathes*,15:541–573.

Carita Paradis, Caroline Willners & Jones Steven. 2009. Good and bad opposites: using textual and psycholinguistic techniques to measure antonym canonicity. *The Mental Lexicon, 4(3): 380–429.*

Carita Paradis, Simon Löhndorf , Joost van de Weijer & Caroline Willners. 2015. Semantic profiles of antonymic adjectives in discourse. *Linguistics, 53.1: 153 – 191.*

Roehm, D., I. Bornkessel-Schlesewsky, F. Rösler & M. Schlesewsky. 2007. To predict or not to predict: Influences of task and strategy on the processing o f semantic relations. *Journal of Cognitive Neuroscience, 19 (8):1259–1274.*

Debela Tesfaye. & Michael Zock. 2012. Automatic Extraction of Part-whole Relations. *In Proceedings of the 9th International Workshop on Natural Language Processing and Cognitive Science*.

Michael Zock. & Debela Tesfaye. 2012. Automatic index creation to support navigation in lexical graphs encoding part of relations. *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon (CogALex-III), COLING 2012.*