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]

1 Introduction

This corpus study addresses the question of the nature and the structure of antonymy 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 (semi-) 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.

In this study, we mined word pair co-occurrence information using language model based on dependency grammar. The model is similar to the standard n-gram co-occurrences extraction algorithms, but instead of using the linear ordering of the words in the text, our algorithm generates co-occurrences frequencies along paths in the dependency tree of the sentence. Hence, our algorithm relies on the dependency grammar as pro-duced by the Stanford dependency parser . The reliance on the dependency grammar allows us to mine at least the following two vital information:

1. To capture long distance co-occurrences between the word pairs
2. To extract co-occurrences specific to a given domain/dimension/context

In order to elaborate the arguments more, consider the following example:

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

In the above sentence, the antonyms hot: cold are co-occurring but in a very distant position in the sentence structure. One can extract such long distance co-occurrences if the size of window is a sentence. Yet, extracting the co-occurrences in the domains using a window of sentences is a challenge, without the dependency information. Without the dependency information mining the terms which the antonyms or/and synonyms is hardly possible.

As introduced before, the goal of this research is to mine co-occurrences specific to a given domain. The idea is based on our hypothesis “The nature and strength of co-occurrence of antonyms and the associated synonyms is dependent on the domain in which they are co-occurring”. In order to mine the co-occurrence information of the antonyms and the synonyms, we have decided to extract the domains first. One way of mining such co-occurrences is to relay on the dependency grammar. The dependency grammar produces the relational information among the constituent words of a given sentence. In our case, since the majority of the antonyms and their associated synonyms are adjectives, the concepts they modify are potential concepts expressing properties of various domains. In the above example, the antonyms cold: hot modify winters and summers respectively and are considered as the concepts expressing the domain temperature or climate. The concepts expressing the domains will be replaced by a term which is more descriptive of the encoded domain (please refer step 2.1 below). Accordingly, winters and summers are re-placed by temperature/climate, which are more representative of the domain. The use of dependency grammar is therefore crucial to mine such concepts, since the concepts modified by the antonyms and the synonyms might appear in a long distance.

2 Method

Using an algorithm similar to the one proposed by Tesfaye & Zock (2012) and Zock & Tesfaye (2012), we mine 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.

We mined word pair co-occurrence based on dependency grammar. The model is similar to the standard n-gram co-occurrences extraction algorithms but doesn’t only rely on the linear ordering of the words in the text. It rather generates co-occurrence frequencies along paths in the dependency tree of the sentence. For this task we used Stanford dependency parser.

2.1 Extracting the co-occurrences in the respective domains

The dependency grammar produces the relational information among the constituent words of a given sentence. In our case, the antonyms and their associated synonyms are adjectives. Hence, the concepts they modify are potential concepts expressing properties of various domains. We extracted the patterns linking the synonyms/antonyms and the concepts they modify and used this same pattern to extract further concepts using the following procedures:

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

2.2 Extracting the Domains

Using the patterns learned at step 2.1 we have identified as many domains as possible for a given pair of synonym and antonym and count the frequency of their co-occurrence in the respective domains. We then clustered words expressing properties of these various domains If the concepts are narrow or too specific to represent the domains. Hence, the concepts expressing a given domain are replaced by a term which is more descriptive of the encoded domain. For instance, winters and summers are replaced by temperature/climate.

In order to cluster the concepts to represent them with the more representative concept, we used word co-occurrences as the clustering feature as follows:

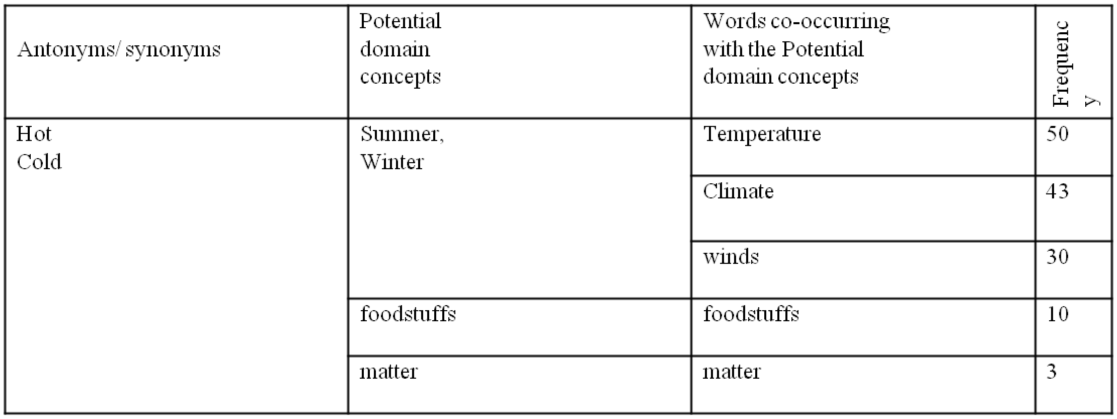
- *Extract other term co-occurrence frequencies within a window of sentences constituting both the antonyms/synonyms and the potential domain concepts*

– *Antonyms: Hot cold , domain concepts: winter summer*

– *Query: : Hot cold winter summer*

- *Count frequency of other term co-occurrences*
- *Create a matrix of the potential domain concepts and the co-occurring terms together with the frequencies*
- *Cluster them using k-means algorithm*

- *Take the term with the maximal frequency (centroid) in each cluster and consider it as the domain term*
- *Test the result employing expert judgment running the algorithm on test set*



Antonyms/ synonyms	Potential domain concepts	Words co-occurring with the Potential domain concepts	Frequenc y
Hot Cold	Summer, Winter	Temperature	50
		Climate	43
		winds	30
	foodstuffs	foodstuffs	10
	matter	matter	3

Table 1: the matrix of the frequencies of terms co-occurring with the potential domain concepts

The algorithm finally calculated the co-occurrence frequency of the antonyms/synonyms with the respective concepts they are referring to (or modify) as presented and produce the result in an excel file.

2.3 Variant Domain Dependent Co-occurrence Extraction

In the previous algorithm, the co-occurrence information is extracted from the same sentence. However, unlike the antonyms, observing synonyms together in the same context (the same sentence and domain) is a rare event. We assumed that using synonym pairs in the same context could create redundancies as they tend to convey related meaning (at least in theory). However, antonym pairs can be used in the same context for expressing contrasting ideas. Hence, we have decided to extract variant domain dependent co-occurrences of the synonyms and antonyms. Variant domain dependent co-occurrence algorithm extracts patterns of co-occurrence information of the synonyms and antonyms in different sentences.

It seems to be more natural to use the substitutes at different times rather than expecting them to appear at the same time. Hence, one can obtain the same information (if the assumption that they are substitute preserve) indirectly by extracting their co-occurrence when they appear separately in different sentences but in the same domain. So, we mined the co-occurrence information of the synonym/antonym in all possible domains and checked if they co-occurred with the same sorts of domains:

  - X(y, frequency)
  - Z(y, frequency)

Where,

> X and Z are antonyms or synonyms and Y being the domains in which the antonyms or synonyms are co-occurring.

The frequency of a pair of the antonym/ synonym in Y domain is counted and the same applies for the other pair. This will help us to measure the degree of co-occurrence of the antonym/synonym pairs from the domain perspective indirectly.

3. Results and Discussions

From the co-occurrences in the same sentence

Based on the result of the experiment the strength of the antonyms/synonyms is varying based on the domains. Hence, the strength of co-occurrence of antonyms and synonyms are dependent (are a function of) on the domains. The language producers showed very consistent way of using the antonyms and the synonyms. For instance, the antonyms: fast slow, quick slow and rapid slow were used in completely different domains with little or no overlap. Fast slow is used in the domains of motion, movement, speed; Quick slow is used for time, march, steps domains.

We have recognized 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: Fast slow: growth, lines , motion, movement, speed ,trains, music, pitch; Quick slow: time, march, steps; Gradual Slow: process, change, transition; Big small: Screen, band; Large Small: Intestine, Companies Businesses; Strong week: Force, Interaction, Team, Ties, Points, Sides, wind

The Synonyms:

- Co-occurred more frequently than most non canonical antonyms in the same sentence but mainly in different domains: Fast quick is more frequent than quick slow, rapid slow
- Few Co-occurrences in same sentences same in the same domains as exhibited by the pairs gradual slow in the domains of process change development
- The strength of the co-occurrence depends on the domains also: Strong Heavy in Wind and Rain domains respectively to express intensity ; large wide in the domains of population and distribution respectively; gradual slow in the domains of process change development.

From the result of variant co-occurrences

The experiment using the variant co-occurrence demonstrated insignificant change in the domains in which the synonyms and antonyms function: 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 have observed that the frequency of co-occurrence significantly changed. For instance, the frequency for the pair gradual Slow was 76 in same sentence experiment and increased to 1436 in the variant co-occurrence experiment.

4. Comparison with related works

The related researches demonstrated that there are antonyms that are strongly opposing (canonical antonyms) (Paradis et al. 2009, Jones et al. 2012). Such antonyms are highly frequent in terms of co-occurrence as compared to other antonyms: large small Vs big small. In this experiment we have observed that the canonical antonyms are the set of antonyms whose domains in which they function are very productive. For instance the number of domains for large small (11704) is by far greater than for big small (120). However this doesn’t make the antonym large small a winner in all the domains. Big Small is the canonical antonym for the domains like Screen as compared to large small. Measuring the strength of antonyms without considering domains in to account provides higher values for the canonicals as they tend to be used in several domains as compared to the non canonicals. If domains are taken in to account, as we did in this experiment, all the antonyms are strong in their specific domains. Large small has higher value without considering domain in to account yet has 0.29 value in the domain of screen where big small has much higher value (0.71).

5. Conclusion

The strength of the antonyms/synonyms is varying based on the domains. The language users showed very consistent way of using the antonyms or the synonyms with little overlaps across the domains. Similar result is observed in both experiments from domain perspective, however, significant differences in frequency. Antonyms frequently co-occurred in the same domains in the same sentence and synonyms co-occurred in different domains in the same sentences (with less frequency) where as synonyms co-occurred more frequently in different sentences in the same domains.

6. Future collaboration

We have planned (up on the availability of funding):

1. To do more experiment using more sets (more number of antonyms and synonyms)
2. Continue the analysis of the results for better understanding and explanation of the nature of the antonyms and synonyms
3. Do the experiment on other languages (like Swedish) and compare across languages
4. To use the patterns learned for the disambiguation of antonyms and synonyms in automatic extractions of the antonyms and synonyms

7. Projected publications

1. Debela Tesfaye and Carita Paradis. “On the use of antonyms and synonyms from a domain perspective” has been accepted for presentation at the 2015 NetWordS Conference as a poster, March 30th - April 1st 2015, Pisa, Italy.

8. Facilities employed and Seminar in the host institution (Lund University)

1. Complete access to 2 labs with multiple high performance computers. We divided the tasks and shared it among the computers to get the results as fast as possible in order to get time to analyze the result during the visit.
2. I am provided with an office with excellent work environment.
3. I gave a talk (on the methods and the results of this experiment) on the seminar organized by the Centre for Languages and Literature, Lund University on January 28, 2015.

9. Reference

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

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

Jones, S. , M.L. Murphy, C. Paradis & C. Willners. 2012. Antonyms in English: Construals, construc-tions and canonicity. Cambridge: Cambridge Uni-versity Press.

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

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

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 Neuro-science 19 (8). 1259–1274.

Tesfaye, D. & Zock, M. 2012. Automatic Extraction of Part-whole Relations. In Proceedings of the 9th International Workshop on Natural Language Processing and Cognitive Science.130-139

Zock, M. & Tesfaye, D. 2012. Automatic index creation to support navigation in lexical graphs en-coding part of relations. Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon (CogALex-III), COLING 2012. 33–52.