ANALYSIS OF LEXICO-SYNTACTIC PATTERNS FOR ANTONYM PAIR EXTRACTION FROM A TURKISH CORPUS

Gürkan Şahin[1], Banu Diri[1] and Tuğba Yıldız[2]

[1]Faculty of Electrical-Electronic, Department of Computer Engineering
Yıdız Technical University, İstanbul, Turkey
{gurkans,banu}@ce.yildiz.edu.tr
[2]Faculty of Engineering and Natural Sciences,
Department of Computer Engineering
İstanbul Bilgi University, İstanbul, Turkey
tdalyan@bilgi.edu.tr

ABSTRACT

Extraction of semantic relations from various sources such as corpus, web pages, dictionary definitions etc. is one of the most important issue in study of Natural Language Processing (NLP). Various methods have been used to extract semantic relation from various sources. Pattern-based approach is one of the most popular method among them. In this study, we propose a model to extract antonym pairs from Turkish corpus automatically. Using a set of seeds, we automatically extract lexico-syntactic patterns (LSPs) for antonym relation from corpus. Reliability score is calculated for each pattern. The most reliable patterns are used to generate new antonym pairs. Study conduct on only adjective-adjective and noun-noun pairs. Noun and adjective target words are used to measure success of method and candidate antonyms are generated using reliable patterns. For each antonym pair consisting of candidate antonym and target word, antonym score is calculated. Pairs that have a certain score are assigned to antonym pair. The proposed method shows good performance with 77.2% average accuracy.

KEYWORDS

Natural Language Processing, Semantic relations, Antonym, Pattern-based approach

1. INTRODUCTION

Extraction of semantic relation pairs from corpus is one of the most popular topic in NLP. Hyponymy, hypernymy, meronymy, holonymy, synonymy, antonymy etc. can be given to example of semantic relations.

Several resources are used to acquire semantic relations. WordNet [1] is the one of the important sources for semantic relations. WordNet is a lexical database for English and consists of so many words and links among these words. Since each word is represented as synonym words called

Jan Zizka et al. (Eds) : ICAITA, SAI, CDKP, Signal, NCO - 2015
pp. 11–17, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.51502

synsets in WordNet, it can be said that main relations of WordNet is synonymy. Apart from synonymy relation, words are connected each other via semantic relation links like hyponymy, hypernymy, meronymy, holonymy, antonymy etc. Words are collected under four different titles as noun, adjective, verb, adverb, respectively in WordNet.

One of the most important semantic relation in WordNet is antonymy. Antonymy represents contrast sense between two words. In fact, there is no exactly consensus on the definition of the antonymy. According to domain experts, some pairs like good-bad, hot-cold etc. represent good antonymy relation, but some pairs like north-south, woman-man do not exactly represent antonymy. This makes difficult to detect opposite pairs. In addition, studies have shown that synonym and antonym words occur with similar context words. This case also reveals difficulty of distinguishing antonyms from synonyms.

In this study, we propose a pattern-based model to extract antonym pairs from Turkish corpus. Only lexico syntactic patterns are used to find antonym pairs. Noun-noun and adjective-adjective antonym initial seeds are prepared and antonym patterns are extracted using seeds. Patterns having a reliable pattern score are selected to generate new antonym pairs from corpus.

The rest of this paper organized as follows: Section 2 presents related works. Extraction of antonym patterns and extraction of new antonym pairs are explained in Section 3 and Section 4, respectively. Finally, we present experimental results in Section 5.

2. RELATED WORKS

Patterns have been widely used to extract semantic relations from corpus. The most popular pattern-based study was made by Hearst [2] in 1992. Hearst used some patterns like “such X as Y” to extract hyponym words from corpus. In this pattern, X and Y represent hypernym and hyponym words, respectively. After experiments, it has been shown that using some patterns, hyponym words can be extracted from corpus with high accuracy.

Various studies have been conducted on extraction of antonym pairs. Lobanova (2010) [3] prepared some adjective-adjective antonym initial pairs and generated antonym patterns occurring with initial pairs from large Dutch corpus. Lobanova used generated antonym patterns to extract new antonym pairs from corpus. This process repeated iteratively. At each iteration, new antonym patterns were generated by using initial pairs and new antonym pairs were used to extract new antonym patterns again. At each iteration, only reliable antonym pairs and patterns were selected. Thus, sharp accuracy decreasing for generated antonym pairs was prevented.

Turney (2008) [4] used a corpus based supervised classification method to separate antonyms from synonyms. Only patterns obtained from corpus were used as features. Co-occurrence frequency between pair and pattern was used as a feature. Support Vector Machines (SVM) was used as classification algorithm. To measure success of method, English as a second language (ESL) questions were used and 75.0% classification accuracy was obtained for antonym pair classification.

Lin (2003) [5] manually prepared some patterns like “from X to Y”, “either X or Y” to discriminate synonyms from antonyms. It was observed that antonym pairs occur with these two pattern very frequently but synonym pairs occur with these patterns rarely.

Mohammad (2008) [6] developed an unsupervised method using degree of antonym to discriminate antonym pairs. According to definition of degree of antonym, the more a pair has antonym degree, the more the pair represents antonymy. Mohammad used corpus statistical features and antonym dictionary category words together. Over test pairs 80.0% accuracy was obtained for antonym pairs.

For Turkish, there are some studies to extract semantic relation pairs from corpus and dictionary definitions [7], [8]. Hyponym-hypernym [9], [10], meronym-holonym [11] and synonym [12] pairs have been automatically extracted from Turkish corpus. For hyponym-hypernym, meronym-holonym and synonym pairs 83.0%, 75.0% and 80.3% accuracies were obtained, respectively.

Although there are some studies about antonym pair extraction from Turkish dictionary definitions, there is no study using Turkish corpus and antonym corpus patterns. Our main motivation is that there is no such a corpus based study for Turkish before.

3. EXTRACTION OF ANTONYM PATTERNS FROM TURKISH CORPUS

LSPs are widely used to extract antonym relation pairs. In this study, antonym patterns are used to extract antonym pairs. Therefore, we have to generate antonym patterns from corpus. To extract Turkish antonym patterns, following processing steps are applied.

- BOUN web corpus was used [13] as a source. The corpus consists of nearly 10 million sentences and 500 million words (tokens). Firstly, we remove all punctuation and special characters from corpus. Corpus is parsed morphologically by Zemberek Turkish NLP tool [14] and each word in corpus is separated to root, root part-of-speech tag and suffixes. For a given word, Zemberek generates multiple parsing results, but only first parsing result is used. Because our corpus is too big, search process can take a long time. For fast search operations, morphologically parsed corpus is indexed by Apache Lucene 4.2.0 searching tool [15] and index file is used for all corpus search operations.

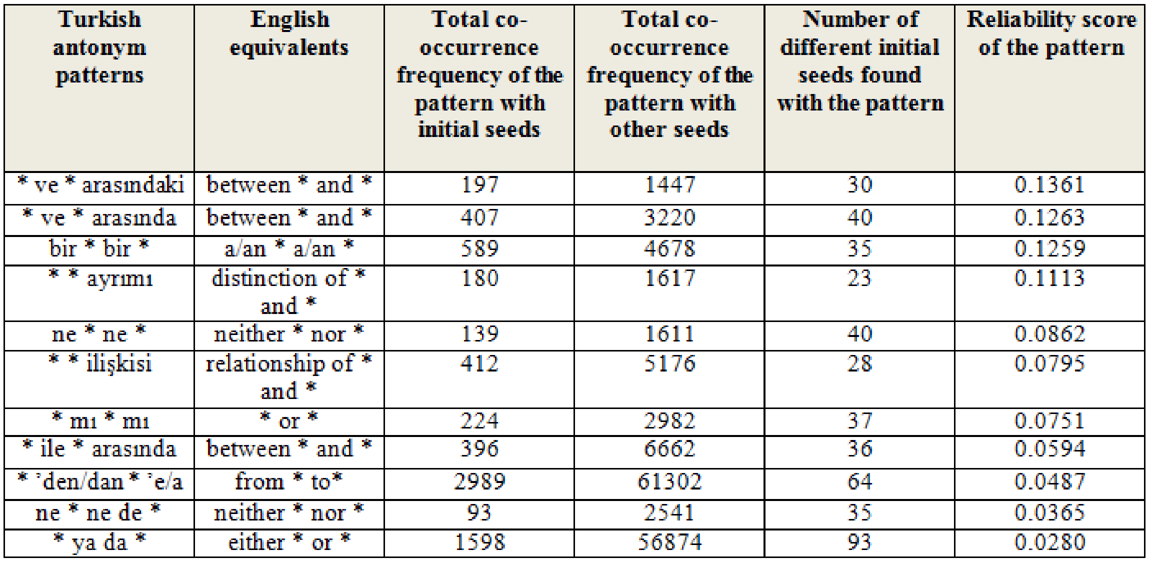
- To find antonym patterns, we generate noun and adjective target words. Antonym equivalents of target antonyms are extracted with using Turkish Antonym Dictionary [16]. 184 antonym pairs called initial seeds are searched in corpus index file and sentences which contain initial seeds are found. In related sentences, initial seeds are replaced with * (wildcard) character. We select patterns having maximum two words between two * characters and others are removed. Thus, we ignore unproductive special patterns.

- Reliability score of each pattern is calculated. To calculate pattern reliability score, we used a formula which is given in equation (1).

$$R_n = \frac{P}{T} \tag{1}$$

In formula, R_n represents reliability score of pattern n. P is total co-occurrence frequency of pattern n with initial seeds. T represents total co-occurrence frequency of pattern n with other antonym pairs(other seeds) in corpus. Total co-occurrence frequency of pattern with initial seeds is divided by total co-occurrence frequency of pattern with other seeds. Then, reliability score is calculated for each pattern. For example, if pattern X occurs with initial seeds 100 times and occurs with other seeds 10.000 times in corpus, reliability score of X equals 100/10.000 = 0.01. But the reliability score may be misleading. If pattern X occurs with initial seeds 7 times and occurs with other seeds 10 times, reliability score equals 7/10 = 0.7. Although reliability score of X is high, X occurs with initial seeds only 7 times. Because co-occurrence frequency of X with initial seeds is too low, pattern X does not have any importance in terms of productivity and generality. For this reason, we calculate reliability score for patterns occurring with initial seeds more than 50 times and other patterns are ignored. To determine pattern reliability score, number of different initial seeds occurring with a pattern is an important parameter. We can say that the more different initial seeds occur with a pattern, the more the pattern is reliable. We assume that pattern X occurs with initial seeds 100 times, but only occurs with 5 different initial seeds. Likewise, pattern Y occurs with initial seeds 100 times, but occurs with 20 different initial seeds. If total co-occurrence frequency of X and Y with other seeds equals 1000, reliability scores of both patterns equal 100/1000 = 0.1. Although pattern reliability scores of X and Y equal each other, Y pattern occurs with more different initial seeds than X. Hence the pattern Y is more general and productive than X. To tackle this problem, pattern reliability score is calculated for patterns occurring with more than 20 different initial seeds and other patterns are not assessed. After calculating reliability score for each pattern according to two conditions given above, all patterns are sorted according to reliability score. Patterns that have reliability score greater than 0.02 are selected to generate new antonym pairs from corpus. Reliable antonym patterns are given in Table 1.

Table 1. Antonym patterns extracted from corpus using initial seeds



Turkish antonym patterns	English equivalents	Total co-occurrence frequency of the pattern with initial seeds	Total co-occurrence frequency of the pattern with other seeds	Number of different initial seeds found with the pattern	Reliability score of the pattern
* ve * arasındaki	between * and *	197	1447	30	0.1361
* ve * arasında	between * and *	407	3220	40	0.1263
bir * bir *	a/an * a/an *	589	4678	35	0.1259
* * ayrımı	distinction of * and *	180	1617	23	0.1113
ne * ne *	neither * nor *	139	1611	40	0.0862
* * ilişkisi	relationship of * and *	412	5176	28	0.0795
* mı * mı	* or *	224	2982	37	0.0751
* ile * arasında	between * and *	396	6662	36	0.0594
* 'den/dan * 'e/a	from * to*	2989	61302	64	0.0487
ne * ne de *	neither * nor *	93	2541	35	0.0365
* ya da *	either * or *	1598	56874	93	0.0280

4. EXTRACTING NEW ANTONYM PAIRS USING PATTERNS

Using antonym patterns in Table 1, antonym equivalent words are extracted for a given target word. Process steps of new antonym pair extraction are given below.

- Firstly, target words are determined and patterns generated by replacing target word with * characters are searched in corpus. Words corresponding to * characters are extracted as candidates of target word. Although reliable patterns are used, not antonym pairs can occur in these patterns. For this reason, antonym equivalents of given a target word are defined as candidates. In pattern structure, any antonym pairs can show two different sequence like X-Y and Y-X. Thus, given a target word is searched in two different positions and words in different * positions are recorded as candidates. For example, target word "iyi" (good) are searched as;

Turkish patterns	**English equivalents**
# iyi ve * arasındaki	# between good and *
# * ve iyi arasındaki	# between * and good
# ne iyi ne de *	# neither good nor *
# ne * ne de iyi	# neither * nor good
…	

- After extracting candidates of target word, antonym score is calculated for each pair consisting of target and a candidate. Pairs having a certain antonym score are assigned to antonym and others are eliminated.

To calculate antonym scores of pairs, we used Lobanova's antonym score formula given in equation (2) [17].

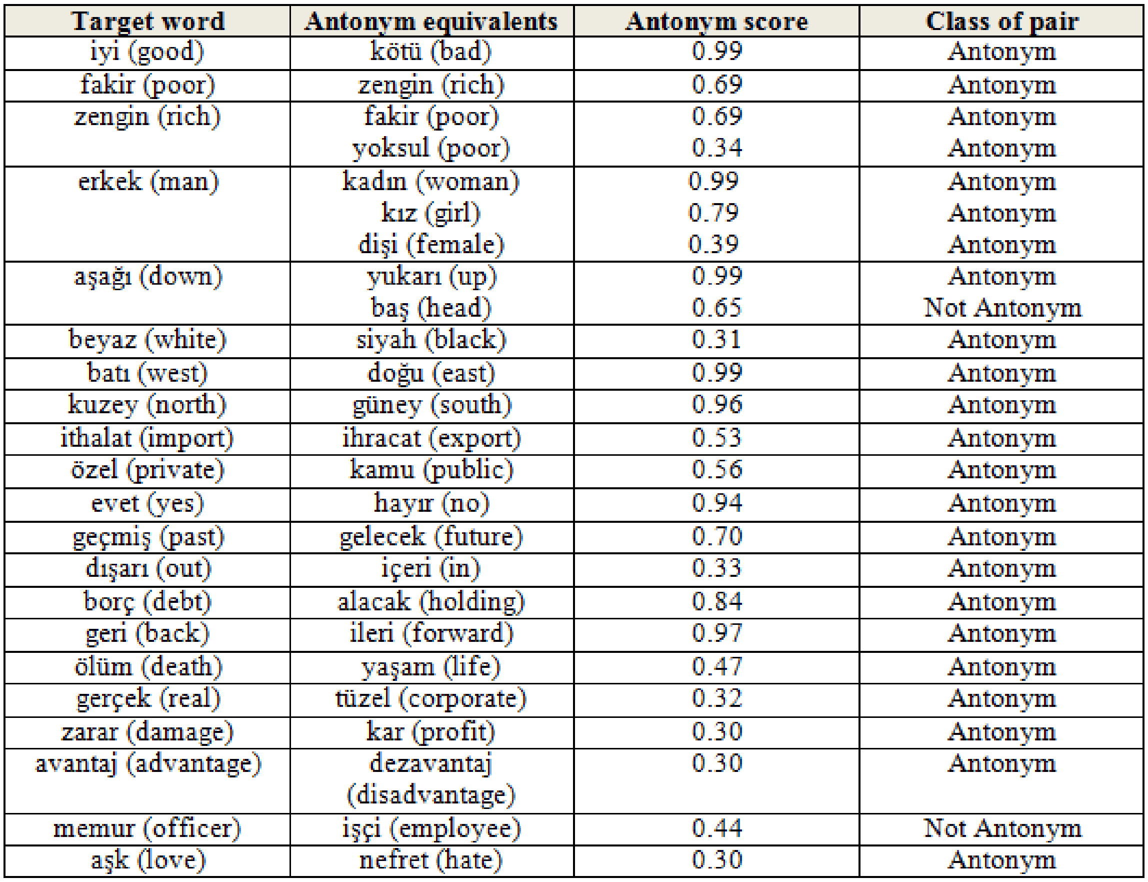
$$P_x = 1 - \prod_{n=1}^{M}\left(1-\frac{Ck}{Tk}\right)^{Ck} \tag{2}$$

In formula, P_x represents antonym score for pair x. M is number of reliable pattern and C_k is co-occurrence frequency of pair x with pattern k. T_k represents co-occurrence frequency of pattern k with other seeds in corpus.

5. Experimental Results

To measure success of model, 196 noun and adjective target words are utilized. Target words were searched together with reliable antonym patterns and candidates were extracted. For each pair, antonym score was calculated. After observations, we defined minimum reliable antonym score as 0.3. When the minimum reliable antonym score is defined less than 0.3, it is shown that accuracy of the method falls sharply. For 45 out 196 target words, our method proposed reliable antonym pairs with 77.2% average accuracy. Class of pairs were manually tagged by 3 Turkish native speakers. 21 target words and candidates, english equivalents are given in Table 2.

Table 2. Target words, candidates, antonym scores and pair classes



Target word	Antonym equivalents	Antonym score	Class of pair
iyi (good)	kötü (bad)	0.99	Antonym
fakir (poor)	zengin (rich)	0.69	Antonym
zengin (rich)	fakir (poor) yoksul (poor)	0.69 0.34	Antonym Antonym
erkek (man)	kadın (woman) kız (girl) dişi (female)	0.99 0.79 0.39	Antonym Antonym Antonym
aşağı (down)	yukarı (up) baş (head)	0.99 0.65	Antonym Not Antonym
beyaz (white)	siyah (black)	0.31	Antonym
batı (west)	doğu (east)	0.99	Antonym
kuzey (north)	güney (south)	0.96	Antonym
ithalat (import)	ihracat (export)	0.53	Antonym
özel (private)	kamu (public)	0.56	Antonym
evet (yes)	hayır (no)	0.94	Antonym
geçmiş (past)	gelecek (future)	0.70	Antonym
dışarı (out)	içeri (in)	0.33	Antonym
borç (debt)	alacak (holding)	0.84	Antonym
geri (back)	ileri (forward)	0.97	Antonym
ölüm (death)	yaşam (life)	0.47	Antonym
gerçek (real)	tüzel (corporate)	0.32	Antonym
zarar (damage)	kar (profit)	0.30	Antonym
avantaj (advantage)	dezavantaj (disadvantage)	0.30	Antonym
memur (officer)	işçi (employee)	0.44	Not Antonym
aşk (love)	nefret (hate)	0.30	Antonym

6. CONCLUSIONS

In this study, antonym pairs and patterns were automatically extracted from Turkish corpus. Noun-noun and adjective-adjective seeds were created and antonym patterns were generated using these seeds. After generating patterns from initial seeds, reliability score was calculated for each antonym pattern. 11 patterns having reliability score greater than 0.02 were selected to produce new antonym pairs. To measure accuracy of method, noun and adjective target words were used as test words. Using these targets with antonym patterns, candidates were found for each target words. For each antonym pair, antonym score was calculated. Pairs having antonym score greater than 0.3 were assigned to antonym and others were eliminated. For 45 out 196 target words, our method proposed reliable antonym pairs with 77.2% average accuracy.

This study has been shown that Turkish antonym relation patterns can be extracted from corpus easily using some manually created antonym seeds. Candidates also can be easily extracted for a given target word with high accuracy with using reliable antonym patterns. Because patterns are used to extract antonym pairs, high co-occurrence frequency of target with patterns in corpus directly influences success of the method. This is a disadvantage for all of pattern-based methods.

In further studies, we aim to use corpus statistical information with patterns. Thus, antonym pairs occurring with patterns at low frequency can be extracted from corpus.

REFERENCES

[1] Christiane Fellbaum (1998, ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

[2] Hearst, M.A.: Automatic Acquisition of Hyponyms from Large Text Corpora. In: 14th International Conference on Computational Linguistics, COLING 1992, Nantes, France, pp. 539-545(1992)

[3] Lobanova, A., van der Kleij, T. and J. Spenader (2010). Defining antonymy: a corpus-based study of opposites by lexico-syntactic patterns. In: International Journal of Lexicography. Vol 23: 19-53.

[4] Turney, P.D. (2008), A uniform approach to analogies, synonyms, antonyms, and associations, Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), Manchester, UK, pp. 905-912.

[5] Dekang Lin, Shaojun Zhao, Lijuan Qin, and Ming Zhou. 2003. Identifying synonyms among distributionally similar words. In Proceedings of IJCAI 2003. Acapulco, Mexico.

[6] Saif Mohammad, Bonnie Dorr, and Graeme Hirst. 2008. Computing word-pair antonymy. In EMNLP, pages 982–991. Association for Computational Linguistics.

[7] Yazıcı, E. ve Amasyalı, M.F., (2011). Automatic Extraction of Semantic Relationships Using Turkish Dictionary Definitions, EMO Bilimsel Dergi, İstanbul.

[8] Zeynep Orhan, İlknur Pehlivan , Volkan Uslan , Pınar Önder,Automated Extraction of Semantic Word Relations in Turkish Lexicon, Journal of Mathematical And Computational Applications, 16, 1, Jan. 2011, pp.13-22,

[9] Yildirim, S., Yildiz, T., (2012). "Corpus-Driven Hyponym Acquisition for Turkish Language", CICLing 13th International Conference on Intelligent Text Processig and Computational Linguistics, 2012.

[10] Şahin, G., Diri, B., Yıldız T., "Pattern and Semantic Similarity Based Automatic Extraction of Hyponym-Hypernym Relation from Turkish Corpus", 23th Signal Processing and Communications Applications Conference, Malatya, Turkey, (16-19 May), 2015.

[11] Yıldız, T., Yıldırım, S., Diri, B., "Extraction of Part-Whole Relations from Turkish Corpora", Computational Linguistics and Intelligent Text Processing, CICLing, Greece, 2013.

[12] Yıldız, T., Yıldırım, S., Diri, B., "An Integrated Approach to Automatic Synonym Detection in Turkish Corpus", 9th International Conference on Natural Language Processing, PolTAL, Springer LNAI proceedings, Warsaw, Poland, (17-19 September), 2014.

[13] Sak, H., Güngör, T. and Saraçlar. M., (2008). "Turkish Language Resources: Morphological Parser, Morphological Disambiguator and Web Corpus", The 6th International Conference on Natural Language Processing, GOTAL 2008.

[14] http://zemberek-web.appspot.com/

[15] http://lucene.apache.org/core/

[16] http://tdk.gov.tr/

[17] Lobanova, A., van der Kleij, T. and J. Spenader (2010). Defining antonymy: a corpus-based study of opposites by lexico-syntactic patterns. In: International Journal of Lexicography. Vol 23: 19-53.