Analysis of Lexico-syntactic Patterns for Turkish Meronym Extraction from Corpus

Tuğba Yıldız, Banu Diri, Savaş Yıldırm

Istanbul Bilgi University
Department of Computer Engineering, Eyüp, 34060 Istanbul, Turkey

Yıldız Technical University
Department of Computer Engineering, Davutpaşa, 34349 Istanbul, Turkey

{tdalyan,savasy}@bilgi.edu.tr

banu@ce.yildiz.edu.tr

Abstract

In this paper, we applied lexico-syntactic patterns to disclose meronymy relation from a huge corpus in Turkish raw text. Once, the system takes a huge raw corpus and extract matched cases for a given pattern, it proposes a list of whole-part pairs depending on their co-occur frequencies. For the purpose, we exploited and compared a list of pattern clusters. The clusters to be examined could fall into three types; general patterns, dictionary-based pattern, and bootstrapped pattern. We evaluated how these patterns improve the system performance especially within corpus-based approach and distributional feature of words. Finally, we discuss all the experiments with a comparison analysis and we showed advantage and disadvantage of the approaches with promising results.

Keywords: meronym, lexico-syntactic patterns, corpus-based approaches

1. Introduction

Meronym has been referred to as a part-whole relation that represents the relationship between a part and its corresponding whole. It is a subject of some disciplines like logic, philosophy, linguistic and cognitive psychology. In many studies, it has been primarily discussed the types of meronym relation, relatedness of meronym relation with other relations and transitivity of meronym relation. One of the most important and well-known study is designed by Winston et al. (1987). They identified part-whole relations as falling into six types: Component-Integral(CI), Member-Collection(MC), Portion-Mass(PM), Stuff-Object(SO), Feature-Activity(FA) and Place-Area(PA).

Recently, there have been many significant studies in automatically extracting meronym relation from a raw text. Some of these methods are based on lexico-syntactic patterns (LSP) that is useful technique especially used in semantic relation extraction. It is the most preferred method due to its simplicity and the success. A set of LSP that indicate hyponymic relations has been applied to unrestricted text by Hearst (1992). Although the same technique was applied to extract meronym relations, it was reported that the efforts concluded without great success.

In computational linguistics, pattern-based approaches have been widely used by other researchers for other semantic relations and various attempts have been made to extend the Hearst patterns. In (Berland and Charniak, 1999), some statistical methods were applied within very large corpus to find parts using Hearst's methods. At the end, five reliable lexical patterns were retrieved using some initial seeds.

A semi-automatic method was presented in (Girju et al., 2003) for learning semantic constraints to detect part-whole relations. The method picks up pairs from WordNet and searches them on text collection: SemCor and LA Times from TREC-9. Sentences containing pairs were extracted and manually inspected to obtain a list of LSP. Training corpus was generated by manually annotated positive and negative examples where the decision tree was applied as learning procedure.

Another attempt is a weakly-supervised algorithm; Espresso (Pantel and Pennacchiotti, 2006) used patterns to find several semantic relations besides meronymic relations. The method automatically detected generic patterns to decide correct and incorrect ones and to filter them with the reliability scores of the patterns and the instances.

In Turkish, recent studies to harvest meronym relations and types of meronym relations for Turkish are based on dictionary definition (TDK) and WikiDictionary (Yazıcı and Amasyalı, 2011; Şerbetçi et al., 2011, Orhan et al. 2011). The other major attempt (Yıldız et al., 2013) modeled a semi-automatically extraction of part-whole relations from a Turkish raw text. The model takes a list of manually prepared seeds to induce syntactic patterns and estimates their reliabilities. It then captures the variations of part-whole candidates from the corpus.

In our study, three different clusters of Turkish patterns are analyzed within a huge corpus. First cluster is based on *general patterns* which are the most widely used in literature. Second one is based on *dictionary patterns* that are extracted from TDK and WikiDictionary. Third one is based on *bootstrapping* of the unambiguous seeds.

The rest of the paper is organized as follows: Section 2 includes the methodology of the study. Analysis of pattern-based approach is introduced in Section 3. Details of challenges and evaluation are explained in Section 4.

2. Methodology

The methodology employed here is to apply the lexico-syntactic patterns to acquire part-whole pairs from a corpus. We evaluate three different clusters of patterns; General Patterns, Dictionary-based Patterns, Bootstrapped Patterns. While general patterns are widely used and well known especially within a huge corpus, the dictionary based patterns are suitable and applicable to dictionary-like resources (TDK, WordNet, Wikipedia, etc.). Although the latter is suitable for the dictionary, we discuss that it can have a capacity to disclose semantic relation even from a corpus. The last approach is to bootstrap patterns using a set of part-whole seeds.

2.1. General patterns

The most precise acquisition methodology earlier applied by Hearst (1992) relies on LSPs. We start with the same idea of using the widely used patterns, General Patterns, to acquire part-whole relations, which are the widely used and well-known patterns from several studies (Winston et al., 1987; Girju et al. 2003; Keet and Artale, 2008). One of these studies is proposed by Winston et al. used frames as "part of", "partly" and "made of" for six different types of meronymic relations. Girju et al. represented that some of patterns always refer to part-whole relation in English text, while most of them are ambiguous. Keet and Artale developed a formal taxonomy, distinguishing transitive mereological (1) part-whole relations from intransitive meronymic (2) ones. All general patterns are listed in Table 1. Although there are also various studies that have used patterns-based approaches, most of them are subsumed by the following patterns.

Patterns	Pattern Specifications
Winston	NPx part of NPy NPx partly NPy NPx made of NPx
Girju	parts of NPy include NPx NPy consist of NPx NPy made of NPx NPx member of NPy One of NPy constituents NPx
Keet	NPx member of NPy (1) NPy constituted of NPx (1) NPx subquantity of NPy (1) NPx participates in NPy (1) NPx involved in NPy (2) NPx located in NPy (2) NPx contained in NPy (2) NPx structural part of

Table 1. Patterns that are used in three different studies

We adopted the all these patterns to Turkish domain. As expected, those patterns which are not suitable and applicable for Turkish language are eliminated. The remaining patterns are evaluated in terms of the capacity and reliability.

To extract prospective sentences that include part-whole relations by using LSPs from a Turkish corpus (Sak et al.,) of 490M tokens, Turkish equivalents of these patterns are constructed in regular expression forms. General patterns, type of patterns, number of cases matched in corpus, number of wholes that matches the pattern, the most frequent wholes are listed in Table 2.

Adoption of the pattern to Turkish domain is difficult due to free word order language with agglutinating word structures. The noun phrases can easily change their position in a sentence without changing the meaning of the sentence. However this replacement can only affect the emphasis. Besides that other part of speech tags can lie between NPs and hence parts can be found away from whole in a sentence. For example, ultraviyole radyasyon (ultraviolet radiation)- güneş enerjisi(solar energy) is part-whole pair in the following sentence.

"Ultraviyole (UV) radyasyon, dünya yüzeyine erişen güneş enerjisinin doğal bir parçasıdır."
(Ultraviolet (UV) radiation is a natural part of the solar energy that access to the Earth's surface.)

Determining a window is crucial for the potential parts. If it keeps too smaller, it might not be even enough to catch parts. However a bigger window leads to many irrelevant NPs extracted with large context and it deteriorates the system. We observed that the window size of 15 allows us to capture more reliable parts and the sentences.

General Patterns	Type	#of Cases	# of Whole	The most frequent wholes
NPx is (a|-) part of NPy	CI, MC, PA	16K	2.4 K	Life Culture Turkey Europe
NPx partly NPy	CI, SO	10	10	-
NPy made of NPx	SO, FA	5K	1K	Shopping Paying Process Trade
parts of NPy include NPx	CI	2	2	-
NPy include NPx	MC, CI	1.5K	770	-
NPx member of NPy	MC	23K	2K	Family Union Group Turkey Commission
One of NPy constituents NPx / NPy constituted of NPx	CI, SO	530	276	-
to have	CI, MC, FA, PA	22K	3.7K	Society World Woman Government Kid

Table 2. A summary for General Patterns

In order to evaluate the approach, we picked up the most frequent wholes for each LSPs. Each whole and its potential parts are ranked according to their frequencies.

To distinguish the distinctiveness, we utilized inverse document frequency (idf) that is obtained by dividing the number of times a part occurs with whole by number of times a part retrieved by all the patterns. We selected first 20 candidates ranked by their scores for evaluation. The proposed parts were manually evaluated by looking at their semantic role.

2.2. Dictionary-based Patterns

The most efficient and reliable way of applying LSP is to extract information from Machine Readable Dictionaries (MRDs). The language of use in dictionary is generally simple, informative, and structured and it highly includes a set of syntactic patterns. Thus, many studies have exploited the dictionary definition recently. For Turkish, the recent studies to harvest meronym relations used dictionary definition (TDK) and WikiDictionary (Yazıcı and Amasyalı, 2011; Şerbetçi et al., 2011, Orhan et al. 2011). In (Şerbetçi et al., 2011), semantic relations are extracted to build semantic network. Another study (Yazıcı and Amasyalı, 2011) presents different automatic methods to extract semantic relationships between concepts using two Turkish dictionaries. They efficiently used regular expressions to extract part-whole relation.

We examined all these findings and provided a summary report for relations, type of patterns, number of cases in corpus, number of wholes that matches the pattern, the most frequent as shown in Table 3.

Relations	Type	#of Cases	# of Whole	The most frequent wholes
Group-of (whole/group/all/set/ flock/union of)	MC	140	111	-
Member-of (class/member/setof) (from the family of Y)	MC	207 192	159 56	- -
Amount-of: (amount/measure/unit-of)	PM	91	81	-
Has-a (Y has the suffix of ‘l(H))	CI, MC, FA, PA	58K	16K	Kid Woman Football Job
Consist-of	CI, MC	12K	2760	Album Collection Exhibition Team
Made-of	SO	6K	1766	Export Payment Application Receiving

Table 3. A summary for dictionary-based patterns

Member-of, made-of and consist-of can be confused with the ones in the general patterns whereas pattern specifications are different from each other. All patterns are applied to Turkish corpus as same as general patterns and a similar process is carried out. Even though these patterns are usefulness especially in dictionary, they could return redundant and incorrect results for Turkish.

2.3. Bootstrapped Patterns

Methodology of bootstrapped patterns is different from that of others described above. The bootstrapped pattern-based approach proposed here is implemented in two phases: Pattern identification and part-whole pair detection. For the pattern identification, we begin by manually preparing a set of unambiguous seed pairs that definitely convey a part-whole relation. For instance, the pair (engine, car) would be member of that set. The seed set is further divided into two subsets: an extraction set and an assessment set. Each pair in the extraction set is used as query for retrieving sentences containing that pair. Then we generalize many LSPs by replacing part and whole token with a wildcard or any meta-character.

The second set, the assessment set, is then used to compute the usefulness or reliability scores of all the generalized patterns. Those patterns whose reliability scores, rel(p), are very low are eliminated. The remaining patterns are kept, along with their reliability scores. A classic way to estimate rel(p) of an extraction pattern is to measure how it correctly identifies the parts of a given whole. The success rate is obtained by dividing the number of correctly extracted pairs by the number of all extracted pairs. The outcome of entire phase is a list of reliable LSP along with their reliability scores.

In order to run second phase, the previously generated patterns are applied to an extraction source that is a Turkish raw text. The instantiated instances (part-whole pairs) are assessed and ranked according to their reliability scores. We experiment with three different measures of association (pmi, dice, t-score) to evaluate their performance in scoring function. We also utilized idf to cover more specific parts. The motivation for use of idf is to differentiate distinctive features from common ones. All formulas, results have been already reported in other study (Yıldız et al., 2013).

Based on reliability scores, we decided to filter out some generated patterns and finally obtained six different significant patterns. The list of the patterns and their examples can be found in Table 4.

Patterns	Examples
NP+gen NP+pos	door of the house /*evin kapısı*
NP+nom NP+pos	House door /*ev kapısı*
NP+Gen (N—ADJ)+ NP+Pos	back garden gate of the house *Evin arka bahçe kapısı*
NP of one-of NPs	the door of one of the houses *Evlerden birinin kapısı*
NP whose NP	The house whose door is locked *Kapısı kilitli olan ev*
NP with NPs	the house with garden and pool *bahçeli ve havuzlu ev*

Table 4. Bootstrapped Patterns and examples

All patterns are evaluated according to their usefulness. We roughly order the pattern as P1, P2, P3, P6, P4, and P5 by their normalized average scores. P1, which is the genitive one, is the most reliable pattern with respect to all measures.

Measures	rel(p1)	rel(p2)	rel(p3)	rel(p4)	rel(p5)	rel(6)
pmi	1.58	1.53	0.45	0.04	0.07	0.57
dice	0.01	0.003	0.01	0.004	0.001	0.003
tscore	0.11	0.12	0.022	0.0004	0.001	0.03

Table 5. Reliability of Patterns

For the evaluation phase, we manually and randomly selected five whole words: book, computer, ship, gun and building. For a better evaluation, we selected first 10, 20 and 30 candidates ranked by the association measure defined above. However the results based on first 20 candidates will be used to fairly compare performance with other clusters of patterns.

3. Challenges

The very basic problem of natural language processing is sense ambiguity. Almost all studies suffer from the ambiguity problem. For a given whole, proposed parts could be incorrect due to polysemous words. Girju et al. (2006) represented that some of patterns always refer to part-whole relation in English text, while most of them are ambiguous. Their listings of unambiguous and ambiguous patterns are given in Table 6. Part-of pattern, genitive construction, the verb "to have", noun compounds and prepositional construction are classified as ambiguous meronymic expressions.

Unambiguous	Ambiguous
parts of NPy include Npx	NPx part of NPy
NPy consist of Npx	NPy has NPx
NPy made of Npx	NPy's NPx
NPx member of NPy	NPx of NPy
One of NPy constituents NPx	NPy NPx
	NPy with NPx

Table 6: Ambiguous and Unambiguous pattern list

For Turkish domain, we could not easily do such classification and find even one unambiguous pattern to extract part-whole relation. Additional methods are needed to cope with the problem and to find more accurate results from extracted pairs.

Another problem is that the patterns can also encode other semantic relations such as hyponymy, relatedness, cause etc. Although use of genitive case is popular for detecting part-whole relations, the characteristic of the genitive is ambiguous. The morphological feature of genitive is a good indicator to disclose a semantic relation between a head and its modifier. In this case, we found that the genitive has a good indicative capacity, although it can encode various semantic interpretations. Taking the example, "Ali's team", and the first interpretation could be that the team belongs to Ali, the second interpretation is that Ali's favorite team or the team he supports. It refers such relations "Ali's pencil/Possession", "Ali's father/Kindship", "Ali's handsomeness/Attribute". Same difficulties are valid for other patterns. To overcome the problem, researchers have done many studies based on statistical evidence.

Even the best patterns could not be safe enough all the time. The sentence "door is a part of car" strongly represents part-whole relation, whereas "he is part of the game" gives only ambiguous relation. The word "part-of" has nine different meanings in Turkish Dictionary. It means that it is nine times more difficult to disclose the relation.

Some expressions can be more informal than written language or grammar. Indeed, in any language, different kinds of expression can be appropriate in many different situations. From the formal to the informal, the written to the spoken, from jargon to slang, all type of expressions are a part of corpus. This variety can cause another bottleneck for applying regular expression or patterns.

4. Evaluation & Analysis

Three clusters of pattern were taken into consideration. The first two patterns, *general patterns* and *dictionary patterns* are predefined list that are obtained by literature and other studies. On the other hand, third cluster of patterns, *bootstrapped patterns* are semi-automatically obtained by giving initial unambiguous part-whole pairs. The main problem of first two clusters is limitedness. We could not execute these patterns for any arbitrary whole. Instead, the most frequent wholes occurred in these patterns were evaluated. Looking at the Table 2 and Table 3, each pattern has its own list of potential wholes.

However, thanks to the simplicity, bootstrapped patterns are so broader that for an arbitrary whole, the system can propose a list of parts. Especially genitive case pattern has enormous capacity and it can produce reliable results.

The tendency of the all patterns is to capture mostly semantic relatedness especially when two words or concepts are associated in some way. How could the relation between train and the rail be classified? Thus, both evaluation and error analysis for the system improvement get harder due to that problem.

The clearest observation is that applying dictionary based pattern to a corpus rather than a dictionary is quite limited. For instance, the pattern "amount-of" obtains 91 cases and it consists of 81 different wholes. Each whole has only 1.1 cases matched on average. The "*has-a*" (lI) pattern, one of dictionary-based patterns, is the most productive pattern. It captures over 500K cases. However, it also suffers from the same typical problems mentioned before. The most reliable pattern from the dictionary cluster is *consist-of*. The case capacity is 12K, average number of cases of each whole is nearly 6 and its average success ratio is 80%.

However, general patterns are more productive. On average, for each whole, about 10 cases can be matched. For this group of patterns, the most reliable pattern is to have ("vardır"). The size of matched cases is 22K, average number of cases is 7, and overall success score is about 75 %.

Even though the first two clusters have a promising result, they have limited capacity. Any system relying on these patterns can just give limited number of part-whole

pairs. On contrary, third cluster of pattern which based on bootstrapping methodologies can produce more than the other. The system with this approach could work for any given whole. When looking at the success rate of the bootstrapped techniques, its general average is %67. We conducted another experiment to distinguish distinctive parts from general ones. Excluding general parts from the expected list, we re-evaluated the result of the experiments. When idf is applied, measures are increased by 4.3% on average as expected.

5. Conclusion

Applying lexico-syntactic patterns to disclose meronymy relation from a huge corpus is very naïve and effective way. We employed the same idea for Turkish language domain. Once, the system takes a huge text and morphologically extracts matched cases for a given pattern, it proposes and ranks a list of parts depending on frequencies and some other statistics. Three different clusters of patterns were taken into consideration to acquire meronymy relations. While first two clusters, *general patterns* and *dictionary based patterns*, are pre-defined, the last cluster consists of those patterns that are iteratively bootstrapped with a small set of unambiguous seeds. All these bootstrapped patterns are weighted by a reliability scores which are calculated with a special function.

Although general patterns are more productive and broader than dictionary ones, both share the similar performance in precision when only looking their limited results. Thus, general pattern has better success in terms of recall. The best score among dictionary based methods is one with success rate of %80. For the general pattern, the best has the score of %75 in precision. The problem for these two patterns is their limitedness of production. Third cluster, bootstrapped patterns, is much broader than the others. It can give response for any arbitrary *whole* thanks to its simplicity and its learning procedure. It also gives successful result when compare to other approaches especially in terms of recall.

The core challenge that we faced during the experiment is ambiguity and polysemous word. Another problem is use of language. Different type of expressions such as formal, informal, written or spoken is the main challenge to apply pattern matching or other string matching-based methodologies. They are among the future study plan to be completed.

References

Winston, M.E., Chaffin, R., Herrmann D. (1987). A Taxonomy of Part-Whole Relations. In: *Cognitive Science*, 11 (4), pp. 417-444.

Hearst, M.A. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. *14th Conference on Computational Linguistics (COLING '92).* Volume 2, pp. 539–545.

Berland, M and Charniak, E. (1999). *Finding parts in very large corpora. 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* (ACL '99). pp. 57-64.

Girju, R., Badulescu, A., Moldovan, D.I. (2003). *Learning Semantic Constraints for the Automatic Discovery of Part-Whole Relations. In:Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2003)*, pp.1–8.

Pantel, P., Pennacchiotti, M. (2006). Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations. *21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pp. 113-120.

Yazıcı, E., Amasyalı, M.F. (2011). Automatic Extraction of Semantic Relationships Using Turkish Dictionary Definitions. *EMO Bilimsel Dergi*, 1, 1, pp. 1-13.

Serbetçi, A., Orhan, Z., Pehlivan, I. (2011). Extraction of Semantic Word Relations in Turkish from Dictionary Definitions. *ACL 2011 Workshop on Relational Models of Semantics (RELMS 2011)*, pp. 11–18.

Orhan, Z., Pehlivan, I., Uslan, V., Onder, P. (2011). Automated Extraction of Semantic Word Relations in Turkish Lexicon. In: *Mathematical and Computational Applications*, 16 (1). (2011), pp. 13–22.

Yıldız, T., Yıldırım, S. and Diri, B. (2013). Extraction of part-whole relations from Turkish corpora. *14th international conference on Computational Linguistics and Intelligent Text Processing* (CICLing'13), Samos, Greece, LNCS, vol.7856, pp.126—138. Springer, Heidelberg.

Keet, C.M., Artale, A. (2008). Representing and reasoning over a taxonomy of part-whole relations. In: *Applied Ontology*, 3(1-2). (2008), pp. 91-110.

Girju, R., Badulescu, A., Moldovan, D.I. (2006). Automatic Discovery of Part-Whole Relations. In: *Computational Linguistics*, 32(1). pp. 83-135

Sak, H., Güngör, T., Saraçlar, M. (2008). Turkish Language Resources: Morphological Parser , Morphological Disambiguator and Web Corpus. In: *Ranta, A., Nordström, B. (eds.) GoTAL 2008*. LNCS, vol. 5221, pp. 417–427. Springer, Heidelberg.