Extraction of Semantic Word Relations in Turkish from Dictionary Definitions

Şerbetçi Ayşe
Computer Engineering Department
Fatih University
34500
Buyukcekmece, Istanbul, Turkey
aserbetci@fatih.edu.tr

Orhan Zeynep
Computer Engineering Department
Fatih University
34500
Buyukcekmece, Istanbul, Turkey
zorhan@fatih.edu.tr

Pehlivan İlknur

firstnoor@gmail.com

Abstract

Many recent studies have been dedicated to the extraction of semantic connections between words. Using such information at semantic level is likely to improve the performance of Natural Language Processing (NLP) systems, such as text categorization, question answering, information extraction, etc. The scarcity of such resources in Turkish, obstructs new improvements. There are many examples of semantic networks for English and other widely-used languages to lead the way for studies in Turkish. In this study, developing a semantic network for Turkish is aimed by using structural and string patterns in a dictionary. The results are promising, so that approximately two relations can be extracted from 3 definitions. The overall accuracy is 86% if we consider the correct sense assignment, 94% without considering word sense disambiguation.

1 Introduction

Nowadays, the internet is the primary media, people use for communicating with each other and sharing their ideas with the rest of the world. Therefore, a massive amount of data is available but it is not understandable to computers. Wide usage of the web brings some requirements to make this data more beneficial for people. Understanding text from a foreign language or accessing relevant ones among millions of documents has become crucially important. However, due to the large size of data, it is very difficult for human to maintain these tasks without rapid computer processing. Automatic text summarization, information extraction and text categorization are all important NLP areas, which aim to help humans benefit from computer systems to perform these tasks.

The process of obtaining robust computer systems capable of handling these tasks involves supporting machines with semantic knowledge. The type of necessary knowledge depends on the target system. Nevertheless, the information of what kinds of relations exist between the words can be very useful for many purposes especially for NLP applications. Starting with the WordNet project in 1985, semantic networks or lexical databases have been among the important study areas in NLP up to the present. WordNet project (http://wordnet.princeton.edu/wordnet/download/).

Obtaining a semantic network for Turkish language is the goal of this study. Since this study is an initial step of developing a semantic network in Turkish, basic relationship of hyponymy and synonymy are primarily handled. For this purpose, the investigation of dictionary definitions and the morphological richness of Turkish language are utilized. Different types of relationships are shown in Table 1. Since these relationships are very basic, they are likely to be used in various kinds of NLP

Proceedings of the ACL 2011 Workshop on Relational Models of Semantics (RELMS 2011), pages 11–18,
Portland, Oregon, USA, June 23, 2011. ©2011 Association for Computational Linguistics

tasks.

Various patterns are extracted from dictionary by using both syntax and string features of the definitions. Each definition represents particular sense of a word, so they can be considered as different words. For more accurate semantic analysis, the connection between words should be established between appropriate senses of the words. To be more concrete, an example can be given on the semantically ambiguous word as; yüz 'face' or 'hundred'. When a has-a relation is detected between the words vücut 'body' and yüz, the appropriate sense for yüz should be selected as 'face', instead of 'hundred'.

Relationship	Example
Is-a(hyponymy)	flower-plant
Synonym-of	initial-first
Antonym-of	quick-slow
Member-of	academician-academy
Amount-of	kg-weight
Group-of	forest-tree
Has-a	office-computer

Table 1: Basic word relationships

The rest of the paper is organized as follows: Section 2 discusses the previous work in this field. Section 3 explains the implementation methods, details and approaches to some NLP problems, like morphology or word sense ambiguity. This section also gives some statistics about the results. The future work to be performed for both improving and extending the network is also discussed in this section. Section 4 evaluates the overall system.

2 Previous Work

Cyc (http://www.opencyc.org) project is one of the first attempts of obtaining computer accessible world knowledge. Many other studies have been performed for constructing large lexical databases or semantic networks by extracting the semantic connections between words.

In fact, both the number and types of the possible relationships are not clearly identified in this area. However, there are some widely accepted basic relationships, which can be considered as the backbone of semantic networks. No matter which method is followed for extracting these connections, most of the studies including WordNet (Miller, 1995; Fellbaum, 1998) and ConceptNet(Havasi et al., 2007) are based on this set of specific relationships such as hyponymy, synonymy, meronymy etc. These are the most basic but also the most informative ones among the common relation types.

Some manual work has been performed at the beginning for constructing this kind of semantic networks, including but not limited with Wordnet. Nowadays, however, semi or fully automatic systems capable of performing these processes are worked on. Different methods have been used from collecting online data to corpus analysis and from defining syntactical rules to string patterns.

ConceptNet collects its data from Open Mind Common Sense Project (http://commons.media.mit.edu/en/), which is a web-based collaboration (Havasi et al., 2007). Over 15,000 authors enter sentences to contribute to the project. Users can answer questions via the web interface, which aim to fill the gaps in the project. However, in the study of Nakov and Hearts (2008), the whole web is treated like a corpus and the occurrences of the noun pairs together are converted into feature vectors to perform a classification for semantic relations.

There are various methods under the subject of string or structural patterns that represent specific semantic relations. Barriere(1997) investigates some syntactical rules in her study and matches the dictionary definitions to these rules for figuring out the relations. Also, in some languages in which prepositions are used frequently, some relations can be extracted depending on the prepositions, like in the study of Celli and Nessim (2009).

In addition, there are some studies which aim to extract some patterns for each relation for the purpose of finding new instances.

Turney's study (2006) is a good example, which uses a corpus based method for finding high quality patterns. It searches the noun pairs through the corpus to extract some row patterns. The patterns are ranked by a ranking algorithm in order to determine the most qualified patterns for the further steps. Espresso (Pantel and Pennacchiotti, 2006) is also concerned in finding patterns to represent relations. It starts with a few reliable seed of relations and iteratively learns the surface patterns in a given corpus.

There is a lot of work to be done for Turkish in this area. Except one project (Bilgin et al., 2004), which was performed and limited within the scope of BalkaNet project, there is no significant work in this area for Turkish.

BalkaNet project is valuable in the sense of being one of the first attempts for developing Turkish Wordnet. It differs from our study in its methodology, which involves translation of basic concepts in EuroWordNet and then using some string patterns to extend the network. In addition, target relationships and obtained results are quite different and will be handled in the following sections.

Another work (Önder, 2009) which aimed to extract the relations from dictionary definitions by using string patterns but was not completed, constructs the basics of our study.

3 Experimental Setup

In this section, the implementation process is discussed in the following order of sub topics:

- Data
- Morphological features of Turkish
- Extracted patterns
- Morphological analysis and disambiguation
- Word sense disambiguation
- Stop word removal
- Results

Using a dictionary can ease the process of extracting semantic relations in a language in many aspects. First of all, every word occurs in the dictionary at least once, hence the probability of missing a word decreases. Secondly, it consists of definitions of the words, which are relatively informative. Lastly, the sentences in a dictionary are generally simple and similar to each other. Therefore, they generally follow a set of syntactic patterns. This enables to perform easy detection of relations.

For all the reasons listed above, a dictionary of Turkish Language Association (TLA) is used in this study. There are 63110 words and 88268 senses in this dictionary. This concludes that nearly 25000 of the words are ambiguous. In Table 2, the distributions of these words among the most frequent parts of speech are given.

The first step is investigating the dictionary definitions manually in order to explore some patterns which are likely to keep a particular semantic relation inside. The patterns should be general enough for obtaining a reasonable recall. In addition, they should be specific enough not to cause low precision. After a rough analysis, the dictionary is scanned for some row patterns to evaluate the results in terms of both accuracy and comprehensiveness. According to the results, either patterns are reorganized or some additional features are determined to be used for increasing the number of matches and decreasing the error rate. Different kinds of features in the dictionary definitions and the words being explained are used. Morphological structures, noun clauses, clue words and the order of the words in the sentence are the examples of these features.

Part of Speech	Number
Noun	56400
Adjective	14554
Adverb	3011
Pronoun	104
Verb	11408

Table 2: The distributions of words in TLA dictionary

Turkish is an agglutinative language which results in a rich but rather complex morphological structure. Thus, the words do keep a very important part of the sense. They can be converted from one part of speech into another by adding derivational suffixes. For example, from the verb gelmek ‘to come’ the adjective gelen ‘the one who comes’ can be derived. This feature of Turkish constructs the most important effect of increasing the number of matches between patterns and definitions. In addition, indefinite noun phrases are detected with the help of morphological analysis and lots of relations are extracted as a result. These are only a few examples of where morphology is used when extracting the relations.

Some clue words in the definitions are also searched for. In dictionaries, some similar words are explained by using the same words and they can represent some specific relations. To be more concrete, the adjectives that represent the opposite of another adjective can be considered. These types of words are usually defined by using the words olmayan ‘not’ and karşıtı ‘opposite of’. For example, in the definition of the word fantasik ‘fantastic’ there exists the phrase gerçek olmayan 'not real'. An antonymy relation can be established between the word fantastic 'fantastic' and gerçek 'real' as a result. For some other types of relations, different words are detected and handled. For example, for member-of relation, sınıfından 'from the class of'; for is-a relation, türü 'type of' are selected.

Additionally, noun clauses, which are defined in the dictionary, are investigated. Most of the time a noun phrase represents an 'is-a' relation. The word balık 'fish' and kılıç balığı 'sword fish' are both in the dictionary and kılıç balığı 'sword fish' is a noun phrase that has balık 'fish' in it. It is obvious that there is a connection between the words kılıç balığı 'sword fish' and balık 'fish'.

Various patterns are obtained by using at least one of the above features. The obtained patterns for each type of relation are shown in Table 3. When analyzing the table, the representatives to be considered are as follows: X and Y are used for representing the words being connected to each other, *punc* represents one of the specified punctuations like comma or full stop, w* represents zero or more sequential words, $w^*_{no\_punct}$ represents zero or more sequential words without any punctuation inside, w_x is a word which keeps a specific part of speech x, depending on the pattern.

The extracted relations for the provided word definitions are not limited with those mentioned in the table. If possible, two or more relations can be extracted from a single definition. For instance, besides the 'member-of' relation between çakal 'jackal' and etoburlar 'carnivora', a 'kind-of' relation is extracted also for çakal 'jackal' with hayvan 'animal', since the definition matches with the fourth pattern of 'kind-of' relation. Although only the relation between pinhan 'latent' and saklı 'hidden' is given, another synonymy relation is also obtained from this pattern between pinhan 'latent' and gizli 'ulterior'.

The morphological structures of the words are obtained by using Zemberek project (http://code.google.com/p/zemberek), which is an open source morphological analyzer for Turkish. The analysis result of the word atan 'be assigned' or 'your ancestor' or 'the one who throws' is displayed with Figure 1.

The morphological ambiguity is handled with two different methods. Firstly, as a pre-processing step, some suffixes are determined, which cannot occur in the dictionary, such as time suffixes. The analyses are pruned from those results that include one or more of these suffixes. Secondly, according to the pattern requirements, the convenient result is selected as the correct one. For example, if a word is required to have a particular chain of suffixes, the first result providing this necessity is selected. If there is no assumption, the first result is selected by default.

The relations are established between the exact senses of the words in order to obtain a reliable network. Therefore, word sense disambiguation should also be performed. One of the words is not ambiguous, since one of its particular senses (definition) is already being handled for most of the relations. On the other hand, for the purpose of determining the correct sense of the remaining word, simplified Lesk algorithm is used(Lesk, 1986). Simplified Lesk algorithm benefits from the similarity measurements between each sense of the ambiguous word and the concept. The algorithm is given in Figure 2 and the details are provided in the http://en.wikipedia.org/wiki/Lesk_algorithm.

In order to obtain more accurate results, stemming and stop word removal is applied for both relation extraction and word sense disambiguation. A connection can be established only if both of the words are not stop words. Stop words are dictionary specific and obtained by counting the occurrences of word stems in the dictionary. Not all frequent stems are assumed to be stop words but the useless ones among the all stems whose occurrences are above an upper limit are ignored. There are 22 stop words specified, including için 'for', başka 'another' and en 'the most'.

The system was evaluated by manual calculation of the accuracy. Equal number of samples is chosen randomly from each pattern. Two types of accuracy were obtained, which are with and without consideration of correct sense assignment.

The obtained results are given in Table 4. The first accuracy column represents the accuracy percentage by considering whether the correct sense could be matched or not. The second column ignores the senses and evaluates the results in terms of the correct word relation only.

Relation	P no	Pattern specification	Example
Hyponymy	1	X: (w*) (w_{adj}) (w*) Y *punc* (w*). where X is noun, Y is a noun root. (X-Y)	göl: Önceden denizken kurumalar, çekilmeler yüzünden göl durumuna gelmiş yer.**(göl-yer)** *lake: a piece of land, previously existing as sea and becoming dry due to droughts, turns into a small body of water**(lake-land)***
	2	X: (w*) (w_{adv}) ($w^*_{no\_punct}$) (Y) *punc* (w*) where X is verb, w_{adv} is a derived adverb, Y is a verb. (X-Y)	hicvetmek: Alay yoluyla yermek.(**hicvetmek-yermek**) *satirize: To criticize by mocking**(satirize-criticize)***
	3	X Y : w*. where X and Y is an indefinite noun phrase (X Y-Y)	ada çayı: Bu bitkiden yapılan sıcak içecek.**(ada çayı-çay)** *sage tea: The tea that is made of this plant**(sage tea-tea)***
	4	X: w* w_{noun} Y *punc* w*. where w_{noun} and Y compose a noun phrase. (X-Y)	post : Tüylü hayvan derisi. **(post-deri)** *fur : Hairy animal skin.**(fur-skin)***
	5	X : w* Y türü*(kind of)* | tipi*(type of)* | çeşidi*(sort of)*. where X and Y nouns (X-Y)	limuzin: İçinde her türlü donanım bulunan lüks, uzun ve geniş otomobil türü.**(limuzin-otomobil)** *limousine: The type of long, wide and luxury automobile in which there exist various equipment**(limousine- automobile)***
Synonymy	1	X : w* *punc* Y where X and Y are nouns, adverbs, or adjectives (X-Y)	pinhan: Gizli, saklı, gizlenmiş.**(pinhan-saklı)** *latent: Ulterior, hidden, covert. (latent-hidden)*
	2	Z : w* *punc* X, Y *punc* w* where X, Y have equal chain of suffixes and they are verbs, adjectives or nouns (X-Y)	razı: Uygun bulan, benimseyen, isteyen, kabul eden **(benimsemek-istemek)** *willing : The one who approves, embraces, wants, agrees on sth.**(embrace-want)***
Group-of	1	X: w* Y bütünü*(whole of)* | topluluğu*(group of)* | tümü*(all of)* | kümesi*(set of)* | sürüsü*(flock of)* | birliği*(union of)* w* where X and Y are nouns. (X-Y)	âlem: Hayvan veya bitkilerin bütünü.**(alem - bitki)** *kingdom : The whole of plants or animals.**(kingdom-plant)***
Antonym	1	X: w* Y olmayan*(not)* | karşıtı*(the opposite of)*. where X and Y are nouns or adjectives. (X-Y)	acı: Bazı maddelerin dilde bıraktığı yakıcı duyu, tatlı karşıtı. **(acı-tatlı)** *bitter: The feeling of pain which some matters leave on tongue, the opposite of sweet. **(bitter-sweet)***
Member-of	1	X: w* Y sınıfı*(class of)* | üyesi*(member of)* | takımı*(set of)*. where X and Y are nouns (X-Y)	senatör: Senato üyesi.**(senatör-senato)** *senator: Member of senate.(senator-senate)*
	2	X : Ygillerden*(from the family of Y)* | Ylerden*(from the family of Y)* w*. where X and Y nouns. (X-Y)	çakal: Etoburlardan, sürü hâlinde yaşayan, kurttan küçük bir yaban hayvanı.**(çakal-etobur)** *jackal: From carnivora, a kind of wild animal smaller than wolf, which lives in flocks.**(jackal-carnivora)***
Amount-of	1	X: w* Y miktarı*(amount-of)* | ölçüsü*(measure-of)* | birimi*(unit-of)* . where X and Y are nouns (X-Y)	amper: Elektrik akımında şiddet birimi.**(amper-şiddet)** *amper: The unit of intensity in electrical current.**(amper- intensity)***
Has-a	1	X: w* Y [w_{noun}] *punc* w*. where Y has the suffix of 'LI', X and Y are nouns (X-Y)	sof : Bir çeşit sertçe, ince yünlü kumaş. **(sof,yün)** *alpaca : A kind of hard, thin, wooled cloth. **(alpaca, wool)***

Table 3: The obtained patterns for each type of relation

1. {Icerik:atan Kok:ata tip:FIIL} Ekler:FIIL_KOK+FIIL_EDILGENSESLI_N

   {Content : be assigned Root : assign Pos: Verb} Suffixes : Verb Root + Passive

2. {Icerik:atan Kok:ata tip:ISIM} Ekler:ISIM_KOK+ISIM_SAHIPLIK_SEN_IN

   {Content : your ancestor Root : ancestor Pos: Noun} Suffixes : Noun Root + Possesive_you

3. {Icerik:atan Kok:at tip:FIIL} Ekler:FIIL_KOK+ FIIL_DONUSUM_EN

   {Content : the one throws Root : throw Pos: Verb} Suffixes : Verb Root + Participle

Figure 1: The morphological analysis result of the word atan*(be assigned | your ancestor | the one who throws)*

```
function SIMPLIFIED LESK(word,sentence) returns best sense of word
        best-sense <- most frequent sense for word
        max-overlap <- 0
        context <- set of words in sentence
        for each sense in senses of word do
           signature <- set of words in the gloss and examples of sense
           overlap <- COMPUTEOVERLAP (signature,context)
           if overlap > max-overlap then
              max-overlap <- overlap
              best-sense <- sense
end return (best-sense)
```

Figure 2: Simplified Lesk algorithm

Relation	Pattern	Number of Relations	Accuracy %	Accuracy(ambiguous) %
Hyponymy	1	20566	84	94
	2	1448	84	89
	3	5127	84	90
	4	3502	74	95
	5	387	90	96
Synonymy	1	2313	76	88
	2	22518	96	100
Group-of	1	435	87	97
Antonym	1	380	99	100
Member-of	1	128	92	97
	2	634	100	100
Amount-of	1	119	81	92
Has a	1	2430	82	89
Total		59987	86,85	94,38
NET		58125		

Table 4: The number of relations and the accuracy results for each relation and each pattern rule

It should be considered that the number of relations extracted per pattern is counted individually in order to show the performance of each pattern separately. Some of the relations can be extracted by different patterns of that relation type, so the net total, which is cleaned from the repetitions, is less than overall total.

The results are promising in terms of both the comprehensiveness and the accuracy. If some more effort can be spent on word sense disambiguation, the accuracy may rise to a considerable ratio. The comprehensiveness is intended to be increased with further work, which is discussed in the following section.

The numbers of relation instances are quite greater when compared to BalkaNet project. There are nearly 34,000 relation instances in the project, including the synonym relations among synset members. In this study 58,000 relations are available. Also, it is more likely to be extendible, since not only string patterns but also structural patterns are benefitted from, which will be increased with future work.

4 Conclusion

The semantic relations between the words are extracted in order to develop a semantic network. Some basic relation types such as is-a, group-of, synonym-of, etc. are targeted to obtain an initial network to be extended with further work.

The words are investigated according to their definition in the TLA dictionary. Some row patterns which consist of morphological features of the words, parts of speech or strings in some specific positions and compound words are defined. After that, the dictionary is scanned for searching the definitions that matches one of these patterns. Depending on the results, patterns are reformed and additional features are inserted with the purpose of increasing pattern quality and number of matches. Exact senses of the words are tried to be matched by applying a word sense disambiguation algorithm.

The study has shown that, by taking advantage of the morphological richness of Turkish language and using some structural patterns, it is possible to construct a reasonable semantic network. This study can pave the way for more complex NLP applications and can be used for improving ordinary processes such as word sense disambiguation. The network can be converted into a knowledge base by inserting more accurate relationships and investigating larger and more comprehensive corpora as the future work.

5 Future Work

There is a set of processes to do both for improving and extending the network. Firstly, in order to eliminate erroneous connections from the obtained network, statistical information such as co-occurrence of the words can be investigated. The assumption here is that if two words are related to each other, the possibility of their being together in a corpus increases. The existing connections can be verified or ranked in terms of their reliability by using such information.

In addition, to remove erroneous sense determination, word sense disambiguation method can be improved. After obtaining a reliable, small network, which will serve as seed, new patterns can be extracted by following Turney (2006) and by using these patterns more instances can be extracted from larger corpora. As an alternative, the words can be first tagged with concrete or abstract labels automatically. This information can limit the types of connections a word can contribute. For example, an abstract word cannot connect to another word with a part-whole relation. For this task, a pre-processing step should be applied to classify the words as concrete or abstract.

In addition, with the purpose of improving the network, some other resources will be benefitted from. The existing patterns will be applied to Wikipedia (http://www.wikipedia.org/) entries, by selecting only the definitions of the concepts. An advantage of this process is that it can be re-performed periodically to keep the network up-to-date and dynamic. Also, the number of relation types will be increased. Currently, only the nouns, noun phrases consisting from two words, adjectives and verbs are handled. Also, only the relationships within the same type of words are extracted that is, a noun can be connected only to another noun, not an adjective or a verb. Finer grained relationships can establish connections among different parts of speech.

References

Barrière Caroline. 1997. From a children's first dictionary to a lexical knowledge base of conceptual graphs. PhD thesis. Simon Fraser University, Canada.

Bilgin Orhan, Çetinoğlu Özlem, and Oflazer Kemal. 2004. Morphosemantic Relations In and Across Wordnets: A Preliminary Study Based on Turkish. Proceedings of the Global WordNet Conference. Masaryk, Czech Republic.

http://people.sabanciuniv.edu/~oflazer/balkanet/twn_tr.htm

Celli Fabio, Nissim Malvina. 2009. Automatic identification of semantic relations in Italian complex nominals. Proceedings of the 8th International Conference on Computational Semantics, Tilburg. pp. 45-60.

Fellbaum, Christiane. 1998. WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Havasi Catherina, Speer Robert, and Alonso B. Jason. 2007. ConceptNet 3: a Flexible, Multilingual Semantic Network for Common Sense Knowledge. Proceedings of the 22nd Conference on Artificial Intelligence.

Lesk, E. Micheal. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In SIGDOC '86: Proceedings of the 5th annual international conference on Systems documentation, pages 24-26, New York, NY, USA. ACM.

Miller, George A. 1995. WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41.

Nakov Preslav, Hearts A. Marti. 2008. Solving Relational Similarity Problems Using the Web as a Corpus, Proceedings of ACL-08: HLT, Columbus, Ohio, USA. pp. 452–460.

Pantel Patrick, Pennacchiotti Marco. 2006. Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations, Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL. Sydney, Australia. pp. 113-120.

Turney D. Peter. 2006. Expressing implicit semantic relations without supervision. Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL. Sydney, Australia. pp. 313-320.

Önder Pınar. 2009. Design and Implementation of the semantic Turkish Language and Dialects Dictionary. MS thesis. Fatih University, İstanbul.