SemEval-2007 Task 12: Turkish Lexical Sample Task

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

Emine Çelik
Department of Computer Engineering, Fatih University 34500, Büyükçekmece, Istanbul, Turkey
eminemm@gmail.com

Neslihan Demirgüç
Department of Computer Engineering, Fatih University 34500, Büyükçekmece, Istanbul, Turkey
nesli_han@hotmail.com

Abstract

This paper presents the task definition, resources, and the single participant system for Task 12: Turkish Lexical Sample Task (TLST), which was organized in the SemEval-2007 evaluation exercise. The methodology followed for developing the specific linguistic resources necessary for the task has been described in this context. A language-specific feature set was defined for Turkish. TLST consists of three pieces of data: The dictionary, the training data, and the evaluation data. Finally, a single system that utilizes a simple statistical method was submitted for the task and evaluated.

1 Introduction

Effective parameters for word sense disambiguation (WSD) may vary for different languages and word types. Although, some parameters are common in many languages, some others may be language specific. Turkish is an interesting language that deserves being examined semantically. Turkish is based upon suffixation, which differentiates it sharply from the majority of European languages, and many others. Like all Turkic languages, Turkish is agglutinative, that is, grammatical functions are indicated by adding various suffixes to stems. Turkish has a SOV (Subject-Object-Verb) sentence structure but other orders are possible under certain discourse situations. As a SOV language where objects precede the verb, Turkish has postpositions rather than prepositions, and relative clauses that precede the verb. Turkish, as a widely-spoken language, is appropriate for semantic researches.

TLST utilizes some resources that are explained in Section 2-5. In Section 6 evaluation of the system is provided. In section 7 some concluding remarks and future work are discussed.

2 Corpus

Lesser studied languages, such as Turkish suffer from the lack of wide coverage electronic resources or other language processing tools like ontologies, dictionaries, morphological analyzers, parsers etc. There are some projects for providing data for NLP applications in Turkish like METU Corpus Project (Oflazer et al., 2003). It has two parts, the main corpus and the treebank that consists of parsed, morphologically analyzed and disambiguated sentences selected from the main corpus, respectively. The sentences are given in XML format and provide many syntactic features that can be helpful for WSD. This corpus and treebank can be used for academic purposes by contract.

The texts in main corpus have been taken from different types of Turkish written texts published in 1990 and afterwards. It has about two million words. It includes 999 written texts taken from 201 books, 87 papers and news from 3 different Turkish daily newspapers. XML and Text Encoding Initiative (TEI) style annotation have been used. The distribution of the texts in the Treebank is similar to the main corpus. There are 6930 sentences in this Treebank. These sentences have been parsed, morphologically analyzed and disambiguated. In Turkish, a word can have more than one analysis, so having disambiguated texts is very important.

Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 59–63,
Prague, June 2007.

```
<?xml version="1.0" encoding="windows-1254" ?>
  - <Set sentences="1">
   - <S No="1">
      <W IX="1" LEM="" MORPH="" IG="[(1,"soğuk+Adj")(2,"Adv+Ly")]"
         REL="[2,1,(MODIFIER)]">Soğukça</W>
      <W IX="2" LEM="" MORPH="" IG="[(1,"yanıtla+Verb+Pos+Past+A1sg")]"
         REL="[3,1,(SENTENCE)]">yanıtladım</W>
      <W IX="3" LEM="" MORPH="" IG="[(1,".+Punc")]" REL="[,( )]">.</W>
    </S>
  </Set>
```

Figure 1: XML file structure of the Treebank

Words	Main English translation	# Senses	MFS	Train size	Test size	Total #of instances
Nouns						
ara	distance, break, interval, look for	7	53	192	63	255
baş	head, leader, beginning, top, main, principal	5	34	68	22	90
el	hand, stranger, country	3	75	113	38	151
göz	eye, glance, division, drawer	3	48	92	27	119
kız	girl, virgin, daughter, get hot, get angry	2	72	96	21	117
ön	front, foreground, face, breast, prior, preliminary anterior	5	21	72	23	95
sıra	queue, order, sequence, turn, regularity, occasion desk	7	30	85	28	113
üst	upper side, outside, clothing	7	20	69	23	92
yan	side, direction, auxiliary, askew, burn, be on fire be alight	5	21	65	31	96
yol	way, road, path, method, manner, means	6	17	68	29	97
Average		**5**	**39**	**92**	**31**	**123**
Verbs						
al	take, get, red	24	180	963	125	1088
bak	look, fac, examine	4	136	207	85	292
çalış	work, study, start	4	33	103	61	164
çık	climb, leave, increase	6	45	138	87	225
geç	pass,happen, late	11	51	164	90	254
gel	come, arrive, fit, seem	20	154	346	215	561
gir	enter, fit, begin, penetrate	6	88	163	84	247
git	go, leave, last, be over, pass	13	130	214	120	334
gör	see, understand, consider	5	155	206	68	274
konuş	talk, speak	6	42	129	63	192
Average		**9.9**	**101.4**	**263.3**	**99.8**	**363.1**
Others						
büyük	big, extensive, important, chief, great, elder	6	34	97	26	123
doğru	straight, true, accurate, proper, fair, line towards, around	6	29	81	38	119
küçük	little, small, young, insignificant, kid	4	14	45	14	59
öyle	such, so, that	4	20	51	23	74
son	last, recent, final	2	76	86	18	104
tek	single, unique, alone	2	38	40	10	50
Average		**4**	**35.2**	**66.7**	**21.5**	**88.2**

Table 1: Target words in the SEMEVAL-1 Turkish Lexical Sample task

Frequencies of the words have been found as it is necessary to select appropriate ambiguous words for WSD. There are 5356 different root words and 627 of these words have 15 or more occurrences, and the rest have less.

The XML files contains tagging information in the word (morphological analysis) and sentence level as a parse tree as shown in Figure 1. In the word level, inflectional forms are provided. And in the sentence level relations among words are given. The S tag is for sentence and W tag is for the word. IX is used for index of the word in the sentence, LEM is left as blank and lemma is given in the MORPH tag as a part of it with the morphological analysis of the word. REL is for parsing information. It consists of three parts, two numbers and a relation. For example REL="[2, 1, (MODIFIER)]" means this word is modifying the first inflectional group of the second word in the sentence. The structure of the treebank data was designed by METU. Initially lemmas were decided to be provided as a tag by itself, however, lemmas are left as blank. This does not mean that lemmas are not available in the treebank; the lemmas are given as a part of "IG" tag. Programs are available for extracting this information for the time being. All participants can get these programs and thereby the lemmas easily and instantly.

The sense tags were not included in the treebank and had to be added manually. Sense tagging has been checked in order to obtain gold standard data. Initial tagging process has been finished by a single tagger and controlled. Two other native speaker in the team tagged and controlled the examples. That is, this step was completed by three taggers. Problematic cases were handled by a commission and the decision was finalized when about 90% agreement has been reached.

3 Dictionary

The dictionary is the one that is published by TDK [1] (Turkish Language Foundation) and it is open to public via internet. This dictionary lists the senses along with their definitions and example sentences that are provided for some senses. The dictionary is used only for sense tagging and enumeration of the senses for standardization. No specific information other than the sense numbers is taken from the dictionary; therefore there is no need for linguistic processing of the dictionary.

4 Training and Evaluation Data

In Table 1 statistical information about the final training and testing sets of TLST is summarized. The data have been provided for 3 words in the trial set and 26 words in the final training and testing sets (10 nouns, 10 verbs and 6 other POS for the rest of POS including adjectives and adverbs). It has been tagged about 100 examples per word, but the number of samples is incremented or decremented depending on the number of senses that specific word has. For a few words, however, fewer examples exist due to the sparse distribution of the data. Some ambiguous words had fewer examples in the corpus, therefore they were either eliminated or some other examples drawn from external resources were added in the same format. On the average, the selected words have 6.7 senses, verbs, however, have more. Approximately 70% of the examples for each word were delivered as training data, whereas approximately 30% was reserved as evaluation data. The distribution of the senses in training and evaluation data has been kept proportional. The sets are given as plain text files for each word under each POS. The samples for the words that can belong to more than one POS are listed under the majority class. POS is provided for each sample.

We have extracted example sentences of the target word(s) and some features from the XML files. Then tab delimited text files including structural and sense tag information are obtained. In these files each line has contextual information that are thought to be effective (Orhan and Altan, 2006; Orhan and Altan, 2005) in Turkish WSD about the target words. In the upper level for each of them XML file id, sentence number and the order of the ambiguous word are kept as a unique key for that specific target. In the sentence level, three categories of information, namely the features related to the previous words, target word itself and the subsequent words in the context are provided.

[1] http://tdk.org.tr/tdksozluk/sozara.htm

Feature	Example
File id	00002213148.xml
Sentence number	9
Order	0
Previous related word root/lemma	tap
Previous related word POS(corrected)	verb
Previous related word onthology level1	abstraction
Previous related word onthology level2	attribute
Previous related word onthology level3	emotion
Previous related word POS	verb
Previous related word POS(derivation)	adv
Previous related word case marker	?
Previous related word possessor	fl
Previous related word-target word relation	modıfıer
Target word root/lemma	sev
Target word POS	verb
Target word POS(derivation)	noun
Target word case marker	abl
Target word possessor	tr
Target word-subsequent word relation	object
Subsequent related word root/lemma	sıkıl
Subsequent related word POS(corrected)	verb
Subsequent related word onthology level1	abstraction
Subsequent related word onthology level2	attribute
Subsequent related word onthology level3	emotion
Subsequent related word POS	verb
Subsequent related word POS(derivation)	verb
Subsequent related word case marker	?
Subsequent related word possessor	fl
Subsequent related word-target word relation	sentence
Fine-grained sense number	2
Coarse-grained sense number	2
Sentence	#ne tuhaf şey ; değil mi ? iyi olmamdan ; onu taparcasına sevmemden sıkıldı .#

Table 2: Features and example

In the treebank relational structure, there can be more than one word in the previous context related to the target, however there is only a single word in the subsequent one. Therefore the data for all words in the previous context is provided separately. The features that are employed for previous and the subsequent words are the same and they are the root word, POS(corrected), tags for ontology level 1, level 2 and level 3, POS, inflected POS, case marker, possessor and relation. However for the target word only the root word, POS, inflected POS, case marker, possessor and relation are taken into consideration. Fine and coarse-grained (FG and CG respectively) sense numbers and the sentence that has the ambiguous word have been added as the last three feature. FG senses are the ones that are decided to be the exact senses. CG senses are given as a set that are thought to be possible alternatives in addition to the FG sense. Table 2 demonstrates the whole list of features provided in a single line of data files along with an example. The "?" in the features shows the missing values. This is actually corresponding to the features that do not exist or can not be obtained from the treebank due to some problematic cases. The

line that corresponds to this entry will be the following line (as tab delimited):

00002213148.xml 9 0 tap verb abstraction attribute emotion verb adv ? fl modıfıer sev verb noun abl tr object sıkıl verb abstraction attribute emotion verb verb ? fl sentence 2 2 #ne tuhaf şey ; değil mi ?iyi olmamdan ; onu taparcasına sevmemden sıkıldı .#

5 Ontology

A small scale ontology for the target words and their context was constructed. The Turkish WordNet developed at Sabancı University[2] is somehow insufficient. Only the verbs have some levels of relations similar to English WordNet. The nouns, adjectives, adverbs and other words that are frequently used in Turkish and in the context of the ambiguous words were not included. This is not a suitable resource for fulfilling the requirements of TLST and an ontology specific to this task was required. The ontology covers the examples that are selected and has three levels of relations that are supposed to be effective in the disambiguation process. We tried to be consistent with the WordNet tags; additionally we constructed the ontology not only for nouns and verbs but for all the words that are in the context of the ambiguous words selected. Additionally we tried to strengthen the relation among the context words by using the same tags for all POS in the ontology. This is somehow deviating from WordNet methodology, since each word category has its own set of classification in it.

6 Evaluation

WSD is a new area of research in Turkish. The sense tagged data provided in TLST are the first resources for this specific domain in Turkish. Due to the limited and brand new resources available and the time restrictions the participation was less. We submitted a very simple system that utilizes statistical information. It is similar to the Naïve Bayes approach. The features in the training data was used individually and the probabilities of the senses are calculated. Then in the test phase the probabilities of each sense is calculated with the given features and the three highest-scored senses are selected as the answer. The average precision and recall values for each word category are given in Table 3. The values are not so high, as it can be expected. The size of the training data is limited, but the size is the highest possible under these circumstances, but it should be incremented in the near future. The number of senses is high and providing enough instances is difficult. The data and the methodology for WSD will be improved by the experience obtained in SemEval evaluation exercise.

The evaluation is done only for FG and CG senses. For FG senses no partial points are assigned and 1 point is assigned for a correct match. On the other hand, the CG senses are evaluated partially. If the answer tags are matching with any of the answer tags they are given points.

Words	FG		CG	
	P	R	P	R
Nouns	0,15	0,50	0,65	0,43
Verbs	0,10	0,38	0,56	0,50
Others	0,13	0,50	0,57	0,44
Average	**0,13**	**0,46**	**0,59**	**0,46**

Table 3: Average Precision and Recall values

7 Conclusion

In TLST we have prepared the first resources for WSD researches in Turkish. Therefore it has significance in Turkish WSD studies. Although the resources and methodology have some deficiencies, a valuable effort was invested during the development of them. The resources and the methodology for Turkish WSD will be improved by the experience obtained in SemEval and will be open to public in the very near future from http://www.fatih.edu.tr/~zorhan/senseval/senseval.htm.

References

Orhan, Z. and Altan, Z. 2006. *Impact of Feature Selection for Corpus-Based WSD in Turkish*, LNAI, Springer-Verlag, Vol. 4293: 868-878

Orhan Z. and Altan Z. 2005. *Effective Features for Disambiguation of Turkish Verbs*, IEC'05, Prague, Czech Republic: 182-186

Oflazer, K., Say, B., Tur, D. Z. H. and Tur, G. 2003. *Building A Turkish Treebank*, Invited Chapter In Building And Exploiting Syntactically-Annotated Corpora, Anne Abeille Editor, Kluwer Academic Publishers.

[2] http://www.hlst.sabanciuniv.edu/TL/