International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:1, No:7, 2007

Effective Features for Disambiguation of Turkish Verbs

Zeynep Orhan, and Zeynep Altan

***Abstract*—**This paper summarizes the results of some experiments for finding the effective features for disambiguation of Turkish verbs. Word sense disambiguation is a current area of investigation in which verbs have the dominant role. Generally verbs have more senses than the other types of words in the average and detecting these features for verbs may lead to some improvements for other word types. In this paper we have considered only the syntactical features that can be obtained from the corpus and tested by using some famous machine learning algorithms.

***Keywords*—**Word sense disambiguation, feature selection.

I. Introduction

IN the history language is thought to be nomenclature and this view has dominated Western thought for a long time. Words have a meaning since they stand for something else in this view. In Saussure period[1], this view has been criticized and the approach of words standing for an idea in the mind has been offered against words standing for something which independently exists in the world, i.e. nomeclaturism. Wittgenstein[2] suggests that different types of word/meaning relationships correspond to different types of games[1].

The thoughts on word and meaning affected the word sense disambiguation (WSD) researches. It is difficult to define this task due to many different views about the words and the meaning. However, WSD can be described as the task of assigning the most appropriate meaning to a polysemous word within a given context. Natural language sentences are given as the inputs of the WSD programs and the expected output of such programs are the assignment of sense tags to the ambiguous words that generally correspond to specific sense definitions listed in a dictionary or another similar source

WSD is a long-standing problem in NLP. It is crucial to figure out what each individual word in a sentence means for understanding natural languages. Words in natural languages are known to be highly ambiguous. This fact is true especially for the frequently occurring words of a language. For example, in the WordNet dictionary, the average number of senses for each noun for the most frequent 121 nouns in English is 7.8, but that for the most frequent 70 verbs is 12.0 [3]. This set of 191 words is estimated to account for about 20 percent of all word occurrences in any English free text. Therefore WSD is a difficult and hard to master task in NLP. However, once it has been achieved it can be useful for many other tasks. It is also an essential part of many NLP applications.

A large range of applications, including machine translation, knowledge acquisition, information retrieval, information extraction, and others, require knowledge about word meanings, and therefore WSD algorithms represent a necessary (at least a helpful) step in all these applications. Starting with Senseval-1 in 1998 [4], WSD has received growing attention from the NLP community. The tests on part of the TREC corpus, a standard information-retrieval test collection, proved that WSD improved precision by about 4.3 percent [5]. Similarly, in machine translation, WSD has been used to select the appropriate words to translate into a target language. Specifically, it has been reported successful use of WSD to improve the accuracy of machine translation[6]. These works are good indicators of the utility of WSD in practical NLP applications.

Manual encoding can be laborious and time consuming. Additionally, manual maintenance and further expansion become increasingly complex causing a scalability problem. Nowadays corpus-based methods are more popular due to these problems. In these approaches, machine-learning techniques are used to automatically acquire disambiguation knowledge. Sense-tagged corpora and large-scale linguistic resources such as online dictionaries are the fundamental components of a typical WSD system. There are two main decisions that have to be considered in the design of a WSD system:

- the set of features to be used and
- the learning algorithm.

The first decision is finding the appropriate set of features and commonly used features include surrounding words in a given window size and their part of speech [7], keywords [3] or bigrams in the context [8] and various syntactic properties [9] etc. Intuitively, a good feature should capture an important source of knowledge critical in determining the sense of an ambiguous word. The features that are included in [10] are

Zeynep Orhan is with Istanbul University, Faculty of Engineering, Department of Computer Engineering, 34850, Avcılar, Istanbul, Turkey (e-mail: zeyneporhan@yahoo.com).

Zeynep Altan is with Maltepe University, Faculty of Engineering, Department of Computer Engineering, 34857, Maltepe, Istanbul, Turkey (e-mail: zaltan@istanbul.edu.tr).

[1] Game Theory: The meaning of a word or phrase is nothing other than the set of (informal) rules governing the use of the expression in actual life. Analogous to the game rules, language usage rules are neither right nor wrong, neither true nor false and their role depend on the particular usage or application of their user.

surrounding words, local collocations[2] ([11], [12], [13], [3]), syntactic relations, parts of speech and morphological forms ([3], [14]). Another set of features given in [15] are ccurrent ambiguous word, current part of speech, contextual features (the words and parts of speech of K words surrounding current word), collocations formed with maximum K words surrounding, head of noun phrase, sense specific keywords (maximum MX keywords occurring at least MN times, bigrams (maximum MX bigrams occurring at least MN times are determined for all training examples, the verbs, nouns, named entities, prepositions, pronouns, determiners before and after the ambiguous word. This set is a high coverage set of features that can be used in WSD, and the most of-ten used features are determined to be turn out to be ccurrent ambiguous word, current part of speech, contextual features and collocations which are also the features most frequently mentioned in the literature.

The next step following the selection of features to encode the context and form the training examples is the decision about the methods that is to be used. There are many different approaches proposed in the WSD research. Bayesian probabilistic algorithms ([16], [17], [14], [18], [19], [20]), neural networks ([21], [18]), and decision trees(DT) ([22]), decision lists(DL) ([21], [23]), memory based learning(MBL)[3] ([24], [21], [3], [25], [26]) etc

The comparisons of different algorithms are discussed in [27] and very interesting results have been observed. In [21] it has been stated that the performance of the algorithms were given as follows: Naive Bayes(NB) & perceptron > DL > MBL. However, in [28] it has been claimed that MBL has a better performance than NB. These results are somehow contradictory and additionally, in [29] a different set of algorithms has been compared and they have suggested that the Support vector machines(SVM) are better than Adb, NB, DT. Considering these results it has been concluded taht the comparisons could be reliable iff the two algorithms were given constant input features and constant algorithm parameter settings, and generally best features or best parameter settings were unpredictable for a particular task and for a particular ML algorithm[27].

In many artificial intelligence applications features were studied carefully and variable sets of features have been successfully used. Automatic feature selection has become a hot topic in many researches. It has been used searching algorithms for feature selection in the domain of cloud types' classification and obtained increased performance [30]. In [31], they have used many different efficient search algorithms for the detection of optimal feature subsets and performed successful experiments on several synthetic datasets. Cardie [32] described a linguistic and cognitive biased approach that considered the application of instance based learning with automatic feature selection for relative pronoun resolution. Domingos [33] selected different features for each instance in the training set by using a context sensitive feature selection algorithm. In [7] decomposable probabilistic models are used in combination with eager NB algorithms. Mihalcea [34] described an algorithm for WSD in which a lazy learner, improved with automatic feature selection, has been used.

II. Turkish Verb Sense Disambiguation

WSD researches are as old as NLP researches, unfortunately, some specific languages such as English and some other European languages were the main concern of the applications. As a matter of this fact, there does not exist too many practical applications for lesser studied languages, like Turkish[35], due to the scarcity of electronically available NLP resources (e.g. WordNets, POS tagging, morphological analyzers, etc.) for them.

We are trying to develop a project for Turkish WSD and as a first attempt in this context we have used seven stories from world classics[4]. These stories were plain texts so we have processed these texts manually by annotating the senses of the ambiguous words and the other features such as POS, affixes and collocations. In many computational lingusitics (CL) or NLP tasks such as machine translation (MT), WSD, or text categorization classification of data is important. In generative classifiers, generally maximum likely hood is used and methods such as NB or EBL deal with the joint probabilities of the words in context. In a previous research [36], these two methods have been tested on a small set of sentences that contain the highly ambiguous verb *git*(to go) and EBL method performed slightly better than NB on this small data set.

We did not have appropriate general purpose corpus for NLP applications in Turkish and manual encoding of necessary information was a very time-consuming and error-prone task. The studies on Turkish NLP have been improving in recent years and in parallel to these improvements there have been some ongoing projects for developping corpus for NLP applications in Turkish. However, some of them do not have a broad coverage or some others are not open to public. Considering these difficulties, we have used the corpus that has been released for academic purposes for the first time by METU[37]. METU Turkish corpus has become available for academic investigations in 2003. The corpus has two forms:

- METU Turkish Corpus is a collection of 2 million words of post-1990 written Turkish samples and this one is the basic corpus.

[2] A local collocation refers to a short sequence of words near w, taking the word order into account. Such a sequence of words need not be an idiom to qualify as a local collocation. Collocations differ from surrounding words in that word order is taken into consideration.

[3] Memory based learning (MBL) is also called instance based (IBL) or exemplar-based learning (EBL)

[4] Stories are: Gulliver, Candide, Ivan Nikiforovic, Tours Papazı, Mozart Prag Yolunda, Mektuplar, Kır Atlı.

TABLE I
FORMAT OF THE METU TREEBANK

```
<?xml version="1.0" encoding="windows-1254" ?>
- <Set sentences="1">
- <S No="1">
 <W IX="1" LEM="" MORPH="" IG="[(1,"nere+Pron+QuesP+A3sg+Pnon+Dat")]" REL="[2,1,(OBJECT)]">Nereye</W>
 <W IX="2" LEM="" MORPH="" IG="[(1,"git+Verb+Pos+Narr+A3sg")]" REL="[3,1,(OBJECT)]">gitmiş</W>
 <W IX="3" LEM="" MORPH="" IG="[(1,"ol+Verb+Pos")(2,"Verb+Able+Aor+A3sg")]" REL="[5,1,(SENTENCE)]">olabilir</W>
 <W IX="4" LEM="" MORPH="" IG="[(1,"Osman+Noun+Prop+A3sg+Pnon+Nom")]" REL="[3,2,(SUBJECT)]">Osman</W>
 <W IX="5" LEM="" MORPH="" IG="[(1,".+Punc")]" REL="[,( )]">.</W>
 </S>
 </Set>
```

- METU-Sabanci Turkish Treebank, a subset of the basic corpus, is a morphologically and syntactically annotated treebank corpus sentences.

They have used XML and TEI(Text Encoding Initiative) style annotation and tried to obtain a corpus similiar to BNC (British National Corpus). The sentences provided many syntactic features that could be helpful for disambiguation. However, there were some errors and inconsistencies in the treebank and have to be corrected. Sense tagging was still problematic and achieved manually. The structure of the treebank is given in Table I.

Wordnets are other important resource types for most of the NLP applications in some languages. WordNet [38] for English provides many valuable relations of the words that can be helpful in various domains such as synonymy/antonym, hyponymy/hypernymy and some other relationships among words. On the other hand, obtaining the appropriate set of senses for a given word can be the bottleneck in WSD, since the majority of the on-line dictionaries list various specific usages of the words along with their general meanings rather than classifying them into some refined set of senses. Only 51% of the synsets of the nouns contains a single word [39] in WordNet which is an outcome of a long and careful study and this ratio is definetly higher in some other types of dictionaries.

Although, wordnets are essential or at least helpful in WSD, they are still in the development phase for some other languages. Balkanet [40] Project that includes a Turkish WordNet[**][41] is one such effort[††]. However, the completed version has not been released for the time being, but it can be helpful for future works.

We applied machine learning algorithms of WEKA[42] in WSD task. The system provides many visualization tools and a detailed analysis of the output. We have extracted the features given in Table III for our experiments. These are the possible set of syntactic features that can be obtained from the data. The input to the algorithms were prepared according to the arff format which is a standart input data format for the WEKA system.

The structure of Turkish language enforces POS to be obtained for a better WSD, unfortunately it is not an easy task and we have used POS of the words provided in the corpus as an effective feature. In our experiments we have used the word *gel* (to come). In TDK[‡‡] dictionary the set of senses for this word has 37 entries excluding idiomatic usages and the compound words. In Turkish WordNet[§§] the same word has 4 senses. Additionally there are many other classifications for the same word in other Turkish dictionaries. Therefore we have studied the senses in an incremental manner. First we have divided the usages of this word into two in which the first one symbolizes the primary sense of coming, arriving, going etc and the second one for the rest of the usages and tested the selected features accordingly. Then the second set has been divided into two as the ones that are totally idiomatic and the others. As the last part we have divided this set into three the added category was the one that has the meaning of time for something or turn of something. We had three set of data S1(2 senses), S2(3 senses), S3(4 senses). The prior distributions of the senses are given in Table II.

TABLE II
SENSE DISTRIBUTIONS

Sense groups	1	2	3	4
S1	0.59	0.41		
S2	0.59	0.14	0.27	
S3	0.59	0.1	0.27	0.04

[**] The Human Language and Speech Technologies Laboratory at the Sabanci University, Istanbul, is the participant in the Balkanet Project

[††] The BalkaNet project's aim is to develop a multilingual database with WordNets for a set of Balkan languages, based on the model of the EuroWordNet project, a multilingual lexical database comprising of eight different European languages, semantically represented in it. Greek, Turkish, Bulgarian, Romanian, Czech and Serbian are included in BalkaNet. Each language's Wordnet are classified according to their semantic relations, but they share common ontology. The aim is to organize a common database for the lesser studied Balkan languages integrating and comparing them cross-linguistically

[‡‡] TDK dictionary, http://tdk.org.tr/tdksozluk

[§§] Turkish WordNet: http://www.hlst.sabanciuniv.edu/TL

TABLE III
SYNTACTICAL FEATURES OBTAINED FROM METU CORPUS

Sentence number
File number
Previous word root
Previous word POS
Previous word inflected POS
Previous word case marker
Previous word possessor
Previous-target word relation
Target word root
Target word POS
Target word inflected POS
Target word case marker
Target word possessor
Target-subsequent word relation
Subsequent word root
Subsequent word POS
Subsequent word inflected POS
Subsequent word case marker
Subsequent word possessor
Subsequent- Subsequent word relation
Sense number

TABLE IV
FEATURE GROUPS USED IN THE TEST DATA

Group	Features
G1	All features
G2	All features related with the previous word
G3	Only the previous word root
G4	All features related with the subsequent word
G5	The previous and subsequent word roots

The algorithms that were selected from the WEKA are AODE(improved version of NB, statistical method), IBk(exemplar-based), and J48(C4.5 algorithm, decision tree). The selected features for these algorithms are given in Table IV. We have tested the effects of the features on the five groups of feature combinations for three sense classification data sets by the three algorithms and obtained the results in Table V.

III. CONCLUSION AND FUTURE WORK

The performance of different machine learning algorithms are either very close to each other or the fluctuations can be ignored. The algorithms offered so far for WSD are far or less provide these type of results. On the other hand, selection of the features has a serious impact on the results. Detecting effective features is more important than the algorithms used. We have studied only a small set of syntactical features and observed that different set of features has dramatically changed the experimental results.

Selecting an appropriate set of features can be helpful in many ways. First of all, by eliminating the useless features we can improve the run time efficiency, decrease the cost of operations and increase the accuracy. The results have shown that using all the features mentioned in Table III is not giving better results than the ones related to the previous word. Sometimes using only the previous root word has a compatitive performance with all the other features.

The above observations are true for the verbs but the features that are important for the other word types(nouns, adjectives, adverbs etc.) have to be examined seperately.

Moreover, these are only the syntactical clues and there may be some other hidden features or pragmatical issues that can not be obtained from the corpus but have to be extracted from some other resources such as ontologies or other types of human knowledge represantation resources.

TABLE V
TEST RESULTS (PERCENTAGES OF THE CORRECTLY CLASSIFIED INSTANCE)

roup	AODE			IBk			J48		
	S1	S2	S3	S1	S2	S3	S1	S2	S3
G1	69.26	65.27	64.90	70.78	69.07	69.45	69.26	65.84	65.27
G2	70.78	66.22	64.90	71.56	65.27	65.27	70.78	65.46	65.27
G3	67.93	65.27	65.65	67.93	65.27	65.65	65.84	65.27	65.27
G4	61.29	59.77	59.58	62.80	62.80	62.62	59.78	59.96	59.77
G5	69.63	69.07	69.07	69.45	68.12	68.12	65.27	64.51	63.95

REFERENCES

[1] Saussure, Ferdinand de. 1974 [1916]. Course in General Linguistics. Tr. Wade Baskin. Glasgow: Fontana & Collins. [Orig.: Cours de linguistique générale.Lousanne et Paris: Payot.]
[2] Canfield J.V. (Editor), 1997, Philosophy of Meaning, Knowledge and Value in the 20th Century: Routledge History of Philosophy Volume 10. British Library Cataloguing in Publication data.
[3] Ng, H.T., and Lee, H.B., 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL-96), Santa Cruz.
[4] SENSEVAL: An Exercise in Evaluating Word Sense Disambiguation Programs A. Kilgarriff. In Proc. LREC, Granada, May 1998. Pp 581--588.
[5] Schutze, H., and Pedersen, J. 1995. Information Retrieval Based on Word Senses. In Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval, 161-175. Las Vegas, Nev.: University of Nevada at Las Vegas.
[6] Ido Dagan , Alon Itai, Word sense disambiguation using a second language monolingual corpus, Computational Linguistics, v.20 n.4, p.563-596, December 1994.
[7] R. Bruce and J. Wiebe. 1999. Decomposable modeling in natural language processing. Computational Linguistics, 25(2):195-207.
[8] Pedersen, T., 2001. A decision tree of bigrams is an accurate predictor of word sense. In Proceedings of the North American Chapter of the Association for Computational Linguistics, NAACL 2001, pages 79-86, Pittsburg.
[9] Fellbaum, C., Palmer, M., Dang, H.T., Delfs, L., and Wolf., S., 2001. Manual and automatic semantic annotation with WordNet. In WordNet and Other lexical resources: NAACL 2001 workshop, pages 3-10, Pittsburgh.
[10] Ng, H. T., Zelle, J., Winter, 1997, Corpus-based approaches to semantic interpretation in natural language processing - Natural Language Processing, AI Magazine.
[11] Kelly, E. and Stone, P. (1975) Computer Recognition of English Word Senses, North Holland, Amsterdam.
[12] Yarowsky, D. 1993. One Sense per Collocation. In Proceedings of the ARPA Human-Language Technology Workshop, 266-271. Washington, D.C.: Advanced Research Projects Agency.
[13] Yarowsky, D. 1994. Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French. In Proceedings of the Thirty-Second Annual Meeting of the Association for Computational Linguistics, 88-95. Somerset, N.J.: Association for Computational Linguistics.
[14] Bruce, R. and J. Wiebe. 1994. Word-sense disambiguation using decomposable models. In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, pages 139-- 146.
[15] Rada Mihalcea, August 2002, Instance Based Learning with Automatic Feature Selection Applied to Word Sense Disambiguation, in Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002), Taiwan.
[16] Pedersen, Ted and Rebecca Bruce. 1997. A new supervised learning algorithm for word sense disambiguation. In Proceedings of the 14th National Conference on Artificial Intelligence (AAAI-97), Providence, RI.
[17] R. Mooney. 1996. Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning. In Proceedings of the 1996 Conference on Empirical Methods in Natural Language Processing (EMNLP-1996), pages 82-91, Philadelphia.
[18] Leacock, C., Towell, G. and Voorhees, E. M., 1993 "Corpus-based statistical sense resolution." In Proceedings of the ARPA Human Languages Technology Workshop.
[19] Gale, W., K. Church, and D. Yarowsky. ``Work on Statistical Methods for Word Sense Disambiguation." In Proceedings, AAAI Fall Symposium on Probabilistic Approaches to Natural Language. Cambridge, MA, pp. 54-60, 1992.
[20] Yarowsky, D. `` Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora." In Proceedings, COLING-92. Nantes, pp. 454-460, 1992.
[21] R. Mooney. 1996. Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning. In Proceedings of the 1996 Conference on Empirical Methods in Natural Language Processing (EMNLP-1996), pages 82-91, Philadelphia.
[22] Pedersen, T., 2001. A decision tree of bigrams is an accurate predictor of word sense. In Proceedings of the North American Chapter of the Association for Computational Linguistics, NAACL 2001, pages 79-86, Pittsburg.
[23] Yarowsky, D. `` Hierarchical Decision Lists for Word Sense Disambiguation." Computers and the Humanities, 34(2):179-186, 2000.
[24] H. T. Ng. 1997. Exemplar-Base Word Sense Disambiguation: Some Recent Improvements. In Procs. of the 2nd Conference on Empirical Methods in Natural Language Processing, EMNLP.
[25] C. Cardie. 1993. A case-based approach to knowledge acquisition for domain-specific sentence analysis. In Proceedings of the Eleventh National Conference on Artificial Intelligence, pages 798-803, Washington, DC.
[26] Veenstra, A. van den Bosch, J., S. Buchholz, W. Daelemans, and J. Zavrel. 2000 Memory-based word sense disambiguation Computers and the Humanities, 34:171-177.
[27] Daeleman, W.,Machine Learning of Language: A Model and a Problem, ESSLLI'2002 Workshop on Machine Learning Approaches in Computational Linguistics, August 5 - 9, 2002, Trento, Italy.
[28] G. Escudero, L. Mrquez, and G. Rigau. 2000, Naive Bayes and Exemplar-Based Approaches to Word Sense Disambiguation Revisited. In Proceedings of the 14th European Conference on Artificial Intelligence, ECAL.
[29] Lee, Yoong Keok, & Ng, Hwee Tou. An Empirical Evaluation of Knowledge Sources and Learning Algorithms for Word Sense Disambiguation. Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP-2002). pp. 41-48, 2002.
[30] D.W. Aha and R.L. Bankert. 1994. Feature selection for case-based classification of cloud types: An empirical comparison. In Proceedings of the AAAI'94 Workshop on Case-Based Reasoning, pages 106-112, Seattle, WA.
[31] A.W. Moore and M.S. Lee. 1994. Efficient algorithms for minimizing cross validation error. In International Conference on Machine Learning, pages 190-198, New Brunswick.
[32] C. Cardie. 1996. Automating feature set selection for case-based learning of linguistic knowledge. In Proceedings of the Conference on Empirical Methods in Natural Language Processing EMNLP, pages 113-126, Somerset, New Jersey.
[33] P. Domingos. 1997. Context-sensitive feature selection for lazy learners. Artificial Intelligence Review, (11):227-253.
[34] Mihalcea, R., Instance Based Learning with Automatic Feature Selection Applied to Word Sense Disambiguation, in Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002), Taiwan, August 2002.
[35] Yılmaz, O., September, 1994, Design and implementation of a verb lexicon and sense disambiguator for Turkish, MS. Thesis, Bilkent University, Ankara, Turkey.
[36] Orhan Z., Altan Z., 2003, "Disambiguation of Turkish Word Senses By Supervised Statistical Methods", International XII. Turkish Symposium on Artificial Intelligence and Neural Networks (TAINN 2003).
[37] Nart B. Atalay, Kemal Oflazer, Bilge Say, The Annotation Process in the Turkish treebank, in Proceedings of the EACL Workshop on Linguistically Interpreted Corpora-LINC, April 13-14, 2003, Budapest.
[38] Fellbaum C., 1998, WordNet: An Electronic Lexical Database. The MIT press.
[39] Ciaramita M., Johnson M. , 2004, "Multi-Component Word Sense Disambiguation" Proceedings of Senseval-3: The Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, pp. 97-100.
[40] Stamou, S., Oflazer, K., Pala, K., Christodoulakis, D., Cristea, D., Tufis, D., Koeva, S., Totkov, G., Dutoit, D., Grigoriadou, M., BalkaNet: A multilingual Semantic Network for Balkan Languages, in Proceedings of the First International WordNet Conference, Mysore India, January 2002.
[41] O. Bilgin, Çetınoğlu, Ö., Oflazer, K., Building a Wordnet for Turkish, Romanian Journal of Information Science and Technology, Volume 7, Numbers 1–2, 2004, 163–172.
[42] WEKA system, http://www.cs.waikato.ac.nz/ml/weka.