Word Sense Disambiguation for Semantic Applications

Zeynep Orhan[1], Zeynep Altan[2]

[1]Department of Computer Engineering, Fatih University 34500, Büyükçekmece, Istanbul, Turkey

[2]Department of Computer Engineering, Maltepe University 34857, Maltepe, Istanbul, Turkey
zorhan@fatih.edu.tr
zaltan@maltepe.edu.tr

Abstract: Natural language processing (NLP) has become the most significant obstacle that has been restricting the applications via the web. Today, very little of the content on the web can be understood by the machines, although vast amount of electronic information has been kept on them. Word sense disambiguation (WSD) is an important intermediate step in many language processing applications. It is basically a mapping function from the context to the set of senses. This function has many parameters that are difficult to explore. The factors effecting the success of WSD systems are generally very sensitive to these parameters. The issues in WSD are examined in the context of Turkish WSD application. Ambiguous words and their sense classifications have been established and by providing manually sense tagged corpora and examining WSD problem from various perspectives, an important contribution has been achieved for the researches in this domain.

1 Introduction

The internet has had a tremendous impact on society and business in the last decade leading to various applications. Natural language processing (NLP) has become the most significant obstacle that has been restricting these applications via the web. Today, very little of the content on the web can be understood by the machines, although vast amount of electronic information has been kept on them. The consequences of the machine understandable web will be incredible and smarter applications will emerge simultaneously. However, theoretical and practical framework of semantic technologies needs to be matured. There are many different issues that have to be handled about semantic applications such as ambiguity.

Ambiguity is a serious problem that may occur in various domains. Whenever an ambiguity arises in the application areas such as mathematics, programming languages etc. where certainty must be ensured, disambiguation can be achieved easily by using some strict rules of that specific domain. If this was not the case, unexpected results could occur. For example, unless the operator precedence is used, the result of 3+2*4 is ambiguous and can be either 11 or 20. However, since multiplication has precedence over addition the answer is obviously 11. This precedence rule definitely solves the ambiguity of such mathematical expressions and does not allow misinterpretations. Unfortunately, disambiguation task is not an easy process in many applications, especially in NLP.

NLP applications can be categorized into two parts as the end tasks and the intermediate ones considering their functionalities. Machine translation (MT), information retrieval/extraction (IR/IE), search engines, etc. are the typical end tasks in which complete solutions have been provided. On the other hand, morphological analysis, parsing and word sense disambiguation (WSD), etc. are intermediate tasks that contribute to the end tasks rather than providing overall solutions [IV98].

Many words have different meanings/senses and generally, ambiguity arises for those words. WSD is the process of selecting the most suitable senses of the ambiguous words that are invoked in that particular usage by considering other contextual features. Internet is a multilingual environment keeping vast amount of information in many domains. Today, WSD is very important for search engines and translation applications. For example, if someone is trying to find documents about “mouse” the accuracy of the results will be determined on the sense of this word. It can be either an “electronic device” or an “animal” depending on the content. Generally nouns and their senses are the important clues in searching. However, if the main concern is a translation task or query and/or question answering system, the disambiguation of verbs are becoming more problematic in addition to the problems of nouns and other word types. Someone may not need to distinguish between the senses of the word “mouse” and translate both senses as “fare” to Turkish; however for searching relevant documents, the distinction must be clear.

The early work on WSD concentrated on hand-coded knowledge [KS75]. However, this can be laborious and time consuming. Additionally, manual systems always suffer from the scalability. The alternative to this approach is the corpus-based methods. Machine learning techniques are used to automatically acquire disambiguation knowledge and utilized in NLP [Ka96]. In WSD researches, supervised machine learning methods are successfully applied. Sense-tagged corpora and large-scale linguistic resources, such as online dictionaries became the fundamental components of typical WSD systems parallel to the development of electronic resources in the last decades. Therefore they are very sensitive to the resources that have been employed. In addition to the resource selection such as corpora and dictionaries, determination of ambiguous words and their sense classification, decision of effective features and algorithms, and evaluation criteria are the major parts of a WSD system that have to be considered.

In this study, the facts mentioned above have been exhaustively scrutinized in the context of Turkish WSD application. Ambiguous verbs and nouns along with their sense classifications have been established and by providing manually sense tagged corpora and examining WSD problem from various perspectives, an important contribution has been achieved for the researches in Turkish. In Section 2 some general issues in Turkish WSD has been explained along with the experimental setup and the results. In the last section, a general evaluation and conclusion have been provided for commenting on the results and future work.

2 Issues in Turkish WSD

English and very few other languages have been widely studied in NLP researches. Lesser studied languages, such as Turkish suffer from the lack of wide coverage electronic resources or other language processing tools. The limitations effect the development of new applications and implementation of new ideas. In WSD studies, the methods benefit from the annotated corpus, therefore, the ambiguous words and their context heavily depend on the corpus. Dictionaries or word nets are other important categories of resources that provide invaluable information, however obtaining the appropriate set of senses and context-sense relations may not be so trivial.

Effective features for WSD may also vary for different languages and word types. Although, some features are common in many languages some others may be language specific. Turkish is based upon suffixation, which differentiates it sharply from the majority of European languages, and many others. Like all of the 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. Whenever the verbs are the main concern, the effective features must be searched in the previous context; on the other hand the nouns may be affected from all neighbouring features. In addition to the syntactic clues that can be obtained from the morphological analysis and parsing, semantic relations that are generally language specific and difficult to acquire may be the bottleneck in WSD applications.

The evaluation task is really very complicated. Many different applications report contradictory results in the literature [Da02]. Standard evaluation criteria are definitely needed; however the evaluation techniques are still being discussed and improved especially in the Senseval Project [Ed02].

2.1 Turkish Corpora

Turkish language processing has been developed in recent years. However, some of the applications do not have a broad coverage or some others are not open to public. This fact limits the selection of appropriate resources for Turkish NLP tasks. There are some projects for providing data for NLP applications in Turkish like METU Corpus Project [Of03]. 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. It has been preferred to use the Treebank part since it fits far or less the purposes of this study.

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. The distribution of the texts in the Treebank is similar to the main corpus. There are 6930 sentences in this Treebank. In Turkish, a word can have many analyses, so having disambiguated texts is very important. There are 5356 different root words and 627 of these words have 15 or more occurrences, and the rest have less. Therefore, most of the root words are so rare and not suitable for WSD experiments.

2.2 Selection and Sense Classification of Ambiguous Turkish Verbs and Nouns

The average number of senses for Turkish words can be significantly high leading to many problems for sense classification. The set of senses that are listed in many Turkish dictionaries are not generally providing a suitable sense list for WSD applications. Additionally, they may have some inconsistencies and rather than providing a sense classification for ambiguous words, they do list some usages of the words that may have overlaps in the definitions. Although this step is crucial and has to be considered in the very early stages of a WSD system, most of the applications do suffer from the lack of appropriate sense classification.

The set of senses that have been used for Turkish words have been obtained after considering many different approaches. In the first trial, the senses in some broad coverage Turkish dictionaries [Tu95, TDK05] and the usages in the Treebank have been considered for manual construction of the sense set. However, the end product of this study can still have many senses and it does not fit to practical applications for many words. In the next approach, translations have been considered. Nevertheless, this approach has its own disadvantages. First of all, this set may not be a suitable one for applications other than the translation task. The words can be mapped to different classes whenever many different languages have been considered. Besides these, the word can have language specific usages that may not be classified in another language. The final decision about the set of senses for the ambiguous words has been set up by considering all the approaches mentioned above and sense tagging has been achieved by using fine-granular (FG) and coarse-granular (CG) senses. Selected words and their senses are given in Table 1.

Table 1. Selected words along with the number of senses for FG and CG senses. The meanings of the words are not provided due to the large number of senses that are applicable.

Verbs	FG # senses	CG # senses	#instances	Nouns	FG # senses	CG # senses	#instances
al	30	6	265	**ara**	10	7	136
bak	11	5	185	**baş**	9	5	102
çalış	6	2	101	**el**	6	5	157
çık	28	7	238	**göz**	8	6	111
geç	19	8	146	**kız**	4	2	89
gel	26	3	298	**ön**	10	3	68
gir	15	4	134	**sıra**	5	2	54
git	17	6	197	**yan**	8	4	96
				yol	10	5	88
				yüz	6	6	61

2.3 Effective Features for WSD in Turkish

Exploring effective features for WSD has been attracted considerable interest in the literature. It has been stated that the appropriate sense of an ambiguous word can be successfully selected whenever N words in the neighborhood have been considered [We55]. Although this point of view reflects a part of the fact, it is highly simplifying the task. Disambiguation process can not be a function solely depending on the N words and many researches about the other effective features have been carried out. The ones that are included in [NL96] are surrounding words, local collocations [NL96] syntactic relations, POS and morphological forms [FGL98, Ng97].

The features are selected by considering three basic word groups in Turkish: The previous and subsequent words that are related with the target word and the ambiguous word itself. The features for these groups include root, POS, case marker, possessor and word's relation with the next word.

2.4 Machine Learning Algorithms

Machine learning techniques have been widely applied to NLP tasks. WSD utilize these methods in different researches. Bayesian probabilistic algorithms [Mo96], neural networks [Mo96], decision trees [Ya00], instance based learning (alternatively memory based or exemplar based learning) [Ng97] etc. are the most frequently used methods in this domain.

There is a system so called WEKA developed at the University of Waikato in New Zealand. It includes many famous machine learning algorithms. The system provides many visualization tools and a detailed analysis of the output [WF99]. Selected features and their combinations that are thought to be effective in Turkish have been tested by some famous machine learning algorithms. These algorithms, AODE (Aggregating one-dependence estimators) which is an improved version of famous Naive Bayes statistical method, IBk (Instance-based learning) where k is 1 and J48 decision tree method were tested by using Weka. The average performance results for FG and CG sets by using AODE, IBk and J48 algorithms are provided in Table 2. In the evaluation k-fold cross validation has been used where k=10. IBk [AK91] is one of the best performing algorithms among many other alternatives tested on the Turkish WSD problem, therefore performance results are given for IBk in Table 3 and Table 4. Average, maximum and minimum accuracy values in percentages by using different feature combinations are provided. Baselines are the percentage of the most frequently used senses of the words in the corpus. Average net gain is the difference between average accuracy and the baseline. The effective features are generally found in the previous context for the verbs, such as the root word, case marker and the relationship. However the nouns are also effected from the features (POS, case marker and relationship) of the target word itself and the subsequent context (root word and POS) in addition to the previous context (root word and POS). In some cases all features together outperform the rest.

Table 2. Average performance results (% accuracy) for FG and CG senses by using AODE, IBK and J48 algorithms for different POS.

	FG			CG		
POS	**AODE**	**IBK**	**J48**	**AODE**	**IBK**	**J48**
Verbs	43	46	45	61	63	60
Nouns	70	73	66	81	82	75

First of all, selection of the corpus has a direct impact on the results. The words are selected considering the frequencies in the corpus. Furthermore, the set of senses, especially the coarse-granular ones, are also determined by considering the corpus. Some senses rarely or do not occur in the corpus. In addition to this, the distributions of the senses are not uniform. Whenever the senses (or at least some of them) occur equally likely, in other words if one sense is not dominant, then the net gains increase. Therefore, although the results of coarse-granular task are better than the fine-granular one, actually they are worse whenever the net gains are considered. These facts emphasize that the resource selection, distribution of the words and their senses and the granularity level are in some way or another effective in the WSD. Furthermore, interpreting the results depending on the accuracy or average values may be misleading and one has to understand the results by considering the number of senses, baselines and net gains. The size of the feature set is not proportionally increasing the performance. In other words having many features does not mean having better performance. Irrelevant features may cause side effects and may decrease efficiency. Finding optimal set of features will be crucial and it may not be the same set for all words and for all languages. However, the impact of selected features is more significant than the algorithms.

Table 3. Performance results (% accuracy) of verbs for FG and CG senses by using IBk algorithm vs. various features

FG	al	bak	çalış	çık	geç	gel	gir	git
Average	26	62	45	23	32	45	50	57
Baseline	14	60	31	15	24	40	46	55
Average Net Gain (Average-Baseline)	12	2	14	8	8	5	4	2
Maximum	38	65	54	36	40	51	55	58
Minimum	15	60	26	15	24	40	44	56
CG	**al**	**bak**	**çalış**	**çık**	**geç**	**gel**	**gir**	**git**
Average	53	65	70	52	44	71	61	75
Baseline	47	64	66	47	35	67	58	74
Average Net Gain (Average-Baseline)	6	1	4	6	9	4	4	1
Maximum	61	69	76	61	50	77	66	76
Minimum	46	62	64	46	36	67	57	73

Table 4. Performance results (% accuracy) of nouns for FG and CG senses by using IBk algorithm vs. various features

FG	ara	baş	el	göz	kız	ön	sıra	yan	yol	yüz
Average	35	38	70	71	68	49	66	42	53	75
Baseline	20	27	67	64	60	45	57	35	43	63
Average Net Gain (Average-Baseline)	15	11	3	7	8	4	9	7	10	12
Maximum	57	71	82	81	83	61	91	62	72	94
Minimum	16	16	70	68	56	46	54	30	40	62
CG	**ara**	**baş**	**el**	**göz**	**kız**	**ön**	**sıra**	**yan**	**yol**	**yüz**
Average	46	59	73	77	86	81	71	54	66	75
Baseline	30	57	69	76	86	83	60	49	64	63
Average Net Gain (Average-Baseline)	16	2	4	1	0	-2	11	5	2	12
Maximum	70	83	87	86	100	92	93	77	82	94
Minimum	29	52	72	75	83	79	54	43	61	62

3 Conclusion And Future Work

The study on Turkish verbs and nouns demonstrates that there are many different factors on sense disambiguation process. Despite examining a small set of words, the results point out some important clues about the ambiguity problem. These clues can be used in order to have successful semantic applications in various domains for Turkish.

We are planning to increase our test data for the future studies. We also intend to develop a basic ontological classification of the words in the corpus. This ontological structure will be added to our algorithm as one or two extra features. Thus, our corpus based approximation presented in this study will be combined with a new estimation, which is a knowledge-based taxonomy deriving a hybrid approach.

References

[AK91] Aha, D.; and Kibler, D.: Instance-based learning algorithms, Machine Learning (6) 1991; S. 37-66

[Da02] Daeleman, W.: Machine Learning Of Language: A Model And A Problem, ESSLLI'2002 Workshop On Machine Learning Approaches In Computational Linguistics, August 5-9, Trento, Italy, 2002, S. 134-145

[Ed02] Edmonds, P.: SENSEVAL: The evaluation of word sense disambiguation systems, *ELRA Newsletter*, Vol. 7 No. 3, 2002, S. 5-14

[FGL98] Fellbaum, C.; Grabowski, J.; Landes, S.: Performance and confidence in a semantic annotation task, In Fellbaum, C. (ed.), WordNet: An Electronic Lexical Database Cambridge (Mass.), The MIT Press, 1998.

[IV98] Ide, N.; Veronis, J.: Introduction To The Special Issue On Word Sense Disambiguation: The State Of The Art, CL., 24(1), 1998, S. 1-40.

[Ka96] Kazakov, D.: Natural Language Processing Applications Of Machine Learning, Ph.D. Thesis, Czech Technical University, Prague, 1996.

[KS75] Kelly, E.; Stone, P.: Computer Recognition of English Word Senses, NAvg.h Holland, Amsterdam, 1975.

[Mo96] Mooney, R. J.: Comparative Experiments On Disambiguating Word Senses: An Illustration Of The Role Of Bias In Machine Learning, Proceedings Of The Conference On Empirical Methods In Natural Language Processing, Association For Computational Linguistics, Somerset, New Jersey, 1996, 82–91.

[NL96] Ng, H.T.; Lee, H.B.: Integrating Multiple Knowledge Sources To Disambiguate Word Sense: An Exemplar-Based Approach. In Proceedings Of The 34th Annual Meeting Of The ACL, Santa Cruz, In Arivind Joshi ve Martha Palmer, Editors, San Francisco, Morgan Kaufmann Publishers, 1996, S. 40–47.

[Ng97] Ng, H. T.: Exemplar-Based Word Sense Disambiguation: Some Recent Improvements, In Procs. Of The 2nd Conference On Empirical Methods In Natural Language Processing, EMNLP, 1997, S. 37-49

[Of03] Oflazer, K.; Say, B.; Tur, D. Z. H.; Tur, G.: Building A Turkish Treebank, Invited Chapter In Building And Exploiting Syntactically-Annotated Corpora, Anne Abeille Editor, Kluwer Academic Publishers, 2003.

[Tu95] Tuğlacı, P.: Okyanus Ansiklopedik Türkçe Sözlük, ABC Kitabevi Yayın ve Dağıtım AŞ, İstanbul, 1995.

[TDK05] TDK, Türkçe Sözlük(Turkish Dictionary): TDK, 2005 http://tdk.org.tr/sozluk.html .

[We55] Weaver, W.: Translation, Mimeographed, (1949) Reprinted In Locke, William N., Booth, A. Donald, (Eds.), Machine Translation Of Languages, John Wiley & Sons, New York, 1955, S. 15-23

[WF99] Witten, I. H.; and Frank E.: DataMining: Practical Machine Learning Toolsand Techniques with Java Implementations,Morgan Kaufmann, San Francisco, 1999.

[Ya00] Yarowsky, D.: Hierarchical Decision Lists For Word Sense Disambiguation, Computers And The Humanities, 34(2), 2000, S. 179-186.