Journal of Computing and Security

April 2014, Volume 1, Number 2 (pp. 123–132)

http://www.jcomsec.org

Word Sense Disambiguation Based on Lexical and Semantic Features Using Naive Bayes Classifier

Amir Hossein Rasekh [a,*]
Mohammad Hadi Sadreddini [a]
Seyed Mostafa Fakhrahmad [a]

[a] *Computer Science and Engineering Department, Shiraz University, Shiraz, Iran.*

ARTICLE INFO.

Article history:
Received: 28 January 2013
Revised: 18 November 2013
Accepted: 02 February 2014
Published Online: 13 September 2014

Keywords:
Natural Language Processing, Word Sense Disambiguation, Text Mining

ABSTRACT

Machine translation is considered as a branch of machine intelligence with about fifty years background. Ambiguity of language is the most problematic issue in machine translation systems, which may lead to unclear or wrong translation. One of the problems involved in *natural language processing* is the semantic and structural ambiguity of the words. The objective of this paper to focused on the *word sense disambiguation*. In here, the existing algorithms for word sense disambiguation are evaluated and a method which is proposed based on the concept, structure and meaning of the words. The experimental results are promising and indicate that this proposed approach significantly outperform its counterparts in terms of disambiguation accuracy.

1 Introduction

Machine translation is a branch of computational linguistics where a machine is employed to understand a text completely and then translate it into another language. By using corpus linguistics techniques, translation of more complex texts is facilitated. Furthermore, these techniques provide a strong way in order to make better recognition and translation of different phrases and terminologies of almost all languages.

Machine translation includes some features which justify its essence it economic perspective. For instance, in spite of working about 1,200 translators at NATO National Headquarters in Brussels and Europe, machine translation is applied in order to increase the speed and decrease the cost of translation. Moreover, translation of text using a machine is faster than a human translator; although, the quality and accuracy of machine translation is lower than the human translator.

Machine translation is a very active topic in the field of machine intelligence. The first experience of the kind was performed on a Russian text to English. Despite many studies already conducted on machine translation, there exist several defects. It is obvious that translation of any text (simple/complex) using a machine is not done easily. Ambiguity in semantic and structure of the language is the most problematic issues in machine translation. One of the most important tasks in this process is to recognize the role of a word in a sentence and then translate it to its correct meaning in the target language.

Disambiguation requires considering semantics, syntax and word processing issues. Unfortunately, there is a big gap between the theoretical linguistics and its application. Theoretical linguistics has failed to present a general model that covers all aspects of the

* Corresponding author.
Email addresses: ahrasekh@shirazu.ac.ir (A.H. Rasekh), sadredin@shirazu.ac.ir (M.H. Sadreddini), mfakhrahmad@shirazu.ac.ir (M. Fakhrahmad)
ISSN: 2322-4460

language.

A major ambiguity is resolved by using a technique known as *"Part Of Speech Tagging"* which is defined as determining and annotating the role of all words in the sentence [1]. In other words, the method specifies the right role of every word in a sentence, which would facilitate the determination of the correct meaning of the word. Note that the word sense disambiguation (WSD) is more difficult than the structural ambiguity resolution, and a stronger method is required to solve this problem. Actually, word sense disambiguation is carried out by identifying the sentence which contains the ambiguous word followed by recognition of the words surrounding the ambiguous word.

All word sense disambiguation (WSD) methods already proposed can be classified in three main categories. Supervised methods which apply a sense-tagged dataset to train a classifier, applying lexical resources such as dictionary or thesaurus; and the unsupervised methods which work on unlabeled set of words and texts [2, 3].

1.1 Disambiguation Using Supervisor

In this approach, a set of texts containing the ambiguous word are used as the training data. The correct sense of each word is also annotated in each text, during the training phase.

This approach can be considered as a classification task on the words. In other words, the algorithm is trained in order to classify a new unseen case of the ambiguous word by determining its correct sense. In supervised learning approach, different well-known classifiers are applied including *Bayesian Network* [4], *Information Theory* [5], *Decision Tree* [6], *K-Nearest Neighbour* [7] and *Neural Network* [8].

1.2 Dictionary-Based Disambiguation

This method is used when no information about the class of the word in the corpus is available. In such cases, the general information about the word is explored from the dictionary. Here, different types of information are used: the first type is called *Lesk* which uses exactly the meaning of the word which is defined in the dictionary; in the second type, the algorithm uses the classification information of the word which is defined in the dictionary; and in the last type, the information about the translation obtained from a dictionary of two different languages is used [9].

1.3 Disambiguation Using Unsupervised Learning

In two previous disambiguation methods, the procedure needs some prior knowledge about the semantic of the words, while in some cases there is no information about the meaning of the words included in the text [2, 5]. The first step in this approach is to cluster the words based on semantic information. Each cluster has a semantic, and each word is assigned to the nearest cluster. Here, the word takes the meaning of the cluster in the sentence.

The objective of this study is to disambiguate words based on their concept, structure and meaning. The rest of the paper is organized as follows: in Section 2, the related work is reviewed, Section 3 is devoted to introduction of this proposed method. The experimental results are presented in Section 4. Finally, the paper is concluded in Sections 5.

2 Related Works

The set of WSD methods can be divided into three major categories, supervised, unsupervised and hybrid schemes. The first category includes methods which are based on supervised learning. These methods use classification systems to determine the correct translation of ambiguous words. The second category includes methods that use unsupervised learning. Text clustering is the main learning process used by the methods in this category. There is also another category of disambiguation methods which propose a combination of supervised and unsupervised learning.

2.1 Supervised Approaches

There are a lot of proposed methods for word sense disambiguation which follow the supervised learning techniques, e.g., Naive Bayesian [10], Decision List [6], Nearest Neighbour [7], Transformation Based Learning [11], Winnow [12], Boosting [13], SVM [14] and Naive Bayesian Ensemble [15]. Among the mentioned methods, the methods that apply SVM and Naive Bayesian Ensemble have been reported to have the best performance for ambiguity resolution tasks, respectively. In order to determine the correct translation of each ambiguous word, all of the above methods construct a classifier, using features that represent the context of the ambiguous word.

Brown et al. proposed a corpora-based disambiguation method which can be applied in machine translation systems. They use data from syntactically related words in the local context of the ambiguous word. In order to obtain statistical data, a word-aligned bilingual corpus is required. Each occurrence of an ambigu-

ous word should be labeled with a sense by asking a question about the context in which the word appears [16].

In the other method proposed by Yarowsky et al. the assumption is that each word is located in a major category. In order to disambiguate word senses they have used the Roget's Thesaurus dataset. By searching the hundred surrounding words as indicators of each category, the most probable category of a word can be determined. During the training phase, a stemming process is performed over all words in order to achieve more useful statistics. Subsequently, by examining the hundred surrounding words for indicators of each category, the indicator words are obtained and weighted. The measure used as the weight of each indicator word is the log of words salience as shown in 1.

$$weight(w\ in\ cat) = log\frac{P(w|cat)}{P(w)} \qquad (1)$$

Where, w is an would word and cat stands for a category. $P(w|cat)$ is the probability that w appear in the context of a word from the category cat and P(w) is the probability of the was occurrence in the corpus as a whole. For the useful words, the computed weight, i.e., the log of salience will be greater than one [17].

The system proposed by Yarowsky et al. is not limited to particular word categories and works in a wide domain. The first drawback of the system is that it cannot disambiguate topic-independent distinction words that occur in many topics. Moreover, the system does not consider the distance of words in the contexts it handles.

Another method for word sense disambiguation was proposed by Dagan and Itai, where the most probable sense of a word using frequencies of the related word combinations in a target language corpus was chosen. In their method, first of all, the system identifies syntactic relations between words using a source language parser and maps those relations to several possibilities in the target corpus using a bilingual lexicon [18].

Justeson and Katz, uses syntactically or semantically relevant clues. This method disambiguates adjectives using only nouns that are combined by the adjectives. The system was evaluated on five of the most frequent ambiguous adjectives in English: 'right', 'hard', 'light', 'old', and 'short' on large sets of randomly selected sentences from the corpus that contained the adjectives. However, for adjectives which can be differently accompanied by the same noun, this method cannot contribute to disambiguation [19].

The system presented by Ng and Lee is based on the Nearest Neighbor method. The prototypes are the instances of the ambiguous word in the training corpus, each containing the following features: singular/plural, POS tags of the current word, three words on either side, support for verbs which have a different verbal morphological feature, a verb-object syntactic feature for nouns, and nine local collection features. These features are calculated for each instance of w in the sense-tagged training data. The results are stored as exemplars of their senses. By calculating the same feature vector for the current word and comparing all the examples of that word, the given word is disambiguated choosing the closest matching instance [7].

The method presented by Brown et al. requires a bilingual word-aligned corpus, which is costly to build. This is one of the drawbacks of this method, which reduces the applicability of the method to other pairs of languages.

The method proposed by Mosavi and Khalafi is somewhat similar to that of the Dagan and Itai which uses a target language model. They use Persian as the target language and consider the co-occurrences of the multiple-meaning words in a monolingual corpus of the Persian language. By calculating the frequencies of these words in the corpus, the most probable sense for the multiple-meaning words is chosen. However, instead of considering syntactic tuples in the target language corpus, they consider just co-occurrences of certain words in that corpus without having a syntactic analysis for the corpus. In this method, no analysis is conducted either on the source or the target language corpus with respect to syntactic. The only task of the proposed algorithm, for obtaining the required statistical information, is performed by determining the true nature of the nearest noun, pronoun, adjective, or verb to the ambiguous word. When applying this method for the comparison of English and Persian machine translation, only a small portion of ambiguous words in English can be correctly translated into Persian [20].

Reddy et al., investigated the WSD by modeling it in a distributed constraint optimization (DCOP) framework. The method requires information from various knowledge sources (including part-of-speech, morphology, domain information, etc), in order to be modeled in a multi-agent setting [21].

Rezapour et al. proposed a supervised learning method for WSD based on K-Nearest Neighbour algorithm. This method extracts two sets of features, including the frequently occurred words, and the set of words surrounding the ambiguous word. In order to improve the classification accuracy a feature weighting strategy is introduced and used [22].

2.2 Unsupervised and Semi-Supervised Approaches

In addition to supervised approaches, unsupervised approaches and their combinations are proposed for word sense disambiguation.

The proposed approaches in unsupervised WSD can be divided into some major categories, which are based on context clustering [23], word clustering [24] and co-occurrence graphs [25], respectively. Among these main approaches, the graph-based methods is recently explored with a certain success. These approaches are based on the notion of a co-occurrence graph. A co-occurrence graph is a graph like G = (V, E), where its vertices V correspond to words in a text and edges E connect pairs of words which co-occur in the same context.

For example, Veronis et al. proposed an ad hoc approach called HyperLex. In the First step, a co-occurrence graph is drawn where the nodes are the words that have occurred in the paragraphs of a corpus in which a target word occurs, and an edge between a pair of words is added to the graph if they co-occur in the same paragraph. A weight is then assigned to each edge, according to the co-occurrence frequency of the pair of words connected by the edge [26].

Sinha and Mihalcea proposed an unsupervised graph-based word sense disambiguation algorithm, which combines several word semantic similarity (i.e., The degree to which two words are semantically related) measures and algorithms for graph centrality (i.e., the algorithms which assign different important degrees to the nodes). That work was the first in addressing the problem of WSD by comparatively evaluating measures of word semantic similarity in a graph theoretical framework [27].

The method proposed by Reddy et al. uses the Personalized PageRank which is a graph centrality algorithm (Agirre and Soroa 2009) over a graph representing WordNet to disambiguate ambiguous words by taking their context into consideration [28].

There are some other unsupervised methods in the literature, which are based on word clustering or context clustering, such as Schutze [23] proposed an ambiguity resolution technique which divides the occurrences of a word into a number of clusters by determining any two occurrences as whether they belong to the same sense or not, which is then used for the full ambiguity resolution task. The approaches proposed by Litkowski [24] and Lin [29] are other examples of unsupervised learning methods. Nigam et al. [30] proposed an unsupervised learning method using the Expectation-Maximization (EM) algorithm for text classification problems, which then was improved by Shinnou and Sasaki in order to apply it to the ambiguity resolution problem [31]. Agirre et al. combined both supervised and unsupervised lexical knowledge methods for word sense disambiguation [32]. In two other methods, Yarowsky [33] and Towell and Voothees, the rule-learning and neural networks are applied, respectively [8].

Galley and McKeown proposed a method consisting of two stages; first, a graph is drawn representing all possible senses of the words under investigation, represented as graph nodes and their semantic correlation are the weights assigned to their connecting edges; second the text is processed sequentially by comparing each word against all words previously read. If a relation exists between the senses of the current word and any possible sense of a previous word, a connection is established between the appropriate words and the senses. The weight assigned to each connection is a function of two factors, i.e., the type of relationship and the distance between the words in the text. For each sense of an ambiguous word, the weights of all connections corresponding to that sense are summed, giving that sense a unified score. The sense with the highest unified score is finally chosen as the correct sense [34].

Mihalcea and Moldovan, proposed an iterative method for WSD. This method differs from other proposed methods in that it follows an iterative process for WSD (same as the method proposed in this paper). This method requires two sources of information; WordNet and a semantic-tagged corpus. The method which will be proposed in the this paper is similar to the method proposed by Mihalcea and Moldovan in that it also disambiguates the words, iteratively. However, here, no additional information source is necessary; and this method works with a set of discovered association rules and knowledge deduced from the rules. Moreover, as the results of the experiments indicate, the response times of these two methods are quite different [35].

3 The Proposed Method

The key task in WSD is to explore the context of the ambiguous word in order to find some of its unique characteristics. For this purpose, there exist some proposed approaches which are described as follows.

The proposed method is a supervised method, where word sense disambiguation is carried out based on concept, structure and meaning of the words. In this study, the *WordNet* as an extra knowledge base to discover the relationships among the target word and its surrounding words in the context. Based on the gathered information the Naive Bayes classifier is

applied in order to determine the correct sense of the word. *Naive Bayesian* (NB) classifier is known as one of the most suitable methods for supervised approaches in WSD [36]. Naive Bayes for Supervised WSD are collected in feature vectors. A feature vector consist of numeric or nominal values to encode linguistic information as input to most Data Mining algorithms that classes of feature vectors extracted is collocation feature vectors. A collocation is a word or phrase in the position of a specific relationship to a target word.

The proposed method is a hybrid method which makes use of different features and techniques, which will be described in the following sections.

3.1 WSD Using Lexical and Structural Features

3.1.1 N-grams

An N-gram is a continuous sequence of N items, i.e., syllables and words, from a given text and can be considered as the composition of words. The size "1" for the N is "unigram"; size "2" is a "bigram"; size 3 is a "trigram"; larger sizes are mentioned to by the value of N, e.g., "4-gram", "5-gram", and so on. Usually, N-gram is collected from a text or speech corpus.

According experiments run by this authors, the best results are obtained for $N = \pm 5$ where the gram separates five items from the left and the right sides of the ambiguous words in each paragraph. For selection of N surrounding words here, prepositions and numbers are ignored and then all of the selected words are classified using the Naive Bayes algorithm.

3.1.2 Word Frequencies

In this method, word frequencies are used instead of the words themselves. For this purpose, first using 5-grams, each word in the text is replaced with its frequency value. Next the Naive Bayes algorithm is applied to all numbers in order to classify them. In this method if the frequencies of two words are exactly equal, one is selected on a random basis. the disadvantage of this method is that it is faced with a large scale sparse matrix.

3.1.3 Word Stems

This proposed method uses the word stems instead of the words. For instance, the term *studying* is replaced by its stem *study*.

3.1.4 Feature Weighting

Here, each are of the surrounding word found by 5-gram takes a weight. The weights are obtained through the two approaches: first, according to the nearest and the furthest surrounding word and, second, according to the frequency of each word. It means that the nearest word takes the highest weight while the farthest one take; the lowest weight. In this method, the value considered for the nearest word is 0.5, while the value of the farthest word is 0.05. The occurrence frequency also takes part in the weighting metric. That is, the furthest word with a high frequency takes a high weight. After the weighting phase, words are classified using the Naive Bayes classifier.

3.2 WSD Through Conceptual and Meaning-based Features

3.2.1 Keywords

Automatic extraction of keywords is evolved to identify a small set of words or key phrases from the document such that the set can describe the concept of the document. This task should be performed automatically or with minimal human intervention [14].

In this study, the words that have the highest frequency and their occurrences are several times more are compared with the Brown corpus are determined as keywords.

Brown corpus is introduced by Kucera and Francis at Brown University in the USA. This corpus includes one million words which is equal to 500 texts, where each text includes 2000 words in a wide variety of topics in English.

For extraction of the keywords, each arbitrary text is compared with the Brown corpus and the words with the highest frequency will be selected as the keywords of that text.

An example here will reveal the problem. For instance, this algorithm outputs the 167.00 for the word *step* from a text with 92 words. For calculation of this value, the word *step* in Brown corpus and the arbitrary text has 130 (of one million words) and 2 (of 92 words) occurrences, respectively. Taken as a proportion of 1,000,000 words, these 2 occurrences represent 2/92 * 1,000,000 = 21739 virtual occurrences. These 21739 occurrences are 167.22 times more than the 130 occurrences in Brown.

In other words, the occurrences of each word in an arbitrary text are evaluated according to the frequency of this word in the Brown corpus. For this purpose, the algorithm takes a paragraph of the text and extracts keywords of the paragraph and then the Naive Bayes algorithm creates a model for classification of the words.

3.2.2 Semantically Related Words

In order to find the words which are semantically related to the target word, the wordnet which is a lexical database for English words is used here. WordNet can be considered as a combination of lexicon and thesaurus. WordNet made in the Cognitive Science Laboratory at Peterson University in order to make semantic connections between the words.

In this method, using semantic hierarchies of words available in WordNet, the words which have close hierarchical distance to the target word are considered for disambiguation. The keywords of a paragraph and the surrounding words of the ambiguous word are then extracted. Finally, a model is developed based on all extracted words using the Naive Bayes classification method. The collocation of feature vector for Naive Bayes, extracted from a window of related word, keywords and n-gram.

The steps defined for this proposed scheme are as follows:

(1) The ambiguous word is passed into the WordNet.

(2) The words which are related to the target word, from different levels of the WordNet are extracted and a weight is assigned to each one of them.

(3) A paragraph is selected for disambiguation. The words extracted in the previous step are explored in this paragraph and then the available ones are added to the features of the paragraph.

(4) The keywords of the paragraph are extracted using the algorithm described in section 3.2.1 and then, all of them are added to the features of that paragraph.

(5) The 5-gram is applied to find 5 words before and after the ambiguous word. All of the surrounding words are added to the features of that particular paragraph.

(6) The steps 1-5 are repeated for all paragraphs of the text. Next, a vector of features is obtained for each paragraph and finally, the correct meaning of the ambiguous word is used as the class of the paragraph.

(7) The Naive Bayes classifier (which is a Supervised method) is used over the prepared data.

(8) Finally, a new unseen paragraph is passed to the created model, and the obtained results are compared with the real class of the data.

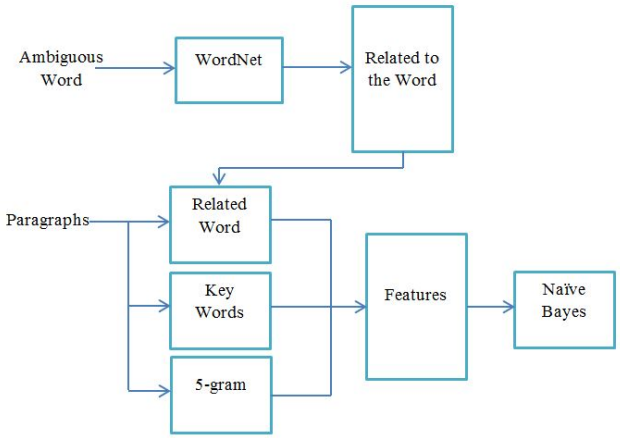

Figure 1. Algorithm Steps

4 Experimental Results

In order to evaluate this proposed scheme, the *two-way ambiguities* (TWA) dataset [1] is applied. TWA is a standard dataset for identification of the ambiguous words. This dataset is collected at the University of North Texas by Mihalcea and Yang. The dataset contains some ambiguous words and the main focus of the TWA is on six different words including *"bass"*, *"crane"*, *"motion"*, *"palm"*, *"plant"* and *"tank"* that the results shown in Table 1.

The dataset is divided into two subsets, training and test (unseen) sets. Using 10-fold cross validation approach, the size of the training set in each one of the iterations is 90% of the whole data and the remaining data is applied for test.

The results shown in Figure 2 are obtained when only WordNet is used; however, in order to improve the prediction accuracy of the method; a combination of features including WordNet relations, keywords and the words around the ambiguous word are used. The accuracy of the proposed methods (i.e., 78.2%) compared to some available methods in the literature are illustrated in Figure 3.

As shown in Figure 3, the proposed method which uses all the propose features outperforms the previous methods in terms of accuracy.

5 Conclusion and Future Works

In this paper, several methods in resolving the semantic and structural ambiguities of words in a text are assessed a new is proposed. One of the best results obtained from the combination of WordNet, the keyword extraction algorithm, and surrounding words of an ambiguous word in the text is evident here. A

[1] http://www.cse.unt.edu/rada/downloads.html/#twa

Table 1. Accuracy of the proposed methods for ambiguous words

Words	N-Gram	Word Freq	Word Stems	Weighting	Key Words	Depended word
Bass	51.85%	57.14%	52.73%	61.9%	75%	85.49%
Crane	63.15%	60.1%	65%	66.9%	80.43%	87%
Motion	47.6%	40%	49.1%	56.4%	66%	74.32%
Palm	68.12%	62.5%	70%	75%	81.25%	84.21%
Plant	53.57%	60.53%	55%	57.89%	62.75%	67%
Tank	60%	50%	62.12%	64.05%	65.71%	71.03%

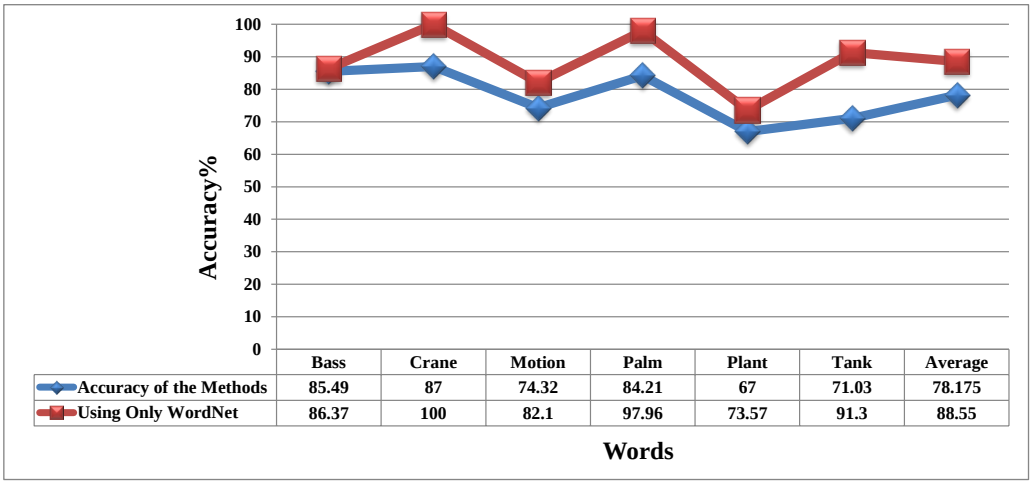

Figure 2. Results of this Proposed Method

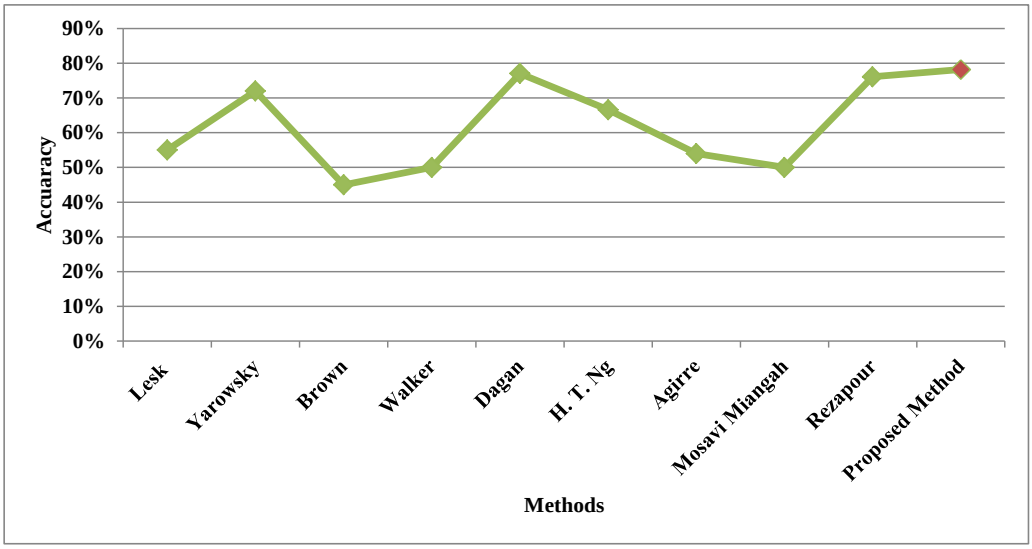

Figure 3. Accuracy Values Compared to Previous Methods

combination of the methods is applied here in order to find the whole set of features applicable for disambiguation. Finally, a model is developed according to the extracted features, where the Naive Bayes algorithm is applied to classify the prepared data. In this paper, the TWA dataset as a benchmark is applied to evaluate this method. The average accuracy of the proposed algorithm is 78.2%. The experimental results demonstrate that the proposed algorithm is very promising in terms of accuracy of disambiguation.

For future; the part of speech tags (POSs) of the words, (i.e. verb, noun, adj, adv, etc.) for word sense disambiguation can be of consideration. Using surface scrolling in WordNet is another solution to increase the accuracy of disambiguation.

Acknowledgements

The authors would like to thank Messrs. *Mohammad Hassan Dianaty* and *Hossein Taghi-Zadeh* for their collaboration in this paper.

References

[1] Walter Daelemans, Jakub Zavrel, Peter Berck, and Steven Gillis. Mbt: A memory-based part of speech tagger-generator. *CoRR*, cmp-lg/9607012, 1996. URL http://dblp.uni-trier.de/db/journals/corr/corr9607.html#cmp-lg-9607012.

[2] Yorick Wilks and Mark Stevenson. Word sense disambiguation using optimised combinations of knowledge sources. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 2*, ACL '98, pages 1398–1402, Stroudsburg, PA, USA, 1998. Association for Computational Linguistics. doi: 10.3115/980691.980797. URL http://dx.doi.org/10.3115/980691.980797.

[3] Chunyu Kit and Yorick Wilks. Unsupervised learning of word boundary with description length gain. In *CoNLL-99*, pages 1–6, Bergen, Norway, 1999.

[4] Guy De Pauw and Walter Daelemans. The role of algorithm bias vs information source in learning algorithms for morphosyntactic disambiguation. In *Proceedings of the 2Nd Workshop on Learning Language in Logic and the 4th Conference on Computational Natural Language Learning - Volume 7*, ConLL '00, pages 19–24, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics. doi: 10.3115/1117601.1117607. URL http://dx.doi.org/10.3115/1117601.1117607.

[5] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA, 1999. ISBN 0-262-13360-1.

[6] David Yarowsky. Decision lists for lexical ambiguity resolution: Application to accent restoration in spanish and french. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 88–95. Association for Computational Linguistics, 1994.

[7] Hwee Tou Ng and Hian Beng Lee. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 40–47. Association for Computational Linguistics, 1996.

[8] Geoffrey Towell and Ellen M Voorhees. Disambiguating highly ambiguous words. *Computational Linguistics*, 24(1):125–145, 1998.

[9] David Yarowsky. Word-sense disambiguation using statistical models of roget's categories trained on large corpora. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 2*, COLING '92, pages 454–460, Stroudsburg, PA, USA, 1992. Association for Computational Linguistics. doi: 10.3115/992133.992140. URL http://dx.doi.org/10.3115/992133.992140.

[10] William A Gale, Kenneth W Church, and David Yarowsky. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26(5-6):415–439, 1992.

[11] Lidia Mangu and Eric Brill. Automatic rule acquisition for spelling correction. In *ICML*, volume 97, pages 187–194, 1997.

[12] Andrew R Golding and Dan Roth. A winnow-based approach to context-sensitive spelling correction. *Machine learning*, 34(1-3):107–130, 1999.

[13] Gerard Escudero, Lluís Màrquez, and German Rigau. *Boosting applied to word sense disambiguation*. Springer, 2000.

[14] Masaki Murata, Masao Utiyama, Kiyotaka Uchimoto, Qing Ma, and Hitoshi Isahara. Japanese word sense disambiguation using the simple bayes and support vector machine methods. In *The Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 135–138. Association for Computational Linguistics, 2001.

[15] Ted Pedersen. A simple approach to building ensembles of naive bayesian classifiers for word sense disambiguation. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 63–69. Association for Computational Linguistics, 2000.

[16] Peter F Brown, Stephen A Della Pietra, Vincent

J Della Pietra, and Robert L Mercer. Word-sense disambiguation using statistical methods. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, pages 264–270. Association for Computational Linguistics, 1991.

[17] David Yarowsky. Word-sense disambiguation using statistical models of roget's categories trained on large corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 454–460. Association for Computational Linguistics, 1992.

[18] Ido Dagan and Alon Itai. Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, 20(4):563–596, 1994.

[19] John S Justeson and Slava M Katz. Principled disambiguation: Discriminating adjective senses with modified nouns. *Computational Linguistics*, 21(1):1–27, 1995.

[20] Tayebeh Mosavi Miangah and Ali Delavar Khalafi. Word sense disambiguation using target language corpus in a machine translation system. *Literary and Linguistic Computing*, 20(2): 237–249, 2005.

[21] Siva Reddy and Abhilash Inumella. Wsd as a distributed constraint optimization problem. In *Proceedings of the ACL 2010 Student Research Workshop*, pages 13–18. Association for Computational Linguistics, 2010.

[22] AR Rezapour, SM Fakhrahmad, and MH Sadreddini. Applying weighted knn to word sense disabiguation. In *Proceedings of the World Congress on Engineering*, volume 3, pages 6–8, 2011.

[23] Hinrich Schütze. Automatic word sense discrimination. *Computational linguistics*, 24(1):97–123, 1998.

[24] Kenneth C Litkowski. Senseval: The cl research experience. *Computers and the Humanities*, 34 (1-2):153–158, 2000.

[25] Dominic Widdows and Beate Dorow. A graph model for unsupervised lexical acquisition. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics, 2002.

[26] Jean Véronis. Hyperlex: lexical cartography for information retrieval. *Computer Speech & Language*, 18(3):223–252, 2004.

[27] Ravi Som Sinha and Rada Mihalcea. Unsupervised graph-basedword sense disambiguation using measures of word semantic similarity. In *ICSC*, volume 7, pages 363–369, 2007.

[28] Siva Reddy, Abhilash Inumella, Diana McCarthy, and Mark Stevenson. Iiith: Domain specific word sense disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 387–391. Association for Computational Linguistics, 2010.

[29] Dekang Lin. Word sense disambiguation with a similarity-smoothed case library. *Computers and the Humanities*, 34(1):147–152, 2000.

[30] Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using em. *Machine learning*, 39(2-3):103–134, 2000.

[31] Hiroyuki Shinnou and Minoru Sasaki. Unsupervised learning of word sense disambiguation rules by estimating an optimum iteration number in the em algorithm. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 41–48. Association for Computational Linguistics, 2003.

[32] Eneko Agirre, German Rigau, Lluis Padro, and Jordi Atserias. Combining supervised and unsupervised lexical knowledge methods for word sense disambiguation. *Computers and the Humanities*, 34(1-2):103–108, 2000.

[33] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 189–196. Association for Computational Linguistics, 1995.

[34] Michel Galley and Kathleen McKeown. Improving word sense disambiguation in lexical chaining. In *IJCAI*, volume 3, pages 1486–1488, 2003.

[35] Rada Mihalcea and Dan I Moldovan. An iterative approach to word sense disambiguation. In *FLAIRS Conference*, pages 219–223, 2000.

[36] Raymond J Mooney. Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 82–91. Philadelphia, PA., 1996.

Amir Hossein Rasekh received his Master's degree from Shiraz University at Computer Science and Engineering Department, Shiraz, Iran, in 2012. He is now a Ph.D. Student in Software Engineering at Shiraz University. His research interests include Natural Language Processing, Text Mining, and Data Mining.

Mohammad Hadi Sadreddini received his B.Sc. degree in Computer Science, M.Sc. in Information Technology and Ph.D. degree in Distributed Information Systems from University of Ulster in UK, in 1985, 1986 and 1991 respectively. He has been working in the department of Computer Science and Engineering at Shiraz University since 1993. His research interests include Natural Language Processing, Association Rules Mining, Bioinformatics, and Distributed Systems.

Seyed Mostafa Fakhrahmad was born in Shiraz, Iran in 1980. He received his B.Sc. degree in Computer Engineering (Software systems) from Kharazmi University of Tehran in 2003, and his M.Sc. degree in Computer Engineering (Artificial Intelligence) from Shiraz University in 2006. He received his Ph.D. in Computer Engineering (Software systems) from Shiraz University, In 2012. After graduation, he was employed as a faculty member in the Department of Computer Science and Engineering in Shiraz University. His research interests include Data Mining and Machine Learning, Fuzzy systems, Text Mining, Natural language Processing and Social Network Analysis. He has already published more than 60 papers in the mentioned areas in international journals and conference proceedings.