Applying Weighted KNN to Word Sense Disambiguation

A. R. Rezapour, S. M. Fakhrahmad and M. H. Sadreddini

***Abstract*—One of the major challenges in the process of machine translation is word sense disambiguation (WSD), which is defined as choosing the correct meaning of a multi-meaning word. Supervised learning methods are usually used to solve this problem. The disambiguation task is carried out using the statistics of the translated documents (as training data) or dual corpora of source and target languages. In this paper we present a supervised learning method for WSD, which is based on K-Nearest Neighbor algorithm. As the first step, we extract two sets of features; the set of words that have occurred frequently in the text and the set of words surrounding the ambiguous word. In order to improve the classification accuracy, we propose a feature weighting strategy. We will present the results of evaluating the proposed schemes and illustrate the effect of weighting strategies proposed. The results are encouraging comparing to state of the art.**

***Index Terms*— Machine Translation, Word Sense Disambiguation, Supervised approaches, K-Nearest Neighbor, Feature Weighting**

I. Introduction

WORD sense disambiguation (WSD) is an interesting topic for researchers and is an important technique for many NLP applications such as Information Retrieval, Machine Translation, and speech recognition and so on. WSD refers to the process of automatically identifying the correct meaning of an ambiguous word (i.e., a multi-meaning word) based on the context in which it occurs.

Word sense ambiguity can be thought of as the most serious problem in machine translation systems. The human mind is able to select the proper target equivalent of any source language word by comprehension of the context. A human being may also automatically consider a group of words, rather than just one word, in order to understand the meaning of a sentence, even if the words of the group are not relevant. In order to simulate this behavior in a machine, a huge amount of data will be required as input and the output may still not be free from errors.

Corpora-based approaches are usually proposed in order to resolve word sense ambiguities. In corpora-based Translation methods translations are generated on the basis of statistical or probabilistic models whose parameters are extracted from the analysis of a bilingual corpus. Statistical translation is based on the study of frequencies of various linguistic units, including words, lexemes, morphemes, letters, etc., in a sample corpus in order to calculate a set of probabilities, so that various linguistic problems such as ambiguity can be solved.

WSD algorithms can be broadly classified into three categories:

--**Supervised Approaches**: these approaches use machine-learning and data mining techniques to train a classifier from sense-tagged corpora. The success of supervised learning approaches to word sense disambiguation is largely dependent on the features used to represent the context in which an ambiguous word occurs..

--**Unsupervised approaches:** these approaches do not use a training corpus and are based on unlabeled corpora.

--**Semi-supervised approaches:** A hybrid of the two other categories.

In This paper, we present a WSD approach that is based on K-Nearest Neighbor (K-NN) algorithm. The proposed scheme is a supervised approach in which sense-tagged data is used to train the classifier.

At the first step, our approach extracts two set of features; the set of words that have Co-occurred with the ambiguous word in the text frequently, and the set of words surrounding the ambiguous word.

The main task performed by the disambiguation method is to assign a sense to an ambiguous word by comparing the context it has occurred in and the texts existing in the training corpus. After illustration of the K-NN approach, in order to improve the accuracy of the WSD method, some weighting schemes will be proposed and discussed.

The rest of this paper is organized as follows. Section 2 is devoted to the introduction of the related work in the literature. Section 3 illustrates the proposed system. Experimental results are presented in Section 4. Finally, Section 5 concludes the paper.

II. Related work

The set of all knowledge-based methods already proposed for ambiguity resolution can be divided into three major categories. The first category includes methods which are based on supervised learning. These methods use classification systems to determine the correct meaning of ambiguous words. The second Category includes methods that use unsupervised learning. Text clustering is the main learning process used by the methods included in this category. There is also another category of disambiguation methods which propose a combination of supervised and

A.R. Rezapour and M.H. Sadreddini are with the Dept. of Computer Eng. and IT, Shiraz University (e-mail: {Rezapour, sadredin} @Shirazu.ac.ir).

S.M. Fakhrahmad is with the Dept. of Computer Eng., Islamic Azad University, Shiraz branch, (corresponding author, phone: +98-9177038028; e-mail: mfakhrahmad@ cse.shirazu.ac.ir).

The partially financial support by the Iranian research institute for Information and Communication Technology (ICT) is kindly acknowledged.

unsupervised learning.

There are a lot of proposed methods for word sense disambiguation which follow supervised learning techniques, e.g., Naïve Bayesian [4], Decision List [5], Nearest Neighbor [6], Transformation Based Learning [7], Winnow [8], Boosting [9], and Naïve Bayesian Ensemble [10]. Among the mentioned methods, the method that uses Naïve Bayesian Ensemble has been reported to have the best performance for ambiguity resolution tasks with respect to data set used [10]. In order to determine the correct meaning of each ambiguous word, all of the above methods build a classifier, using features that represent the context of the ambiguous word.

Brown et al. (1991) proposed a corpora-based disambiguation method which can be applied in machine translation systems[11]. 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 ambiguous word should be labeled with a sense by asking a question about the context in which the word appears. The system was tested by translating 100 randomly selected Hansard sentences, each containing 10 words or less in length and obtained the accuracy of 45%.

In [12], Yarowsky et al. assumes that each word is located in a major category. In order to disambiguate word senses they have used the Roget's Thesaurus data set. 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, firstly, a stemming process is performed over all words in order to achieve more useful statistics. Then, by examining the hundred surrounding words for indicators of each category, the indicator words are obtained and weighted.

The system proposed in [12] is not limited to particular word categories and works in a wide domain. This system achieves accuracy of between 72% and 99%. The first challenge of the system is that it cannot disambiguate topic-independent distinction words that occur in many topics. Another problem is that it does not consider the distance of words in the contexts it handles.

Another method for word sense disambiguation was proposed in [13] by Dagan et al. (1994). The method chooses the most probable sense of a word using frequencies of the related word combinations in a target language corpus. In this 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. Two evaluations were carried out for this method, one using Hebrew sentences and the other using German sentences. The accuracy of the system was 91% and 78% for Hebrew and German sentences, respectively.

The other method of word sense disambiguation proposed in [14] by Justeson et al., 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 and the accuracy of the system reached 97%. However, for adjectives which can be differently accompanied by the same noun, this method cannot be helpful in disambiguation.

The system presented by Ng and Lee (1996) in [15] 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 by all the examples of that word, the given word is disambiguated choosing the closest matching instance. The accuracy of the system on test sets from Brown corpus and WSJ corpus was reported to be 58% and 75.2%, respectively. The results were calculated on a task including 121 nouns and 70 verbs, using fine-grained sense distinctions from WordNet.

The method presented by Brown et al. [11] requires a bilingual word-aligned corpus, which is costly to build. This is one of the challenges of this method, which makes difficult the applicability of the method to other pairs of languages.

The other method proposed by Mosavi et al. in [16] is somewhat the same as the method presented in [13] 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 performed either for the source or the target language corpus from the syntactic viewpoint. The only task of the proposed algorithm, for gaining the required statistical information, is determining the nearest noun, pronoun, adjective, or verb to the ambiguous word, whether it is a noun, a verb, an adjective, or an adverb. When applying this method for the comparison of English and Persian, only a small portion of ambiguous words in English can be correctly translated into Persian.

In addition to supervised approaches, unsupervised approaches and combinations of them have also been proposed for word sense disambiguation. For example, [17] proposed an ambiguity resolution technique which divides the occurrences of a word into a number of classes by determining for any two occurrences whether they belong to the same sense or not, which is then used for the full ambiguity resolution task. The approaches proposed by [18, 19] are other examples of unsupervised learning methods. [20] Had proposed an unsupervised learning method using the Expectation-Maximization (EM) algorithm for text classification problems, which then was improved by [21] in order to apply it to the ambiguity resolution problem. [22]

ISBN: 978-988-19251-5-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Combined both supervised and unsupervised lexical knowledge methods for word sense disambiguation. [23] and [24] used rule-learning and neural networks respectively.

.

III. The Proposed Method for WSD

In this section, we introduce and illustrate a new system for word sense disambiguation. The proposed scheme includes two major parts; the first part performs a feature extraction process and converts each paragraph included in the corpus into a vector of feature values. The main part of the system is a K-NN classifier used for WSD. In order to improve the accuracy of the WSD method, we propose some weighting methods at the end of this section.

A. K-Nearest Neighbor (K-NN)

K-NN is a supervised learning algorithm in which the classification is accomplished based on learning by analogy, that is, by comparing a given test vector with training vectors that are similar to it.

When an unknown vector is introduced, K-NN classifier finds *k* most similar training vectors that are closest to the unknown vector. These *k* training records are the *k* "nearest neighbors" of the unknown vector. K-NN determines the label of the unknown vector by using its k nearest neighbors.

In k-NN, the number *k* is a positive integer number and can be determined experimentally. If *k* = 1, then the unknown vectors is simply assigned to the class of its nearest neighbor, otherwise it is classified by the majority vote of its neighbors.

The distance between a test vector and the training vectors in K-NN classifier is commonly based on the Euclidean distance**.** The Euclidean distance between two typical vectors

X1 = (x_{11}, x_{12}, …, x_{1n}) and ***X2*** = (x_{21}, x_{22}, … , x_{2n}), is defined as follows:

$$dist(x_1, x_2) = \sqrt{\sum_{i=1}^{n}(x_{1i} - x_{2i})^2} \qquad (1)$$

B. Feature extraction

Feature extraction is a very important step in developing WSD system, which will then have a high effect on the system performance. In this problem, features are the set of words which exist in the context of the ambiguous word that is under investigation. We extract two sets of features from the corpus; the set of words that have occurred frequently in the text and the set of words surrounding the ambiguous word. For this purpose, first of all, we omit all stop-words (i.e., the words which are not quite valuable, such as articles, all types of pronouns, etc) as well as the punctuations from the context. The processing we perform over the text is not case sensitive, i.e., does not make difference between upper case and lower case letters.

Set of frequent words

The first subset of the extracted features includes frequent words. Every word that frequently occurs in the training corpus is represented as a feature in this set. The value of a typical feature is dependent on the number of times it has co-occurred with the ambiguous word (in the same paragraph), ignoring the position of its occurrence in the paragraph.

A word can be selected as a feature of this category, if it meets the following pair of conditions:

1. Every word w_i that occurs at least *n* times in the paragraph that the ambiguous word is in its *k*_th sense. Note that the proper value of *n* should be determined experimentally.

2. $prob(k|w_i) \geq p$

Where *p* is a predefined value determined through experiments, and the value of prob(k|w_i) is calculated as follows:

$$prob(k|w_i) = {N(k, w_i)}/{N(w_i)} \qquad (2)$$

Where N(k,w_i) refers to the number of paragraphs in which the word w_i occurs with the ambiguous word that is in its *k*_th sense, and N(w_i) refers to the total number of paragraphs in which the word w_i occures.

The second condition is checked for the words that satisfy the first condition . The first condition tries to prevent selecting the words based on spurious and rare occurrences. The second condition is used to reduce the probability of selecting words that are frequent, but co-occur with all senses of the ambiguous word.

After detecting a number of words that satisfy the above pair of conditions, *m* words that co-occur more frequently with the *k*_th sense of the ambiguous word are selected as features to construct the dataset (If the number of these words for a given sense k exceeds m). Notice that the value of m is determined via experiments.

Set of the surrounding words

Every feature in this set is assigned a weight value according to its positional distance to the ambiguous word. For this purpose, we select *s* words on either sides of the ambiguous word (where s is determined experimentally) and for all of them, check the pair of conditions discussed in the previous section. Among all the words satisfying both conditions, m words that co-occur more frequently with the *k_th* sense of the ambiguous word are selected for the dataset (If the number of these words for a given sense k exceed m).

This set of features do not get binary values (showing whether the word exists or not), instead their positional distance to the ambiguous word will be considered in assignment of their weight.

As a heuristic, we assign the value of $\frac{1}{|i|}$ to a word which has the distance of i to the ambiguous word. Our meaning of distance here is the number of words between two words in a text.

C. Running K-NN

After feature extraction is accomplished, the dataset will be constructed using the extracted features. Hence, the dataset schema consists of two sets of features as described above as well as a class label (i.e., the word sense) of each data instance. Then after selecting a subset of data as test

ISBN: 978-988-19251-5-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

instances, using the K-NN algorithm, every test vector is compared with all training vectors to find the *k* most similar training vectors (i.e., the k nearest neighbors of the test vector). In order to determine the class label of a test pattern using KNN, we utilize follow strategy:

Majority voting

In this approach, a kind of voting is carried out between the k nearest neighbors. Among all possible senses for an ambiguous word, the sense that has been stated by the majority of neighbors will be selected.

It is clear that the value of k shall be selected from odd numbers in experiments in order to avoid tied votes.

D. *Feature Weighting*

As a matter of fact, the features extracted from the corpus do not have the same effect on the final results. Indeed, the importance of each feature has a direct relation with its occurrence frequency. Hence, we propose the following heuristic to weight the extracted features:

$$w_{f_i} = \left(\log_{N(k)} N(k, f_i)\right) * prob(k|f_i) \qquad (4)$$

Where N(k,f_i) refers to the number of paragraphs (for set of the frequent words) or sentences (for set of the surrounding words) in which the feature f_i co-occurs with the *k*_th sense of the ambiguous word, and N(k) refers to the number of paragraphs or sentences in which ambiguous word is in its *k*_th sense, and prob(k|f_i) is computed as follows:

$$prob(k|f_i) = {N(k, f_i)}/{N(f_i)} \qquad (5)$$

Where N(f_i) refers to the number of paragraphs in which f_i occurs.

In order to use the feature weights in computing the distance of a pair of vectors (say x1 and x2), the Euclidean distance is changed as follows:

$$dist(x_1, x_2) = \sqrt{\sum_{i=1}^{n} w_{f_i} (x_{1i} - x_{2i})^2} \qquad (6)$$

Where w_{fi} is the weight assigned to the feature f_i and x_{ji} is the value of the *i*-th feature in the *j*-th vector (*j*=1 or *j*=2).

IV. Experimental Results

In order to evaluate the proposed scheme, we used TWA[25] sense tagged data which is a benchmark corpus developed at University of North Texas by Mihalcea and Yang in 2003. TWA is a Sense tagged data focusing on six words each having two different senses (including " bass", " crane", "motion", " palm", " plant" and " tank").

As a highly used methodology in machine learning and data mining, we used 5 fold cross-validation to estimate the performance of the algorithm. Thus, for each ambiguous word, the set of all related samples were divided into five equal folds. Four folds were used to extract the features and to train k_NN classifier, while the remaining folds were used as test data. In other words, the training and the test data involve 80/20 splitting ratio of the available text.

The above procedure is repeated 5 times so that each fold is used as the test data once. The average accuracy of the proposed method across the 5 fold cross validation is reported in tables 1and 2.

TABLE I
ACCURACY VALUES OF THE PROPOSED SCHEME ON TWA DATA SETS BEFORE APPLYING FEATURE WEIGHTING, IN THREE CASES; USING JUST THE SET OF FREQUENT WORDS (SET1), USING JUST THE SET OF SURROUNDING WORDS (SET2) AND USING BOTH SETS SIMULTANEOUSLY

	Accuracy		
	Set1	Set2	Set1&set2
Ambiguous words	Majority voting	Majority voting	Majority voting
bass	90.7	89.7	90.7
crane	76.8	75.8	76.8
motion	70.1	72.6	71.1
Palm	76.1	81.1	78.1
plant	59	56.9	59.6
tank	69.2	65.7	67.7
Average	73.7	73.6	74

TABLE II
ACCURACY VALUES OF THE PROPOSED SCHEME ON TWA DATA SETS AFTER APPLYING FEATURE WEIGHTING, IN THREE CASES; USING JUST THE SET OF FREQUENT WORDS (SET1), USING JUST THE SET OF SURROUNDING WORDS (SET2) AND USING BOTH SETS SIMULTANEOUSLY

	Accuracy		
	Set1	Set2	Set1&set2
Ambiguous words	Majority voting	Majority voting	Majority voting
bass	90.7	89.7	90.7
crane	76.8	75.8	76.8
motion	75.6	72.6	75.6
Palm	78.6	82.1	78.6
plant	61.2	56.9	63.8
tank	70.1	65.7	71.1
Average	75.5	73.8	76.1

In order to compare the results with other disambiguation methods, we executed some of the existing corpora-based methods (the methods proposed in [11], [12], [13], [15], [16]) over the same data. The results are shown in average in Table3.

TABLE III
THE ACCURACY RESULTS OF DIFFERENT WORD SENSE DISAMBIGUATION METHODS USING TWA DATA SETS, COMPARED TO THE PROPOSED METHOD IN TWO CASES (CASE 1: MAJORITY VOTING, CASE 2: WEIGHTED VOTING)

	Brown method	Yarowsky method	Degan method	Ng method	The proposed method
Accuracy	72.3	71.8	77.4	73.7	76.1

ISBN: 978-988-19251-5-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

V. Conclusion

In this paper, we proposed a supervised learning method for word sense disambiguation based on K-Nearest Neighbor algorithm. Using TWA as a benchmark dataset, we first extracted two sets of features; the set of words that have occurred frequently in the text and the set of words surrounding the ambiguous word. Then, using 5 fold cross validation approach the dataset was divided into training and test parts for a k_NN classifier. In order to improve the classification accuracy of K-NN, we proposed and evaluated a feature weighting strategy. As shown through a set of experiments, the effect of the weighting scheme was encouraging and led to promising improvements in most cases.

References

[1] Statistical Post-Editing of a Rule-Based Machine Translation System, Proceedings of NAACL HLT 2009: Short Papers, pages 217–220, Boulder, Colorado, June 2009. c 2009 Association for Computational Linguistics, pp 217-224

[2] Arabic/English Word Translation Disambiguation using Parallel Corpora and Matching Schemes, 8th EAMT conference, 22-23 September 2008, Hamburg, Germany, pp 6-11.

[3] On the use of Comparable Corpora to improve SMT performance Sadaf Abdul-Rauf and Holger Schwenk, Proceedings of the 12th Conference of the European Chapter of the ACL, pages 16–23, Athens, Greece, 30 March – 3 April 2009. c 2009 Association for Computational Linguistics, pp 16-23

[4] Gale, K. Church, and D. Yarowsky.: A Method for Disambiguating Word Senses in a Large Corpus. Computers and Humanities, vol. 26, pp. 415-439 (1992).

[5] Yarowsky.: Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French. In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, pp. 88-95 (1994).

[6] T. Ng and H. B. Lee.: 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, pp. 40-47 (1996)

[7] Mangu and E. Brill.: Automatic rule acquisition for spelling correction. In Proceedings of the 14th International Conference on Machine Learning pp. 187-194.

[8] R. Golding and D. Roth. A Winnow-Based Approach to Context-Sensitive Spelling Correction. Machine Learning, vol. 34, pp. 107-130 (1999)

[9] Escudero, Gerard, Lluís Màrquez & German Rigau.: Boosting applied to word sense disambiguation. Proceedings of the 12th European Conference on Machine Learning (ECML), Barcelona, Spain, 129-141 (2000)

[10] T. Pedersen.: A simple approach to building ensembles of Naive Bayesian classifiers for word sense disambiguation. In Proceedings of the 1st Annual Meeting of the North American Chapter of the Association for Computation Linguistics, pp. 63–69, Seattle, WA (2000)

[11] Brown, P. F., DellaPietra, S. A., DellaPietra, V. J., and Mercer, R. L. (1991). Word Sense Disambiguation Using Statistical Methods. Proceedings. Annual Meeting of the Association for Computational Linguistics, pp. 264–70.

[12] Yarowsky, D. (1992). Word Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora. Proceedings of 15th International Conference on Computational Linguistics, pp.454–60.

[13] Dagan, I. and Itai, A. (1994). Word sense disambiguation using a second language monolingual corpus. Association for Computational Linguistics, 20(4): 563–96.

[14] Justeson, J. J. and Katz, S. M. (1995). Principled disambiguation: discriminating adjective senses with modified nouns. Computational Linguistics, 21(1): 1–28.

[15] Ng, H. T. and Lee, H. B. (1996). Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Exemplar-Based Approach. Proceeding of the 34th Annual Meeting of the Association for Computational Linguistics (ACL-96), Santa Cruz.

[16] T. M. Miangah and A. D. Khalafi, Word Sense Disambiguation Using Target Language Corpus in a Machine Translation Systems, Literary and Linguistic Computing, Vol. 20, No. 2, 2005

[17] Schütze, H.: Automatic WS discrimination. Computational Linguistics, 24(1):97-124 (1998)

[18] K. C. Litkowski.: Senseval: The cl research experience. In Computers and the Humanities, 34(1-2), pp. 153-158 (2000)

[19] Dekang Lin.: Word sense disambiguation with a similarity based smoothed l brary. In Computers and the Humanities: Special Issue on Senseval, 34:147-152 (2000)

[20] Nigam, McCallum, Thrun, and Tom Mitchell.: Text Classification from Labeled and Unlabeled Documents using EM. Machine Learning, 39(2/3):103–134 (2000)

[21] Shinnou , Sasaki.: Unsupervised learning of word sense disambiguation rules by estimating an optimum iteration number in the EM algorithm, Proceedings of the 7th conference on Natural language learning at HLT-NAACL,p.41-48, Edmonton, Canada (2003)

[22] E. Agirre, J Atserias, L.Padr, and G.Rigau.: Combining supervised and unsupervised lexical knowledge methods for word sense disambiguation. In Computers and the Humanities, Special Issue on SensEval. Eds. Martha Palmer &Adam Kilgarriff. 34:1,2 (2000)

[23] David Yarowsky.: Unsupervised word sense disambiguation rivaling supervised methods. In Meeting of the Association for Computational Linguistics, pages 189.196 (1995)

[24] Towell and E. Voothees.: Disambiguating Highly Ambiguous Words. Computational Linguistics, 24 (1), pp. 125-146 (1998).

[25] WWW.cse.unt.edu/~rada/downloads.html.

ISBN: 978-988-19251-5-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)