Word Sense Disambiguation of Persian Homographs

F. Jani and A.H. Pilevar
Language Engineering Lab., Computer Engineering Dept., Bu Ali Sina University, Hamedan, Iran

Keywords: Thesaurus, Word Sense Disambiguation, Persian Language Corpus.

Abstract: This paper seeks to elaborate on the disambiguation of Persian words with the same written form but different senses using a combination of supervised and unsupervised method which is conducted by means of thesaurus and corpus. The present method is based on a previously proposed one with several differences. These differences include the use of texts which have been collected by supervised or unsupervised method. In addition, the words of the input corpus were stemmed, and in the case of those words whose different senses have different roles in the sentence, the role of the word in the input sentence was considered for disambiguation. Applying this method to the selected ambiguous words from "Hamshahri", which is a standard Persian corpus, we achieved to a satisfactory accuracy of 97 percent in the results, and evaluated the presented method as a better and more efficient in comparison with the similar methods.

1 INTRODUCTION

There are words in some languages which cannot be pronounced unless one knows their meaning in the context. Such words are called homographs. For converting a text to equivalent speech, the system must be able to disambiguate homographs. Therefore, according to the above definition, a word with several senses is called homograph, and disambiguating homographs is a matter of word sense disambiguation (WSD).

WSD is related to attributing a proper sense to a certain word in a text or speech (Gausted, 2004). In contrast to other languages, and English above all, for which many WSD systems have been designed, it should be noted that, due to a lack of tagged corpora, no such system has been so far developed in a large scale and with a high accuracy for the Persian language.

With regard to the strategies and approaches used, there are three methods for solving the problem of assigning the appropriate meaning to an ambiguous word in context:

I. Using an explicit lexicon such as thesaurus, with knowledge based method.
II. Disambiguation of word senses based on, trained sense tagged corpus, for extracting relevant information.
III. Combining the aspects of both I, and II methodologies.

This paper seeks to elaborate on the disambiguation of Persian homographs using a combination of supervised and unsupervised method which is conducted by means of corpus and thesaurus. The method of paper (Ide and Veronis, 1998) is extended from different aspects and presented in this article. These differences include the use of texts which have been collected by supervised or unsupervised method. In addition, the words of the input corpus were stemmed and, in the case of those words whose different senses have different roles in the sentence, the role of the word in the input sentence was considered for disambiguation.

The organization of this paper is as follows: The presented algorithm is discussed in Section 2. In section 3, the corpus and thesaurus used in the experiments are described. Section 4 provides a report on the proposed method and its comparison with the previous method as implemented on a selection of homographs. Finally, section 5 offers an analysis as well, the obtained conclusions.

2 THE PROPOSED METHOD

The method which is described in (Ide and Veronis, 1998) is used in this article and the algorithm is extend for recognizing the ambiguous words. The presented algorithm is based on conceptual

Jani F. and Pilevar A..
Word Sense Disambiguation of Persian Homographs.
DOI: 10.5220/0004031703280331
In *Proceedings of the 7th International Conference on Software Paradigm Trends* (ICSOFT-2012), pages 328-331
ISBN: 978-989-8565-19-8

categorization of context such that to be able to use it for disambiguation of the word senses. Therefore, in the first step we determine, to which category the ambiguous word belongs. Then we construct the related context discriminator and, since each word sense belongs to one category, the correct sense can be predicted (Ide and Veronis, 1998; Makki and Homayounpour, 2008). The thesaurus is used for determining the conceptual categories.

In this step, we extract sentences for each sense which contain the ambiguous word and then begin to disambiguate with regard to words that are collocated with one of the senses of the ambiguous word, but it must be noted that, due to the ambiguity of the homograph, we cannot extract sentences merely through the ambiguous word. Thus, we should use thesaurus to find the synonymous words for each sense of the ambiguous word and extract the sentences containing those words for each conceptual category from the input corpus. However, in reality, some words in thesaurus are not only ambiguous in themselves, but also they cannot often replace the ambiguous word in the sentence; therefore, in this method, a very limited number of words in the thesaurus were considered, and to increase accuracy, some sentences were extracted from the corpus in a supervised manner. More precisely, a sentence in which the ambiguous word or its synonym resides is extracted, because by considering a window around the ambiguous word we may go beyond the scope of one sentence and it should be noted that two contiguous sentences do not necessarily share a conceptual connection. In fact, our aim is to extract the words which are collocated with the ambiguous word in the sentence as well as their probability of occurrence.

Another difference of this method with the previous one is the use of texts whose words have been stemmed. The reason is that in the collected texts, the same word may occur in different morphological forms and in an unstemmed mode a separate probability is calculated for each form. If the input sentence is the same, every time a word is likely to appear in one of its forms, it has been considered as a stem for it. After identifying the categories, similar to (Ide and Veronis, 1998) the discriminators are constructed in the following steps:

1. For the conceptual categories, determine the representatives of the contexts.
2. In contexts, calculate the weights of the words
3. For new contexts, use the calculated weights in step 2, for predicting sense of ambiguous word.

In the proposed algorithm, the above steps are developed as follows:

2.1 Determine the Representatives of the Contexts

In the supervised mode, the sentences containing ambiguous words were extracted from the Hamshahri Corpus and added in separate conceptual categories according to the meaning of the ambiguous word. In the unsupervised mode, the sentences containing a synonym of the ambiguous words in different conceptual categories were extracted. Thus, for each conceptual category several sentences were obtained in a classified form.

2.2 Calculate the Weights of the Words

For every word (w) in the collected sentences the probability Pr(TCat|w) is calculated for each category (TCat) of the ambiguous word by use of the law of conditional probability. Similar method is proposed in (Makki and Homayounpour, 2008; Yarowsky, 1992), the salient words with larger probabilities (Formula 1) are selected, and their logarithms are considered as weights.

$$\Pr(w|TCat)/\Pr(w) \qquad (1)$$

In (Ide and Veronis, 1998) this probability as well as its logarithm is considered for all of the words in the collected texts. The presented article is doing the same, but because Pr(w) is equal for all conceptual categories, only Pr(w|TCat) is considered in the proposed method, which is caused a higher speed in the system.

2.3 Predicting Sense of Ambiguous Words

For any newly entered ambiguous word, the system assigns a score value. This score value calculated based on Formula 2 which adds the words category weights.

$$\begin{aligned}&Score(TCat)=\\&\quad\sum \log(\,(\Pr(w|TCat)*\Pr(TCat))\,/\,\Pr(w))\end{aligned} \qquad (2)$$

For each ambiguous word, the highest assigned category score is selected as its proper sense. Also those words of the test context that have not appeared in any of the collected contexts, for a certain category, scores are calculated by Formula 3.

$$\Pr(w|TCat)=\log(1\,/\,N(TCat)\,) \qquad (3)$$

In formula 3, N is the frequency or number of context, and TCat is the category of the corpus. The relation between probability and frequency is reverse, because for the unknown conceptual categories, there is a small chance for some words to occur in the collected contexts.

3 THE IMPLEMENTED CORPUS AND THESAURUS

In this paper, Hamshahri Corpus was used; the corpus is made up of Hamshahri Newspaper's archive in Iran, and containing 190209 texts with a wide range of topics such as political, cultural, sport, arts, etc. For the supervised mode, due to the lack of a proper sense-tagged corpus in Persian, some parts of the corpus were tagged manually. Furthermore, a part of the intended texts was extracted through an unsupervised method and by means of a thesaurus. A Persian thesaurus which contains words and expressions is used. The utilized Persian thesaurus is very similar to Roget's thesaurus (Fararooy, 1997; Fararooy, 2004), with the same number of heads, sections, classes, and indexes.

4 EXPERIMENTS AND RESULTS

We have applied our system for fifteen different Persian homographs. In presented method, we implement a hybrid technique because, for each homograph a category is extracted from the thesaurus, and for the texts collection corpus is used. The accuracy of the proposed method is shown in Table 1.

The first column shows the homograph, the second column shows the accuracy of the basic method in (Ide and Veronis, 1998), and the two right-hand side columns depict the accuracy of our proposed method.

Having stemmed the words in the collected sentences and calculated the probability of the occurrence of each word in different conceptual categories, a brief inspection discloses that the probabilities have gained more reasonable and appropriate values and the probability of index words has increased due to their different forms being homogenized. Moreover, regarding the score which is finally given to different conceptual categories, the appropriate sense in the proposed method differs large from other sense as compared with the previous method.

In addition, removing the ambiguous words from the thesaurus causes information to be gathered which is conceptually more relevant to the intended concept.

As another factor, considering the role of the word in the sentence in cases where the different senses of the word have different roles has led to an accuracy of 100 percent. This method can also be used to eliminate morphological ambiguities. Every system of natural language processing needs morphological units since it may encounter with words whose morphological structure is impossible to determine without knowing their meaning, for instance: word "سرم (saram)". It has two meanings: the first one is "من + سر" ("my head") which has a morphological structure consisting of /م/ (possessive pronoun) and /سر/ (noun). The second meaning, however, entails another structure, namely a single noun "سرم (serom)" ("a medical drip"). The basis of such methods lies on the highest probability or calculating the similarity according to training data. Therefore, its efficiency is extremely sensitive to the number of training instances.

Table 1: comparison of the basic method with the proposed role based and none role based.

Ambiguous words	Basic %	Role based %	None role based %
سبک (sabk /sabok)	92.33	100	97
نفس (nafs / nafas)	91.37	95	95
خرد (khord / kharad)	89.63	100	95.2
Average	91.11	98.33	95.73

Training instances are the representative of the ambiguous words. Therefore, the training data must contain sufficient number of samples from various texts.

5 CONCLUSIONS

This paper proposed a method for word sense disambiguation of homographs in Persian language, based on the method proposed in (Ide and Veronis, 1998). Basically the method is: sentences containing the ambiguous word are collected from the corpus, the appropriate sentences are gathered for each conceptual category, synonymous words are obtained from the thesaurus, and then the sentences containing these synonyms are extracted. In the basic method there is no need for sense-tagged data, but for increasing accuracy and with the aim of evaluating the efficiency of the basic method as compared with the supervised method a number of the intended texts here were gathered in a supervised manner. In order to increase the accuracy, the words in the obtained corpus were stemmed after extracting the sentences, and it caused a noticeable increase in efficiency and accuracy.

After obtaining the sentences in a classified form for each conceptual category, the probability of occurrence of each word was calculated in that category. These probabilities lead to a score for the input test sentence in each category which is, in fact, indicative of the probability of each sense. Finally, the conceptual category with a higher probability is taken as the output of the system. Moreover, an analysis of the corpus showed that regarding some ambiguous words different senses may appear in different roles in the sentence; therefore, by considering the role of the ambiguous word in the input test sentence we can produce an accuracy of 100 percent in such cases.

Also, the results showed that the modifications applied to the basic algorithm in (Ide and Veronis, 1998) improved the accuracy of word sense disambiguation by 6 percent. The proposed method can also be used for morphological disambiguation.

REFERENCES

Fararooy, J., Thesaurus and Electronic transfer of Persian language content, *In: 2nd workshop on Persian language and computer*, Tehran, Iran, 2004.

Fararooy, J., Thesaurus of Persian Words and Phrases, 1999.

Gausted, T., Linguistic Knowledge and Word Sense Disambiguation, *PhD dissertation, Groningen University*, 2004.

Ide, N., Veronis, J., Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art, *Computational inguistics* 24(1), 1–40, 1998.

Makki, R., Homayounpour, M., Word Sense Disambiguation of Farsi homographs Using Thesaurus and Corpus, *Amirkabir University of Technology*, Tehran, Iran, 2008.