ACSIJ Advances in Computer Science: an International Journal, Vol. 4, Issue 5, No.17 , September 2015
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

Persian Word Sense Disambiguation Corpus Extraction Based on Web Crawler Method

Mohamadreza Mahmoodvand[1], Maryam Hourali[2]

[1] Artificial Intelligence MSc. Student ICT Department, Malek-e-Ashtar University of Technology
Tehran, Iran
Mr.mahmoodvand@yahoo.com

[2] Assistant Professor ICT Department, Malek-e-Ashtar University of Technology
Tehran, Iran
MHourali@mut.ac.ir

Abstract

Finding an appropriate dataset for natural language processing applications is one of the main challenges for researches of this field. This issue is more problematic in Non-Latin languages especially Persian language. Access to an appropriate dataset that can be used in development of practical programs in language processing field, helps us to validate the obtained results and provide the feasibility for comparison and precise analysis of the research studies in this field. This paper presents the procedure for extracting a standard dataset in Persian language. This dataset can only be used for research studies in the field of word-sense disambiguation in Persian language. The required documents that include the ambiguous words of interest are collected by a crawling robot; then these words are processed and registered in Persian dataset for ambiguous words. In this research, three prevalent Persian ambiguous word are used for extracting appropriate phrases that included these words. Finally, a framework for creating the proper configuration for application in word-sense disambiguation problems is presented. By using of this method, we have a solution for absence of suitable word sense disambiguation corpus in Persian language.

Keywords: *Natural language processing, Word sense disambiguation, Information Extraction, Corpus, Machine learning*

1. Introduction

Knowledge of natural language processing has a significant role in development of different sciences especially computer science. By increasing progress of speech and natural language processing methods, now it is feasible to work on edge of science and produce practical programs that it was not possible before; programs that interact with human and exclude the computer inputs from the limited scope of input devices; these kind of programs accept and understand the easiest input command namely the human language, like human who is familiar with this kind of voice command and response for years. With further development of the natural language processing knowledge, it is possible to step in a way that computers would be able to understand, think and act like humans.

Natural language processing is a very interesting approach for human-machine communication and when the practical programs of theses field become completely operational, it can lead to surprising developments. This area of knowledge like other areas has encountered different bottlenecks in its progress path that has slowed down the development trend of this filed. Recognizing the entities, identifying meaning of words, segmentation and text summarization are some of the important tasks in natural language processing. In order to resolve these problems and accelerate the development trend, researchers have concentrated on these issues to provide a comprehensive solution for these problems.

Word-sense disambiguation also is an important subject in natural language processing. The main goal of the language is to present a certain subject to the audience. This subject is extracted from the meaning of words in that language. Without proper identification of words' meaning in a sentence, it is not possible to have a true understanding of writer's intention. For proper identification of concepts in a text, the computer should be able to identify the role and meaning of words. This issue seems to be a more serious problem when words have different meaning regarding the role and meaning of their neighboring words. In different languages, linguists, computer scientist and artificial intelligence experts are gathered together in order to provide a solution for word disambiguation in their own language. Regarding that different practical applications have been developed in Persian language, there is a serious need to provide a solution for word disambiguation in Persian language.

With the advancement of technology, services are becoming more digitalized. Financial transactions and economic activities are developed on computer network

infrastructures. As different activities grow in computer world, an essential need become more evident day by day. Processing speed is increased day to day and with hardware and software enhancement of computer systems, it is possible to take the outputs faster than before. All the computer systems require input information. Users of computers have limited sources of input to enter the information into the computer or deliver their command to the computer. Keyboard, mouse, and touch screens are only part of equipment that users can employ to provide the inputs. All of these equipments require physical activity of the user. It may come to mind that these kinds of physical activities do not cause any problem, but there is an easier solution which is voice command.

By using voice processing knowledge besides natural language processing knowledge, it would be possible to convert user commands to understandable commands for the computer in fractions of a second and observe the results. In this case, there is no need to have basic knowledge of how to work with computers and therefore a larger group of people would be able to work with computers, and people with physical disabilities would be able to interact with computers.

The role of natural language processing is not only limited to recognition of user's voice command, and includes a wide range of practical applications. Data mining and text mining in natural language can have better and more acceptable results by employing natural language processing methods. For instance in semantic web branch, true understanding of user's inquiries can be achieved. In search engines, most of the results are ranked based on the similarity between the input character string by the user and the text in webpage. Often the user's purpose of searching a phrase is to find the webpage in which the same phrase or similar phrase including the same concept exists. This goal can only be achieved by using the recognized methods in natural language processing field. In the following, natural language processing knowledge and specially word sense disambiguation task are investigated.

2. Word Sense Disambiguation

Word-sense disambiguation is an activity during which the proper meaning of a word is selected. Word disambiguation has an important role in development of different parts of natural language processing. Fields like machine translation, answering questions, text classification, and information retrieval are fields that use word-sense disambiguation [1]. The employed methods depend on the expected application. First it is required to do more investigation on word disambiguation and its implementation should be explained without any dependence to its applications.Word-sense disambiguation algorithms receive a certain word with a list of its meanings as the input and give one of the meanings as the right meaning as output.

Nature of the input word and its probable meaning list depends on the application that we expect from the word-sense disambiguation algorithm. For example in machine translation for translating English to Spanish, the list of candidate meanings is different translations of an English word in Spanish language; or if applications like speech combination are considered such as SIRI personal assistant of apple company, the set of meanings can be homographs. If our task is to automatically index medical texts, list of the candidate word meanings would be limited to database of medical subjects.

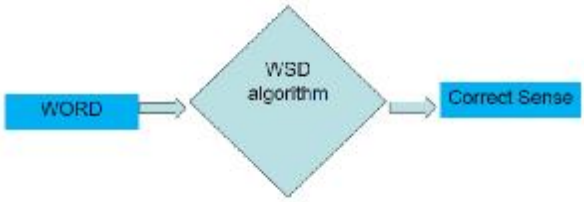

Fig. 1 WSD input and output[12]

When we intend to use word disambiguation in a general mode, we can use thesaurus or word-net for a set of meanings.Word sense disambiguation algorithms are generally categorized in two forms:

1- In simple lexical (related to words) tasks, a set of target words and also the set of meanings of each word are selected. In such cases, the set of words and set of meanings are small.

For simple lexical tasks, supervised machine learning algorithms are used. For every single word, a set of texts are collected and for each of them the label for right meaning of target word is identified manually. Classifiers are trained using these labeled texts. The early work was concentrated on these simple tasks and was intended to disambiguate target words like "BASS" or "BANK".

2- The other kind of task includes all the words. In this case all the text, glossary and meaning list are available for each input and the goal is that every word in the text is disambiguated and its meaning is determined. In this case there is no appropriate set for training. In addition, if the glossary includes words with many meanings, training the classifier with one phrase will have no useful result.

3. Word Sense Disambiguation Methods

Common methods for word-sense disambiguation include:
a)Supervised method: in this method we require a large and comprehensive dataset which include semantic labeling of ambiguous words.

b)Semi-supervised method: in this method we require a small fraction of the semantically labeled dataset.

c)Thesaurus method: in this method, texts thesaurus which is accessible in databases like Word-net or dictionaries are used in order to disambiguate the words.

Generally methods which use a large dataset including semantic labeling of ambiguous words, gives the best results. Therefore, the role of providing suitable dataset for disambiguation problem is very effective in methods that are employed in this field. Like other problems in artificial intelligence research work, preparing a suitable dataset for training and validating the designed system is one of the main challenges.

4. Supervised WSD

If we have a data that manually its meanings are labeled correctly, in this case we can employ supervised machine learning methods for word disambiguation problem. First, we extract the proper features from the text that can help us to have better prediction of the word meaning and then it's time to train the classifier with these features and this procedure is followed to extract the correct meaning of word.
The training output is a classifier that can receive the texts without labeling and return the correct meaning of target word [2]. For simple lexical tasks, there are many labeled texts for different words. For labeling all the words, meaning match is used. In this method, there is a collection of texts and every general word in the sentence is labeled by its corresponding meaning which is obtained from the dictionary.
Supervised algorithms include huge volume of word disambiguation methods. This algorithm can be categorized as artificial intelligence classification methods. This algorithm similar to other supervised methods consists of training and testing data sets in which the classifier should be trained using training dataset. Then the classifier should determine the class of each test data. As in other supervised methods, the training data have a significant influence in final results. The body of semantically labeled texts is used as training data. Therefore as the volume of training data is increased and labeling is performed more precisely and specifically, the final result will be more acceptable. Regarding the limitations of this algorithm, the best results in word-sense disambiguation is obtained by using supervised methods. In order to determine classes of data, different classifiers are used among which simple Bayesian classifier, decision list, etc are well-known.
In [3] supervised methods are used for word-sense disambiguation in Kannada language. For this purpose, data banks of texts for this language are built and the meaning of target word is labeled and simple Bayesian classifier is used for classifying the samples. As it is clear, since supervised methods have demonstrated acceptable performance in word-sense disambiguation problems, most often this method is used for disambiguation in different languages. Another example is [4] that introduces a supervised method for disambiguation in polish language. For obtaining the best result, 6 classifier models are employed. In [5] another sample of supervised method is applied. In this method, neural network and evolutionary algorithms are utilized. In this research, the main focus is on disambiguation of words which have the same spelling but different meaning and while the words are different but are connected. In other words, the meanings are from the same family. As it was mentioned, due to acceptable performance and high validation percentage of supervised methods, these types of methods are often applied in practical problems. These methods are very popular among computational linguistics researchers [6-8].

5. Persian corpus

The most frequently used dataset in Persian language are Dr. Bijan Khan's and Hamshahri's datasets. These datasets are widely used is research field of natural language processing for Persian language. Bijan Khan's corpus is prepared at language laboratory of Tehran University [9]. This corpus has 2.6 million words. These texts are collected from daily news. The words in this corpus include semantic role labeling and the set for role of Persian words has 40 members. Each of documents has conceptual labeling which shows the dependency of that document to political, literal, etc texts. Texts are categorized into 4300 conceptual groups.
Other common corpus in Persian is Hamshahri corpus [10]. This set is created by crawling robot in Hanshahri website and several preprocessing and labeling steps are also carried out. This configuration has two versions. The second version is newer and has more documents of Persian language texts. The number of documents in second version is 318000 and include conceptual classification. The time period for news of Hamshahri configuration is 12 years.

Although these two configurations are widely used in research fields of Persian natural language processing and information retrieval, but are not appropriate for applications like word sense disambiguation. In word sense disambiguation methods, the ambiguous target word should be present in the document. For applying machine learning methods, the number of such documents should be an acceptable quantity. In addition, the lengthiness of these documents makes them unpractical for this special application. Based on the present methods, only the target word and a limited area around the target word are used. This issue causes that a large volume of information remain useless.

Table 1: occurrence of Persian ambiguous words in hamshahri corpus[11]

Word	*Sense candidates*	*occurrence*
"شیر"	Valve – milk - lion	2066
"تار"	Music instrument – cobweb –beat – gloomy - string	1203
"کرم"	Worm – generosity – crème - cream	574
"مهر"	Persian month – seal – marriage - affection	780

Table 1 shows four target words and their meanings. Also number of target words in texts of Hamshahri corpus which is larger and more comprehensive compared to Bijen Khan corpus is indicated [11]. As is it observed, number of target word "شیر" in all of the texts in this configuration is 2066 times. For other target words the status is similar and number of repetition for target words is few. Therefore it is predicted that if machine learning methods are applied to these texts, unacceptable results will be obtained and the results will not be extendable to all texts and target words in Persian language. Therefore there is a serious need for an appropriate corpus that includes sufficient number of target and neighbor words. The proper configuration for word-sense disambiguation applications should have the following two features:

a) Sufficient number of documents to use in machine learning methods.
b) Include the required information around the target word.

For this purpose, this research is concentrated on creating an appropriate corpus for word sense disambiguation applications in Persian language.

6. Applied Approach

In order to create an appropriate corpus, first it is necessary to determine the links of suitable pages that include the target words in their documents. For this purpose, during an initial evaluation the target community for documents is indicated. Proposed method in [12] focus on one Persian ambiguous word and gathered phrases contained that target word. The best sources for getting access to Persian documents are news websites and official news agencies that publish the news and information in the web. Thus, these news bases are a suitable choice for news and provide a basis for the activity of crawling robot. Mehr[1], Isna[2] and Khabaronline[3] are considered as three sources of documents in this research. Web portal of these news agencies have many indexed documents in Google search engine and this factor result in numerous selection options in selecting the documents. Table 2 presents a comparison between number of indexed pages and pages that include three ambiguous target words .

Table 2: indexed pages of news agencies that contain Persian ambiguous words

Word	**total indexed**	**"شیر" included**	**"راست" included**	**"تار" included**
Isna	870000	19800	20500	7600
Mehrnews	790000	25000	10500	4500
Khabaronline	730000	31000	44000	7100

According to table 2, "shir" is most popular ambiguous word in Persian language. It has three meaning classes in Persian."milk","valve" and "lion" are these classes. Using the Google search engine, it is possible to search and find links of web pages that include the target word. After determining addresses of pages, they will be referred to the crawling robot and it will search for documents that include the target word among the linked web pages. In word-sense disambiguation methods, only the target word is not enough to use in machine learning methods but neighboring words of target word should be extracted using a window with sufficient size. For this purpose, the crawling robot is programmed in a way to extract the sentences that include the target word. Thus, each of final documents has sentences that include the target word.

In order to implement the crawling robot and receive the links of news agency web pages, a web-based user interface is designed, by which the user selects the ambiguous target word of interest. This target word can be any word in persian. Then web pages addresses which have been obtained by Google search engine are entered into the indicated sections. Subsequently addresses are sent to the robot and by crawling into the destination addresses,

[1] Mehrnews.com
[2] Isna.ir
[3] Khabaronline.ir

finally sentences that include the target word are added to database. At last results and number of added documents are shown to the user. Figure 2 demonstrates a view of the designed user interface for crawling robot.[12]

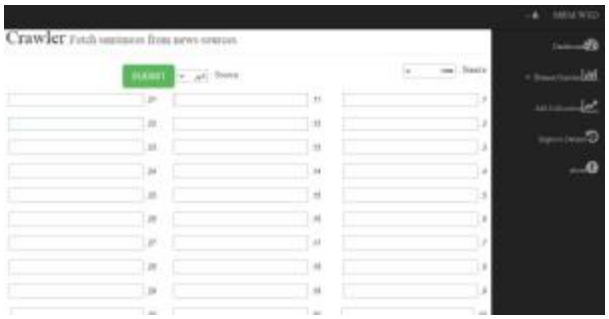

Fig. 2 crawler UI for fetching links and select target word

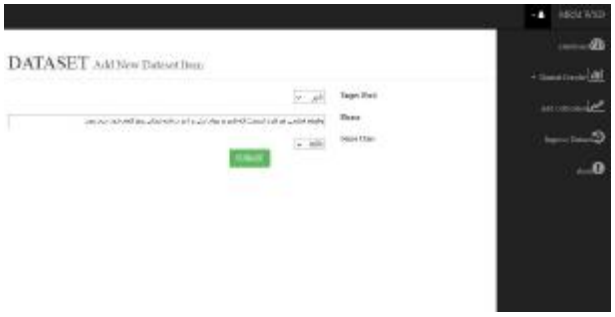

Fig. 3 UI for adding new documents to dataset

There is a other way to add target word included documents into our dataset. Other UI is designed in responsible to easily add new documents that contain target word into the database directly. In this approach we can store examples that language experts mentioned. Fig 3 represents related user interface.

7. Conclusion

One of the main challenges that researchers encounter in natural language processing field is the absence of appropriate texts resources and configurations. This shortage is more evident in rich languages like Persian. In order to resolve this problem different solution are presented, for example the well-known Hamshahri and Bijan Khan text configurations that are applied in many Persian natural language processing research works. However, different fields of natural language processing exist that face shortage of resources and configurations. For instance in Persian word-sense disambiguation field there are no resources to utilize as a configuration in supervised methods [13]. In this research, a framework for building a proper configuration for word-sense disambiguation problem is presented. Based on this framework first the target words are determined. Then Google search engine is employed to find web pages that include the target words and list of web pages links are fed to a crawling robot. The robot will search through these documents and extracts the sentences that include the target word. Finally, the obtained sentences by crawling are added to the data base.

Table 3: sentences and phrases distribution based on target word and sources

agencies	"شیر"	"راست"	"تار"	Total
Isna	492	508	802	1802
Khabaronline	1562	545	135	2242
Mehr	55	762	507	1324
Total	2109	1815	1444	5368

Table 4: independent links used to make corpus

agencies	"شیر"	"راست"	"تار"	Total
Isna	68	128	185	381
Khabaronline	300	132	61	493
Mehr	10	296	114	420
Total	378	556	360	1294

Finally 5368 sentence or phrases gathered from 1294 independent links and from three news agencies as sources and also three ambiguous target word in Persian. Table 4 and 5 represents results of corpus making process. Mehr, Isna and Khabaronline news agencies are considered as the main sources for searching the target word.

Table 5: comparison between proposed method and hamshahri corpus in occurance of two ambiguous words

Corpus	"شیر"	"تار"
Hamshahri	2066	1203
Proposed method	2109	1444

According to table 5 result it is obvious that proposed method can gathered more phrases included target words then hamshahri corpus. This result for proposed method achieved only from 738 independent links and with providing other links this result will be improve. Number of indexed web pages from these news agencies obtained from Google search engine that include target word “Shir” is about 100000. In average, about 6 sentences or phrases that include the target word are extracted from every document. This is while in Hashahri corpus which is the greatest standard corpus in Persian language, the quantity

of target word “Shir” in the entire hamshahri corpus is about 2000 sentences. Regarding the current situation, the presented method can be employed to extract the proper documents to solve word-sense disambiguation problem. By applying this method it would be possible to collect sufficient words for employing supervised and semi-supervised machine learning methods.

References

[1] T. M. Miangah. “Word Sense Disambiguation Using Target Language Corpus in a Machine Translation System”, Literary and Linguistic Computing, 2005

[2] A. Rasekh, M. Sadreddini, S.M. Fakhrahmad, "Word Sense Disambiguation Based on Lexical and Semantic Features Using Naive Bayes Classifier," Journal of Computing and Security,vol., no., pp.1,2, 123-132 April. 2014

[3] S. Parameswarappa, V.N Narayana,"Target Word Sense Disambiguation system for Kannada language," Advances in Recent Technologies in Communication and Computing (ARTCom 2011), 3rd International Conference on , vol., no., pp.269,273, 14-15 Nov. 2011

[4] B. Broda, W. Mazur, "Evaluation of clustering algorithms for Polish Word Sense Disambiguation," Computer Science and Information Technology (IMCSIT), Proceedings of the 2010 International Multiconference on , vol., no., pp.25,32, 18-20 Oct. 2010

[5] A. Azzini, C. Pereira, M. Dragoni, A. Tettamanzi, "Evolving Neural Networks for Word Sense Disambiguation," Hybrid Intelligent Systems, 2008. HIS '08. Eighth International Conference on , vol., no., pp.332,337, 10-12 Sept. 2008

[6] N. Riahi, F. Sedghi, "A Semi-Supervised method for Persian homograph Disambiguation," Electrical Engineering (ICEE), 2012 20th Iranian Conference on , vol., no., pp.748,751, 15-17 May 2012

[7] L. Pengyuan, "Another View of the Features in Supervised Chinese Word Sense Disambiguation," Computational Intelligence and Security (CIS), 2011 Seventh International Conference on , vol., no., pp.1290,1293, 3-4 Dec. 2011

[8] B. Ilgen, E. Adali, A.C. Tantug, "The impact of collocational features in Turkish Word Sense Disambiguation," Intelligent Engineering Systems (INES), 2012 IEEE 16th International Conference on , vol., no., pp.527,530, 13-15 June 2012

[9] M. BijanKhan, ‘The role of the corpus in writing a grammar: an introduction to a soft-ware’, Iranian Journal of Linguistics,Vol. 19, No. 2. 2004

[10]A. AleAhmad , H. Amiri, E. Darrudi , M. Rahgozar , F. Oroumchian, Hamshahri: A standard Persian text collection, Journal of Knowledge-Based Systems, Vol. 22 No.5, p.382-387, Elsevier, July 2009

[11] M Hamidi, A Borjiz, SS Ghidary , Persian word sense disambiguation, 15th icee proceeding, 2007

[12]M.R. Mahmoodvand, M. Hourali , Development of Persian corups sufficient for WSD purpose, in 2nd national conference on computer engineering and IT management, 2015

[13]A. R. Rezapour, S. M. Fakhrahmad and M. H. Sadreddini , Applying Weighted KNN to Word Sense Disambiguation, World Congress on Engineering, 2011

First Author Master of science in artificial intelligence at Malek-e-Ashtar University of Technology ;He interested in natural language processing ,Web mining, web based applications and machine learning.

Second Author Assistant professor in ICT Department of Malek-e-Ashtar University of Technology, Interested in natural language processing, ontology ,data mining and fuzzy-logic .