7th International Conference on Computer and Knowledge Engineering (ICCKE 2017), October 26-27 2017, Ferdowsi University of Mashhad

Semi-supervised approach for Persian word sense disambiguation

Mohamadreza Mahmoodvand
ICT Department
Malek Ashtar University of Technology
Tehran, Iran
Mr.mahmoodvand@yahoo.com

Maryam Hourali
ICT Department
Malek Ashtar University of Technology
Tehran, Iran
mhourali@mut.ac.ir

Abstract— **Word-sense disambiguation is one of the key concepts in natural language processing. The main goal of a language is to present a specific concept to the audience. This concept is extracted from the meaning of words in that language. System should be able to identify role and meaning of words in order to identify the concepts in texts properly. This issue becomes more problematic if there are words that take different meanings because of their surrounding words. Regarding that different practical programs have been developed in Persian language, it is vital now to find a solution for word-sense disambiguation in Persian language. Lack of training data is the biggest challenge in the course of word-sense disambiguation in Persian language. In order to face this problem, machine learning approach with minimal supervision is employed in this research. The applied method tries to disambiguate word senses by considering defined features of target words and applying collaborative learning method. Extracted corpus from published news by news agencies is used as the reference corpus. Evaluating the program by the available corpus on three considered ambiguous words, the implemented method has been able to properly identify the meaning of 5368 documents with 88% recall, 95% precision and 93% accuracy rate.**

Index Terms— **Natural Language Processing, Word Sense Disambiguation, Machine Learning, Semi-Supervised Learning, Text Mining**

I. INTRODUCTION

Word-sense disambiguation means to determine the right meaning of a word which is ambiguous at the first sight and it seems to have several semantic classes. Generally, words with more than a single meaning can be set as target words and our aim is to disambiguate their sense. Humans can easily identify the right meaning of a word based on deduced concept from the sentence, but computer needs more information than only the input sentence in order to identify semantic class of an ambiguous word. By applying disambiguation methods it is possible to extract concepts of a sentence and give it to computer to decide about semantic class of ambiguous word. Word-sense disambiguation is highly taken into consideration in prevalent languages like English and lots of research has been carried out in order to disambiguate words in English language.

If there are lots of instances of metaphor and simile in a text, in would be more difficult to identify right meaning of the target word. Persian language is one of those languages that are rich in metaphor and simile usage and this issue results in more difficult identification of word meaning. Also writhing style of Persian language makes it more difficult to identify the target words while other languages do not have this obstacle. Therefore in order to implement a comprehensive word-sense disambiguation system in Persian language, a preprocessing structure is required to screen out suitable documents at the beginning step.

Presented framework is an indigenous model for solving word-sense disambiguation problem that tries to provide a solution for semantic disambiguation and to present a method for confronting current challenges in this field. The presented framework is prepared to perform semantic disambiguation of all ambiguous words in Persian language and it is only required to give the system the initial and suitable inputs and subsequently it can perform semantic disambiguation of target words in Persian texts and sentences. One of the important features of this framework is that it is expandable; it means that it is possible to add new algorithms to its different components as novel research findings are explored in this field in order to achieve improved results. Thus, it would be feasible to evaluate different word-sense disambiguation methods in Persian language and to select the most effective one.

II. RELATED WORKS

Most of conducted researches in word-sense disambiguation area are implemented by applying supervised methods. The most important issue in implementing this method is providing a suitable dataset which is labeled based on the right meaning of ambiguous words and also contains big volume of information [1]. Features which are based on the neighboring words of target word are also very important. Based on these neighboring features it would be possible to guess the correct meaning of target word more accurately. There is [1] a comparison of available supervised disambiguation methods for English language in medical information subject is provided. In medical information, the target window includes all the paragraph that contains the target word but for simple texts in English language, the window only includes 4 to 10 neighboring words of target word. According to obtained results, simple Bayesian method

978-1-5386-0804-3/17/$31.00 ©2017 IEEE

has performed successfully with 98% accuracy in identifying ambiguous words. Because of deficit in suitable dataset for word-sense disambiguation purposes, most of researches try to make an algorithm to enable them select this suitable dataset.

In [2] an additional software is designed in order to use supervised learning methods that selects suitable documents from semantic network of words, dictionaries and web. Simple Bayesian method is considered as a prevalent supervised disambiguation method [3]. This method is applied to words line and interest which are two common ambiguous words in English language. In simple Bayesian method it is supposed that each of feature vectors points to one of the semantic classes independently and it is unique. In the mentioned research work, binary feature vectors are employed that only determine whether each set of feature words exist in the body of document or not. In fact, in simple Bayesian method we intend to point to a semantic class that can achieve maximum probability rate among other features for target word. Similar to decision list method, in simple Bayesian method it is also probable to obtain zero frequency, and therefore to eliminate this obstacle smoothing ratio should be used. Finally, simple Bayesian method could disambiguate the meaning of these two words respectively with 89% and 88% accuracy. Results of applying simple Bayesian method in different languages are acceptable [4-6].

Boundary between supervised and semi-supervised disambiguation methods is vague because in semi-supervised methods like supervised methods a collection of labeled data and machine learning are employed. In fact part of the process in these methods is the same. But in semi-supervised methods due to lack of suitable resources for training, only a small part of training data is accessible and a big part of available data are given as input to the semi-supervised methods without any labeling or additional information[7]. However the main goal of a semi-supervised method is to create a semantic classifier with minimum amount of available data. For this purpose, algorithm starts the training process using little amount of labeled samples and then evaluates part of data which do not have labeling. After determination of semantic classes, newly produced data are added to the primary dataset and training process is repeated. For creating the primary dataset, labeling can be done by human [8] or automatic selection based on heuristic criteria can be employed [9].

Lack of labeled dataset for word-sense disambiguation is addressed in [10]. A semi-supervised method for correct classification of words with the same spelling is applied in this research. The presented method is a course-grain method and is only concentrated on close meaning of words with the same spelling. Decision list method is used for supervised learning and triple learning method is applied for semi-supervised learning. For identification of 1500 target words in the employed corpus, the applied method had 93% success rate for disambiguation of two Persian homograph words.

Sometimes even a small dataset of labeled data suitable for disambiguation purposes is not available. In such cases, the only solution is to apply corpus based and unsupervised methods. An Example of these methods is used in [11] and the final goal is to use this method in a machine translation system. Even extracting the meaning of target word is done automatically. For determining words semantic classes a bilingual dictionary is used and semantic dependence graph of words is formed based on the employed Persian corpus. In this research second version of Hamshahri corpus is used as reference corpus. This graph includes different meanings of ambiguous words and semantic dependency between them. Also a new method for reinforcement of semantic dependency criteria of words is presented. One of the features of this method is its independency to source and destination languages and it can be used for every language pair in machine translation.

III. Approach Configuration

The presented framework has independent parts that cooperate with each other purposefully. The framework is configured in a way to confront challenges in word-sense disambiguation research field in Persian. At the beginning, suitable corpus for use in machine learning methods is selected and then indigenous preprocessing is performed to screen out suitable algorithms and enhance purity of training data. Using knowledge engineering, semantic labels are ascribed to each of the training documents and features of each target word are determined. In the following, a supervised machine learning method is selected and it is trained using initial labeled seeds (initial data set) and then it is placed inside a semi-supervised learning structure. At last, phrases including ambiguous words that are not used in each of the framework stages and are not labeled will be given to the system in order to be examined and final semantic label for each of these sentences is determined. General structure for different components of the framework is demonstrated in figure 1.

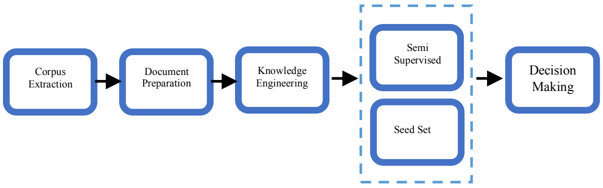


Fig. 1. Proposed method structure

A. Corpus Extraction

Although two frequently used corpora in Persian language Dr. Bijan Khan corpus[12] and Hamshahri corpus [13] have wide applications in Natural language processing and information extraction research fields, but these corpora are not appropriate for word-sense disambiguation applications. In word-sense disambiguation applications there is a need to documents in which the ambiguous target word is present. There should be an acceptable number of such documents so that machine learning applications become feasible. In addition lengthy documents are not appropriate for such applications. Based on available methods, only target word and a limited number of surrounding words are used. Therefore vast amount of information remain useless.

In order to resolve the problem of training data deficit for word-sense disambiguation purposes in Persian, crawling method is used to build the required corpus. Using Google search engine it is possible to search and find webpage links in which target word exists. After finding webpage addresses, they are referred to crawler robot and crawler robot will search for documents in which the target word is present. In word-sense disambiguation methods, the target word is not enough for machine learning applications but a window of surrounding words with suitable size should also be extracted. Hence the crawler robot is programmed in a way to completely extract sentences that include the target word. Thus, each of the finally obtained sentences involves the target word.

In order to implement the crawling robot and receive addresses of news agencies web pages, a web based user interface is designed that user can select the ambiguous target word using this interface [14].

We use three target words in evaluation process:

1. Target Word 1: TW1, "شیر", “Shir”
2. Target Word 2: TW2, "راست",”Rast”
3. Target Word 3: TW3, "تار", “Tar”

This target word can be one of the candidate target words in this research “Shir ”, “Rast ”, “Tar ” or any other word. It is necessary to mention that target words can be defined based on user opinion and by providing features of desired target word, its related documents can be extracted and subsequently semantic disambiguation of those documents can be carried out. Afterwards, web pages addresses obtained from Google search engine can be entered into specified sections of user interface. Then addresses are sent to crawler robot and after crawling in destination addresses, sentences that involve the target word are added to data base and added documents are presented to the user. Thus, by spending little time an appropriate corpus for training a word-sense disambiguation system in Persian is produced and documents of this corpus include phrases and sentences the involve any desired target word in Persian language.

B. *Document Preparation*

Regarding this fact that corpus is obtained by crawler robot and without human interference, it is clear that this corpus should be examined to make sure that final corpus does not include any excess letters and phrases, since presence of excess information in the documents will cause problems for the training step and will burden the training process with high amount of unnecessary computational load. Data preprocessing stages in natural language processing research field include removing conjunctions, stop punctuations, removing excessive writing signs such as Arabic signs and etc. usual preprocessing methods are built based on this foundation. Since there was no corpus available for evaluation of word-sense disambiguation methods, texts extracted from web were used in this research. This issue will add an extra stage to preprocessing phase. In web world, texts are presented in Hyper Text Markup Language (HTML); also Cascade Style Sheet (CSS) language is used to give the texts the preferred design and style. This issue will cause that in addition to extracted texts from web that include the target word, contents of tags in these languages also to exist in the extracted samples. In order to create a suitable database, each of the extracted texts should be refined and unnecessary tags should be removed. This issue is considered as a preprocessing stage in this research.

In this research Persian language is considered as the reference language for word-sense disambiguation. This choice has resulted in more challenges compared to other languages such as Latin languages. One of these challenges is Persian writing style. In Persian language, some words and letters stick together which cause problems in identification of target word in Persian texts. For example if the target word comes in plural form, some alphabets sticks to target word and followed the stem. This issue causes problems for determining correct position of target word and therefore sequential features cannot be applied. To resolve this problem, segmentation of target words is applied in this research. For this purpose, during extraction of texts including the target word, each of the target words is refined in order to remove all the excess words, letters and signs sticking to target word. Thus all the target words in the text are presented without any prefix or suffix. This procedure will improve identification accuracy of target words in text and will enhance validity of results.

C. *Knowledge Engineering*

Knowledge engineering part of this framework is the only part that needs input information from user. In fact this part is dedicated to labeling of training documents and defining database for features of target word. After preprocessing phase is completed, data are ready to participate in training process of word disambiguation system but before that it is required to label some of these data. Feature definition is carried out in two general methods. In the first method we look for features that exist in a certain distance from the target word and depend on their location in the sentence. In second method, vector features are defined that indicates the presence or absence of related words with a special sense of target word in the sentence. To achieve acceptable results, we have tried to take the most advantage from both described features. For this purpose a structure has been created to define required features. Another advantage of this structure for identification of word features is that it would possible to use N-grams as the sign of disordered features. Since Persian language has different characteristics compared to other languages, it is difficult to extract and select suitable features for word-sense disambiguation applications. In Persian language, a set of phrases as N-grams exist that their presence in a sentence indicates a specific meaning of target word. It is not possible to use N-grams in disordered features, because distance of words in these features is the same as writing space and if the space between words is not determined clearly, features identification process in text will encounter problems. Therefore for defining a gram-3 feature, three words should be considered cascading and successively that will impose a heavy computational load and is very time consuming. For instance there is huge set of idioms in Persian. Defining these features has less computational load compared to simple ordered features and final answer will be achieved faster.

D. *Semi Supervised Learning*

After preparing a set of initial labeled documents and a database for target words features, now all the required material for employing machine learning methods to train disambiguation system is provided. In this framework, in order to confront training data deficit in Persian language, a semi-supervised learning method is employed. In semi-supervised method there is no need to have a vast amount of data for training. Training process starts with a small initial dataset. After training process is terminated, classifier will receive raw input data and make decisions about their semantic labels. After input document is labeled, this document is added to the initial dataset and training process will continue again with new and improved dataset. This procedure is continued until all the raw unlabeled data are finished. To classify the data a supervised learning method like simple Bayesian, decision list, nearest neighborhood, etc. is used. In this research, decision list method [15] is employed in supervised learning phase.

The goal of the decision list is finally to determine the correct meaning of input data or input word. To achieve this goal, decision list will use patterns in the text containing the target word. To identify the patterns, extracted features in the data bank are used. Finally, the last decision is the one with the maximum points in the decision list and also features of that decision should exist in the text including the target word. To identify the correct meaning of words, distribution of each feature should be evaluated in the training data. For this purpose, a search algorithm is run over all the training data and presence of words features is evaluated. During this process, number of occurrences for each feature along with true meaning of target word is obtained. In fact the main goal of decision list method is to find a big and most repeated set of features that results in the nearest estimate to correct meaning of word. In this stage of learning process, data are screened out and most valuable features with maximum occurrence rate are introduced. After providing the population, data frequency and available features, it is time to calculate probability rate. Calculating the probability rate helps to have a deep understanding of one feature among other features and to rank the features. The formula for calculating the probability rate is defined as below:

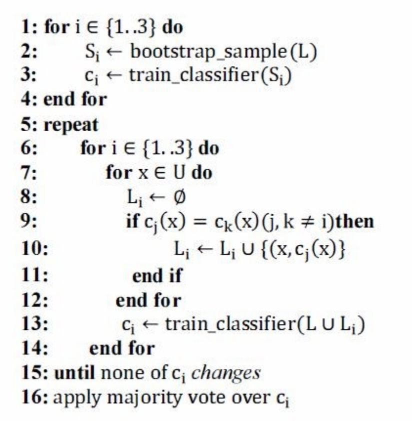
$$\text{Abs (Log (}\frac{P(\,Sense_1 \mid Collocation_{\,i}\,)}{P(\,Sense_2 \mid Collocation_{\,i}\,)}\text{))} \qquad (1)$$

The presented equation calculates ratio of occurrence of a particular feature provided that target sentence include the correct meaning number 1 to the occurrence of the same feature provided that correct meaning is a any number other than 1. Also logarithm of this equation is calculated to have a better understanding of obtained values and it should be noted that outputs of this equation are only positive values. Based on this equation a correct estimation of features strength can be achieved and on this basis more probable features obtain more points and will have higher probability rate and finally these values will be used for proper decision making. In order to prevent occurrence of zero probability, smoothing ration method is used in this research. This ratio is considered as the default occurrence probability for each of the features. If there is no match between a feature and a particular semantic class, its probability will be considered according to smoothing ratio. Therefore in all the calculations for matching between features and semantic classes, smoothing ration is added to the equations to facilitate calculation of probability rates.

In the last stage a suitable decision list is built and is ready to identify correct meaning of target words. Based on fundamental concept of decision list, features that are on top of this list are more reliable features for semantic disambiguation of target words. A character string like a sentence or a phrase that include the target word is the input to decision list. After receiving the input, identification process for feature matching is started on the input from the beginning of decision list where highest features are present. If the considered feature is found in the input content, meaning of that feature class is ascribed to the input sentence, otherwise the next feature in the decision list with less probability rate will be examined. This procedure is continued until one of features match the input content. If all the features in the decision list are examined and none of them match the input phrase, the input sample will considered as unidentifiable. This procedure helps to reduce the computational load and also prevalent features in the language are evaluated before other features which results in less computational load and faster response.

Figure 1 illustrates a general overview of a semi-supervised learning system. After that supervision unit performed its duty and identified the correct meaning of the target word, a decision should be made about the problem output. By using only a single decision system, it is not possible to set up justice between all the features. Also in order to create semi-supervision structure, triple training method (Li & Zhou 2005) is used in this research. In this method in order to reduce singularity of the corpus, first the main corpus is divided into three equal sections. These three sections are created randomly. The reason for making these three sections is to create corpora that have natural dispersion with respect to all the available texts in real world. Sometimes one corpus may have documents more than the other two corpora which is inevitable to create balanced distribution of samples. Each of the created corpora is named with S_1, S_2 and S_3 symbols which indicate produced samples from labeled corpus. The labeled corpus is shown with symbol L.

In triple learning method, in addition to three available corpora, three classifiers are also required that use a supervised learning method on the available corpora. As it was mentioned before, in this research a decision list supervised learning method will be used for each of the classifiers. At the end, three samples of classifiers which have been trained with decision list training method will be produced. These samples will be shown by C_1, C_2, C_3 symbols. In order to train theses classifiers, number one, two and three corpora will be used respectively. In other words, training corpus for each classifier is unique and will not include duplicate documents. Training process for decision list will be performed for each classifier completely and each classifier will produce it own output.



```
for $i \in \{1..3\}$ do
    $S_i \leftarrow$ bootstrap_sample(L)
    $c_i \leftarrow$ train_classifier($S_i$)
end for
repeat
    for $i \in \{1..3\}$ do
        for $x \in U$ do
            $L_i \leftarrow \emptyset$
            if $c_j(x) = c_k(x)(j, k \neq i)$then
                $L_i \leftarrow L_i \cup \{(x, c_j(x))\}$
            end if
        end for
        $c_i \leftarrow$ train_classifier($L \cup L_i$)
    end for
until none of $c_i$ changes
apply majority vote over $c_i$
```

Fig. 2. Tri-Training algorithm

After triple training operation is finished and output is obtained, a new data that has correct label for word meaning is added to initial dataset. This new data is added to initial corpus and all the triple training steps are repeated. In other words the corpus is again divided into three equal and random sections and three decision lists are trained using these three corpora and finally by applying an input phrase including the unlabeled target word, final decision about correct meaning of target word is announced. Thus a new semantic labeled data will be added to dataset. This procedure of training and producing new data is continued until no other unlabeled raw data remains and all the data are classified. Triple learning algorithm is presented in figure 2.

IV. EVALUATION

The presented framework is evaluated by three target words. Ambiguous words play the basic role in word-sense disambiguation problems. These ambiguous words exist in all the live natural languages in world. There may be something shared between these ambiguous words in some languages. Considering that target language of this research is Persian language, three target words are used to train and evaluate the presented methods. These words have dispersed meanings and are significant and widely used words in Persian texts.

TABLE I. DOCUMENT STATS

Total	TW3	TW2	TW1	News Agencies
1802	802	508	492	Isna
2242	135	545	1562	Khabaronline
1324	507	762	55	Mehrnews
5368	1444	1815	2109	Total

Features of each target word are presented in the form of a predefined structure that facilitates identification of target words in the documents. Two types of features are defined in this structure: ordered and disordered. If the feature is of ordered type, search is performed only in the specified boundary for the feature. If the feature is found in the specific distance from the target word, then the sentence indicates the specific meaning of that feature. Each feature can only indicate a single meaning. Table 4 presents defined features for each of the specified target words.

TABLE II. TARGET WORDS AND DEFINED SENSE CLASSES

Sense Classes	Target Words
MILK – VALVE – LION	TW1
MUSIC INSTRUMENT –AMBIGUOUS – STRING	TW2
RIGHTSIDE – POLITICAL – TRUTH – FIRM	TW3

Evaluation of algorithms which are implemented by semi-supervised learning methods is crucial and it is different than other learning methods. One way for evaluating these algorithms is to use expert people. This means that labeled data by semi-supervised classifier are examined carefully by an expert and finally validity of results is judged by the expert. Another common method for evaluation of semi-supervised classifiers is to use Held-out Data.

TABLE III. ORDERED AND DISORDERED FEATURES

Total	Disordered Feature	Ordered Feature	Target Word
137	103	34	**TW1**
94	62	32	**TW2**
62	55	7	**TW3**
293	220	73	**Total**

As it was mentioned above, in this research held-out data are employed in order to evaluate the implemented word-sense disambiguation method. Documents used as held-out data are chosen randomly among labeled documents. Twenty percent of data with semantic labeling are considered as held-out data. Using this evaluation method, it would be possible to obtain a proper understanding of real performance level of designed and implemented word-sense disambiguation system. In this research, validation, accuracy and retrieval criteria are used for evaluation purpose. Calculating each of these criteria is done based on a predefined equation. It should be noticed that these criteria are applied for evaluation of binary classifiers by default, thus for datasets that finally will have more than two classes, these equations have to be changed accordingly.

TABLE IV. DOCUMENTS STATS

Features	Unlabeled	Labeled	Documents	Target Word
137	1419	691	2109	**TW1**
94	1095	720	1815	**TW2**
62	722	722	1444	**TW3**
293	3236	2133	5368	**Total**

An improved method for calculating assessment indexes is to use weighted average. Thus a better understanding of word-sense disambiguation system is obtained. A weight is defined for each class based on its frequency in held-out dataset. In final averaging of indexes for each class, this weight is multiplied at the indexes and then averaging is carried out.

Finally the class with the most frequency in held-out dataset will have the most weight and most effect in final index. This method for calculating assessment indexes is appropriate for this research because in many cases a semantic class of a target word is more prevalent than other semantic classes; therefore this fact justifies choice of weighted average for this research. Assessment indexes which have been calculated by the weighted average method are presented in tables 6, 7 and 8 based on target word and held-out data values.

Among 5368 extracted documents from Mehr, Isna and Khabar-Online news agencies, 1444 documents include "TW3", 1815 documents include "TW2" and 2109 documents include "TW1". Available data are labeled in order to be used as training and evaluation data. Finally 2133 documents were labeled successfully that 691 documents were related to target word "TW3", 720 documents were related to target word "TW2" and 722 labeled documents were related to semantic classes of target word "TW1". 1294 obtained documents from news agencies pages were extracted from web context. On average, from each piece of news 4 documents containing the target word were obtained.

Share of labeled data with respect to all raw data for target words "TW3", "TW2" and "TW1" is 47%, 40% and 34% respectively. In order to identify target words properly between the documents, 293 features are defined totally where 220 features are of disordered type and 73 features are of ordered type. 62 features are defined for target word "TW3", 94 features for target word "TW2" and 137 features are defined for target word "TW1". Among labeled data for target word "TW3", 436 samples are labeled "Music", 240 samples are labeled "String" and 46 samples are labeled "Hide". Labeled data for target word "TW2" have 4 semantic classes from which 354 samples are labeled "Political", 244 samples are labeled "Right side", 101 samples are labeled "Truth" and 21 samples are labeled "Truth". Three semantic classes are also defined for target word "TW1" in which 612 samples are labeled "Milk", 74 samples are labeled "lion" and 5 samples are labeled "Valve". The most frequent semantic class for target word "TW3" is "Music" with 60% share among all the available documents. Most frequent semantic class for target word "TW2" is "Political" with 49% share among all documents and most frequent semantic class for target word "TW1" is "Milk" with 88% share among all the available samples for this target word.

V. Conclusion

In the evaluation procedure, share of the held-out data is set to be variable in order to obtain a proper understanding of system performance. By increasing share of the held-out data, a slight decrease in assessment indexes is observed, but this decrease is negligible which indicates stable performance of designed system for high volume of evaluation data. In this research, share of held-out data and training data is 20% and 80% respectively and obtained values for accuracy, retrieval and validity indexes for held-out and training data are respectively 95.3, 88.3 and 93.73. In other words 95% of labeled samples point to correct semantic class of target word in the sentence. Among selected target words, the best result is obtained for documents including target word "TW1". These documents had the best performance with 96.56% validation rate

TABLE V. Compare results of disambiguation

Accuracy	Recall	Precision	Target Word
96.56	95.6	98.7	**TW1**
90.938	78.4	92.84	**TW2**
93.7	91.03	94.03	**TW3**

The most important reason for this acceptable performance is that target word "TW1" was limited to only three semantic classes and other notable reason is that 88% of documents had a certain semantic class and this issue had a significant effect in identification of correct semantic class.

TABLE VI. Mean results of disambiguation based on held out dataset

Accuracy	Recall	Precision	Training Data	Held-out Data
95.55	92.11	98.76	90%	10%
93.73	88.39	95.31	80%	20%
93.84	86.96	96.26	70%	30%

Among selected target words, the best result is obtained for documents including target word "TW1". These documents had the best performance with 96.56% validation rate. The most important reason for this acceptable performance is that target word "TW1" was limited to only three semantic classes and other notable reason is that 88% of documents had a certain semantic class and this issue had a significant effect in identification of correct semantic class. The next rank was for target word "TW3". Semantic classes of this word had a more balanced dispersion compared to target words "TW1". 93.73% of semantic classes in documents including target word "TW3" were correctly identified. Target word "TW2" obtained the least performance in this research. The most important reason is variety of semantic classes for word "TW2". This word has 4 semantic classes. In addition to variety of semantic classes for this target word, the dispersion of samples in these classes is more balanced that results in harder identification for word-sense disambiguation system. At last, 90.93% of documents including target word "TW2" were correctly disambiguated.

References

[1] Liu H., V. Teller And C. Friedman (2004) A multi-aspect comparison study of supervised word sense disambiguation, Journal of American Medical informatics association

[2] Fernández A., C. Valdés, R. Claramunt, A. Batalla And J. Tormo (2004) Automatic acquisition of sense examples using Exretriever

[3] Pedersen T. (2000) A Simple Approach to Building Ensembles of Naive Bayesian Classifiers for Word Sense Disambiguation, NAACL 2000 Proceedings of the 1st

North American chapter of the Association for Computational Linguistics conference, 63-69

[4] Borah, P., G. Talukdar And A. Baruah (2014) Assamese Word Sense Disambiguation using Supervised Learning, Contemporary Computing and Informatics (IC3I)

[5] Fan D., Z. Lu, R. Zhang And X. Li (2008) Word Sense Disambiguation method based on probability model improved by information gain , Intelligent Control and Automation

[6] Wai P. (2011) Myanmar to English verb translation disambiguation approach based on Naïve Bayesian classifier, Computer Research and Development (ICCRD)

[7] Navigli R. (2009) Word sense disambiguation: A survey, ACM Computing Surveys (CSUR), 2009

[8] HEARST M. (1991) Noun homograph disambiguation using local context in large text corpora, In Proceedings of the 7th Annual Conference of the UW Centre for the New OED and Text Research: Using Corpora

[9] Yarowski D. (1995) Unsupervised word sense disambiguation rivaling supervised methods, In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics

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

[11] Feili, H. And M. Soltani (2010) Word sense disambiguation based on statistical methods, 15th National CSI Computer Conference

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

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

[14] Mahmoodvand M. And M. Hourali (2015) Persian Word Sense Disambiguation Corpus Extraction Based on Web Crawler Method, Advances in Computer Science : an International Journal, Vol 4. Issue 5. 101-106

[15] Yarowsky D. (1994) Homograph disambiguation in speech synthesis, ESCAIIEEE Workshop on speech synthesis