Persian Word Sense Disambiguation

Mandana Hamidi[1], Ali Borji[2], Saeed Shiry Ghidary[3]
[1]Department of Computer and Information Technology, Azad University of Qazvin
[2]School of Cognitive Sciences, Institute for Studies in Theoretical Physics and Mathematics
[3]Computer Engineering Department, Amirkabir University of Technology
mandana.hamidi@gmail.com, borji@ipm.ir, shiry@ce.aut.ac.ir

Abstract: *Word Sense Disambiguation (WSD) aims to identify the correct sense of an ambiguous word in a sentence. In this work we investigate the performance of two state-of-the art approaches: k-NN and Naïve Bayes for the purpose of Persian word sense disambiguation. These methods have been evaluated on two highly frequent and ambiguous words from "Hamshahri"- by some means a standard corpus for Persian language. We performed experiments on both stemmed and non-stemmed version of the corpus. The results show the superiority of k-NN algorithm over Naïve Bayes in almost all cases. Although the results demonstrate good performance, further investigation should be done, by trying other classification methods and also other features used in the literature.*

Keywords: Word sense disambiguation, Persian language, WSD.

1 Introduction

In general terms, word sense disambiguation (WSD) involves the association of a given ambiguous word in a text with a definition or meaning (sense) which is distinguishable from other meanings potentially attributable to that word [1]. It is not usually considered as a goal itself, but acts like an intermediate task to benefit other natural language tasks like machine translation, information retrieval, etc.[2].

The task involves two steps: (1) the determination of all the different senses for every word relevant (at least) to the text under consideration, and (2) a mechanism to assign each occurrence of a word to the proper sense [3].

Much recent work on WSD relies on predefined senses for step (1), including:

- a list of senses such as those found in everyday dictionaries,
- a group of features, categories, or associated words (e.g., synonyms, as in a thesaurus),
- an entry in a transfer dictionary which includes translations in another language.

Some important features used in WSD literature are part of speech of words (noun, verb, etc), collocation features which take the location of words into consideration and co-occurrence features with the only thing being important, is the location of words. In this study we have done a basic research using simple co-occurrence statistics for the purpose of Persian word sense disambiguation. In future studies, part of speech features will be used to enhance performances [13].

The precise definition of a sense is, however, a matter of considerable debate within the community. The variety of approaches for defining senses has raised recent concern about the comparability of much WSD work, and given the difficulty of the problem of sense definition, no definitive solution is likely to be found soon [4].

Step (2), the assignment of words to senses, is accomplished by using two major sources of information:

- the *context* of the word to be disambiguated, in the broad sense: this includes information contained within the text in which the word appears, together with extra-linguistic information about the text such as situation, etc,
- *External knowledge sources,* including lexical encyclopedic resources, as well as hand-devised knowledge sources, which provide useful data to associate words with senses.

All disambiguation work involves matching the context of the instance of the word to be disambiguated with either information from an external knowledge source (*knowledge-driven* WSD), or information about the contexts of previously disambiguated instances of the word derived from corpora (*data-driven* or *corpus-based* WSD).

Although there has been much effort on WSD task for English language, very little work has been

reported for Persian language yet (maybe because there is no standard stemmer and POS-tagger for Persian language yet). As two languages are very different in many aspects and different methods has to be developed for many tasks like machine translation, stemming, etc, it would be interesting to check if current methods that proved to have good performance on English language also show good accuracy on Persian language or not.

The rest of the paper is organized as follows: in section two classifications methods used in the experiments are reviewed in brief. Details of experiments and dataset are explained in section three. Section four shows some results and finally conclusions and future works are discussed in section five.

2 Methods

In supervised corpus-based methods a system is presented with a training set consisting of a set of input contexts labeled with their appropriate senses (disambiguated corpus). The task is to build a classifier which correctly classifies new cases based on their context of use [5]. Here two classification methods which are simple to understand but have good performance on disambiguation are briefly explained.

2.1 Naive Bayes Classifier

This method was been introduced in [6]. In this frame the context of a word w is treated as a bag of words without structure. What we want to find is the best sense s_k for an input context c_{new} of an ambiguous word w. This is obtained as:

$$s' = \arg\max_{s_k} P(s_k \mid c_{new}) = \arg\max_{s_k} \frac{P(c_{new} \mid s_k) \times P(s_k)}{P(c_{new})} \quad (1)$$

$$= \arg\max_{s_k} P(c_{new} \mid s_k) \times P(s_k)$$

The independence assumption (Naive Bayes assumption) is that occurrence of a word v in a context is independent of other words. Thus the probability of context c_{new} given sense s_k becomes:

$$P(c_{new} \mid s_k) = P(\{v_i \mid v_i \in c_{new}\} \mid s_k) = \prod_{v_i \in c_{new}} P(v_i \mid s_k) \quad (2)$$

This assumption (often referred to as a bag of words model) has two consequences:

- the structure and order of words in context is ignored,
- the presence of one word in the context doesn't depend on the presence of another.

This is clearly not true, but there are a large number of cases in which the algorithm works well.

Finally,

$$s' = \arg\max_{s_k} P(s_k) \prod_{v_i \in c_{new}} P(v_i \mid s_k) \quad (3)$$

Thus the supervised algorithm is:

- TRAINING: Calculate:

$$P(s_k) = \frac{C(s_k)}{nr\ of\ contexts}\ ;\ P(v_i \mid s_k) = \frac{C(v_i, s_k)}{C(s_k)} \quad (4)$$

$C(s_k)$: *number of contexts labeled with sense* s_k

$C(v_i, s_k)$: *number of contexts with label* s_k *and containing word* v_i

- TEST: Calculate for a new context cnew the appropriate sense:

$$s' = \arg\max_{s_k} P(s_k \mid c_{new}) \quad (5)$$

$$= \arg\max_{s_k} P(s_k \mid c_{new}) \times \prod_{v_i \in c_{new}} P(v_i \mid s_k)$$

2.2 k-NN or Memory Based Learning

At training time, a k-NN model memorizes all the contexts in the training set by their associated features. Later, when faced with a new context $\vec{c}_{new}$, the classifier first selects k contexts in the training set that are closest to $\vec{c}_{new}$, then picks a sense for $\vec{c}_{new}$.

This supervised algorithm is:

- TRAINING: Calculate $\vec{c}$ for each context c.
- TEST: Calculate:

$$A = \{\vec{c} \mid sim(\vec{c}_{new}, \vec{c}) \text{ is maximum}, |A| = k\} \quad (6)$$

that means A is the set of the k nearest neighbors contexts of $\vec{c}_{new}$.

$$Score(c_{new}, s_j) = \sum_{c_i \in A} (sim(\vec{c}_{new}, \vec{c}_i) \times a_{ij}) \quad (7)$$

$$a_{ij} = \begin{cases} 1 & \text{if } c_i \text{ has sense } s_j \\ 0 & \text{Otherwise.} \end{cases}$$

Then the final sense s' for context $\vec{c}_{new}$ is calculated as follows:

$$S' = \arg\max_j Score(c_{new}, S_j). \quad (8)$$

3 Experiments

3.1 Dataset

In our experiments, both approaches were evaluated on the Hamshahri corpus[1]. This corpus is collected from news archive of Hamshahri newspaper in Iran. The collection contains 190,206 articles covering the following subject categories: politics, city news, economics, reports, editorials, literature, sciences, Society, foreign news, sports, etc. The size of the documents varies from short news (under 1 KB) to rather long articles (e.g. 140 KB) with the average of 1.8 KB per article [7].

Because of the lack of a sense-tagged Persian corpus we had to manually label part of the corpus. In order to check the effect of stemming on performance we also manually formed a stemmed version of the same part. This part was chosen in a way to capture the variation of the whole corpus and to be a good representative.

We used a set of four highly frequent and ambiguous words in our experiments. All sentences containing each of these words were saved to separate text files. After elimination of stop words and punctuation symbols and also stemming each file, we extracted contexts and labelled them manually according to the sense of the target word. Extracted contexts were then fed to the classifiers for the purpose of classification.

Selected words are described in the left hand–side of table 1. Numbers in front of words senses is their frequency.

Since our goal is to obtain a classifier for each word, each row represents a classification problem. Second and third columns of the table 1 show number of training examples and the percentage of the most frequent sense for each word, i.e. the accuracy that a naïve "Most–Frequent–Sense" classifier would obtain [8].

Table 1: Set of Selected Words

Word	*Senses*	*Exs.*	*MFS %*
"شیر"	lion:651,milk:956,valve:459	2066	46.27
"تار"	music instrument:494, cobweb:113, beat:76, gloomy: 232, string: 288	1203	41.06
"کرم"	worm:204, generosity:145, crème:97, cream:128	574	35.54
"مهر"	Persian month:311, seal: 218, marriage portion:130, affection:121	780	39.87

3.2 Representation of Contexts

A widely used model in corpus based sense disambiguation is vector space model [9, 10]. In this model text is considered as a vector of some features. In [5] a number of some used features in literature are listed. A common denominator between the methods is that they excavate information using co-occurrence and collocation statistics. Unlike co-occurrence, collocation type of features makes use of exact position of words in the context.[7, 5,11]

The famous saying: "meaning is use" means that to understand the meaning of a word one has to consider its use in the frame of a context. A context C is represented as a vector $\vec{c}$ of some features. The context size can vary from one word at each side of the focus word to a more "window" or even the complete sentence.

We represent a context as: $\vec{c} = (w_1, ..., w_{|W|})$ where w_i is term frequency of word v_i if it occurs in context C, or 0 otherwise, where v_i is a word from the entire text of |W| words. Here the features are all the words in the contexts meaning that the context size is equal to the length of a dictionary made by all the words from all contexts.

The similarity between two contexts C_a and C_b (of the same word or different words) is the normalized cosine between the vectors $\vec{c}_a$ and $\vec{c}_b$ [5, 12]:

$$\cos(\vec{c}_a, \vec{c}_b) = \frac{\sum_{j=1}^{m} w_{a,j} \times w_{b,j}}{\sqrt{\sum_{j=1}^{m} w^2_{a,j} \times \sum_{j=1}^{m} w^2_{b,j}}} \quad (9)$$

and $sim(\vec{c}_a, \vec{c}_b) = \cos(\vec{c}_a, \vec{c}_b)$.

[1] http://www.insf.org/download/hamshahri/

4 Results

For performing classification with k-NN, we randomly partitioned contexts to train and test sets. Size of train and test sets were selected as 3/5 and 2/5 of all examples. To classify each word in test set a weighted vote of three similar contexts from training set were used. The results reported are averaged among five different runs.

For the Naïve Bayes classifier we used 10-fold cross validation approach. Each time one part was considered as test and training was done with other remaining parts. Then the trained classifier was used to classify the test part. This process was done for all parts and then performance was averaged over all parts.

Table 2 shows the accuracy of methods for selected words both on stemmed and non-stemmed corpus. Each number in a cell is the average accuracy achieved over all context sizes.

To test the effect of context size we performed same experiments from small values for context size to large ones. As it can be seen from figures 1 and 2, increasing context size leads to higher accuracies. As increasing context size increases performance it needs a more amount of computation as well. Tables 3 and 4 show the exact performances in percentages.

It's worth mentioning that when growing context sizes, because of avoiding interference of contexts with different senses, we put contexts with same senses besides together and then increased the context size.

KNN classifier has better performance than Naïve Bayesian classifier in almost all context sizes over both words. Stemming also outperformed results in both classifiers over two words.

Table 2. Disambiguation Performance

Word	*Accuracy (%)*				
	MFS	*Stemmed*		*Non-Stemmed*	
		k-NN	*NB*	*k-NN*	*NB*
"شیر"	46.27	82.50	75.25	70.87	61.11
"تار"	41.06	83.49	68.48	76.26	59.88
"کرم"	35.54	76.80	59.32	66.03	54.50
"مهر"	39.87	67.80	57.5	63.40	56.13
Avg.	*40.68*	*77.64*	*65.13*	*69.14*	*57.90*

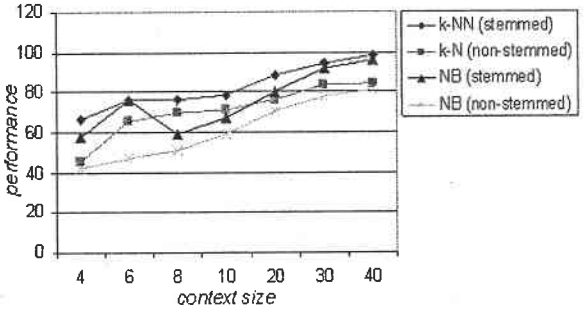

Figure 1. Disambiguation for Word "شیر"

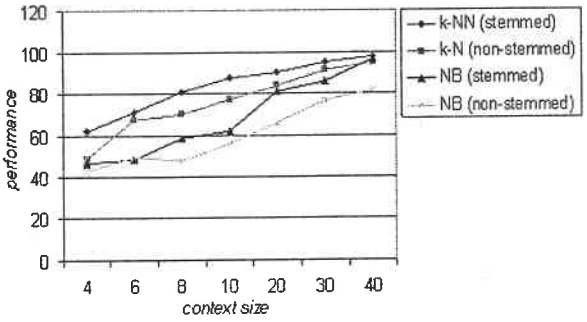

Figure 2. Disambiguation for Word "تار"

5 Conclusions and Future Work

In this work we used k-NN and Naïve Bayes classification methods for sense disambiguation of Persian words. K-NN approach showed greater performance over Naïve Bayes for all words in both stemmed and non-stemmed cases. It means that stemming could make the performance better by reducing the number of features. Experimenting with larger datasets reveals the effect of stemming more reliably. Increments in context size leads to better performance but in the expense of higher computation complexity.

Although the results are very promising, further investigation should be done. As a future work, one can consider using other classification methods like decision trees, artificial neural networks, support vector machine along with other features like: POS (Part of Speech) of words, tf-idf, collocation, , for disambiguation.

It is also necessary to develop a standard tagged corpus (Stemmed and POS-tagged) as a benchmark for further Persian word sense disambiguation studies.

Table 3. Effect of Context Size on Classification Performance for Word "شیر"

Context size	*4*	*6*	*8*	*10*	*20*	*30*	*40*
k-NN (stemmed)	66.24	75.87	76.27	78.92	87.99	94.42	97.81
k-NN (non-stemmed)	45.09	65.46	70.10	71.40	75.88	83.78	84.42
NB (stemmed)	57.31	76.03	58.82	67.59	80.16	91.27	95.61
NB (non-stemmed)	41.76	47.21	51.12	59.20	70.27	77.51	80.76

Table 4. Effect of Context Size on Classification Performance for Word "تار"

Context size	*4*	*6*	*8*	*10*	*20*	*30*	*40*
k-NN (stemmed)	62.09	71.34	80.81	87.34	90.13	95.12	97.65
k-NN (non-stemmed)	49.09	67.46	69.90	77.24	83.89	91.28	94.98
NB (stemmed)	46.86	48.53	58.82	61.76	80.65	85.88	96.89
NB (non-stemmed)	42.86	49.21	47.62	56.03	65.73	76.38	81.34

References

[1] M. Ide and J. Veronis, "Word Sense Disambiguation: The State of the Art". *Introduction to the special issue on WSD: the state of the art. Computational Linguistics*, pp1-40, 24(1) 1998.

[2] G. Serban and D. Tatar, "UBB system at Senseval3 ", *Proceedings of Workshop in Word Disambiguation*, ACL 2004, Barcelona, pp 226-229, July 2004.

[3] H. Schutze, "Automatic Word Sense Discrimination". *Computational Linguistics*, pp97-123, 24(1) 1998.

[4] E. Kelly and S. Philip, *Computer Recognition of English Word senses*. North-Holland Pub, Amsterdam 1975.

[5] D. Tatar, "Word Sense Disambiguation By Machine Learning Approach: A Short Survey". *Studia Univ. Babes Bolyai, Informatica*, Volume XLIX, Number 2, 2004.

[6] W. Gale, K. Church and D. Yarowsky, "A method for disambiguating word senses in large corpus". *Computers and the Humanities*, pp 415-439, (26) 1992.

[7] F. Oroumchian, E. Darrudi and M.R. Hejazi, "Assessment of a modern Farsi corpus,". *Proceedings of The 2nd Workshop on Information Technology & its Disciplines* (WITID), ITRC, Iran, 2004.

[8] G. Escudero, L. Marquez and G. Rigau, "Naive Bayes and Exemplar-Based Approaches to Word Sense Disambiguation Revisited,". *In Proceedings of the 14th European Conference on Artificial Intelligence, ECAI*, Berlin, Germany, 2000.

[9] N. Hwee Tou and L. Hian Beng, "Integrating multiple knowledge sources to disambiguate word sense: An examplar based approach". *Computational Linguistics* , 24-27 40-47. 1996

[10] P. Edmonds and S. Cotton. "Senseval-2: Overview,". *In Proceedings of the Second International Workshop on evaluating Word Sense Disambiguation Systems*, Toulouse, France. 2001.

[11] D. Martinez and E. Agirre. "One sense per collocation and genre/topic variations,". *In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, Hong Kong. 2000.

[12] D. McCarthy, R. Koeling, J. Weeds, and J. Carroll. "Finding predominant word senses in untagged text,". *In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, Barcelona, Spain. 2004.

[13] S. Tasharofi, H. Hojjat, H. Amiri and F. Raja, "Creating a Feasible Corpus for POS Tagging", Technical Report, February, 2006.