STM JOURNALS

Journal of Artificial Intelligence Research & Advances
ISSN: 2395-6720 (Online)
Volume 5, Issue 1
www.stmjournals.com

A Survey on Word Sense Disambiguation for Persian Language

Suliman Ameryani**, Nirav M. Raja, Vishal Prajapati
Department of Information Technology, G.H. Patel College of Engineering and Technology, Anand, Gujarat, India

Abstract

Natural language processing (NLP) is considered as one of the most important subfields of artificial intelligence. It is used to provide an interface between computer and human languages. NLP has found its application in several areas such as speech processing, summation of text, machine translation, searching, information grouping, parts of speech tagging (POS), sentiment analysis etc. Natural language processing is posing some challenges in speech recognition, ambiguity in preposition attachment, co-reference resolution, information extraction, question answering, word sense disambiguation (WSD) etc. Many algorithms have been proposed so far to address the mentioned problems in NLP and some techniques for WSD too. Word sense disambiguation (WSD) is the process of finding correct sense for a word from a collection of semantic tags (senses) in a sentence. It has applications in machine translation, information retrieval, information extraction and knowledge acquisition. Along with the other languages like English, Spanish, French and Hindi that are the focus points for most of the researchers, some work has been done for Persian language too. This paper aims to describe some aspects of NLP and approaches that applied on WSD for Persian language with analysis.

Keywords: *Natural language processing, word sense disambiguation, ambiguity, Persian language*

****Author for Correspondence*** E-mail: suliman.melad@gmail.com

INTRODUCTION

WSD is conceding to be an open and AI-hard problem in natural language processing. The natural languages are essential components of everyday communications. However many ambiguous words exists in natural languages. We have many polysemous words having different senses based on its usage in the context. Ambiguous words are easy to understand by human beings but for a machine, it is an immense task to choose the most appropriate meaning of the word having multiple meanings. For example in Persian language the word سیر "seer" and سیر "seer" have different meanings as it means Garlic and being full. The ambiguity in natural language makes the word sense disambiguation a challenging task. The problem of WSD has been explored ever since computers have been used to solve problems in human language, with some of the earliest work in the 1940s on machine translation [1]. Word sense disambiguation contribute many applications in major areas like information retrieval (IR), information extraction (IE), machine translation (MT), question answering (Q&A), text mining, phone help systems, knowledge mining/acquisition, bioinformatics, semantic web, lexicography etc. [1].

Natural Language Processing

Natural language processing is a domain of computer science, artificial intelligence and computer linguistics. NLP provides interface between computers and human languages. NLP is facing some sort of challenges such as natural language understanding, natural language generation, speech recognition, dialog systems, etc. [2].

Natural Language Processing Tasks

There are some important tasks done by NLP to understand a language by machine in more accurate way, which is given below. In NLP task there are multiple categories like syntax, semantics, discourse and speech [3].

Syntax

1. *Part of Speech Tagging (POS):* POS tagging is the basic task in every context of natural language processing. POS tagging falls under syntax task of NLP.
2. *Parsing:* Parsing is a conventional syntax task of NLP that involves breaking down the sentence into its component parts of speech and details about forms, tense, function and grammatical relationship.
3. *Word Segmentation:* Word segmentation is a basic procedure of separating a sentence into isolated words.

Semantics

1) *Lexical Semantics:* It defines computational meaning of individual words in sentence.
2) *Machine Translation:* The task of automatically converting one natural language to another is known as machine translation.
3) *Name Entity Recognition:* In a given sentence, determining the proper nouns, such as people or places is known as name entity recognition. In a language like English, we can easily determine named entity by capitalization of word. However, in Persian language, capitalization of word does not exist, so it is a complex task.
4) *Natural Language Understanding:* For a machine to understand natural languages with all the sematic relationships, noun, verb, tenses, stemmed words and other parts of speech is a complex task as compared to human.

Question Answering

In natural language processing, question answering system tries to find the answer for the question asked by human beings. For a machine it is hard to find answer for unrestricted questions like "what is the meaning of life?"

Relationship Extraction

It is the task of finding out the relations between the words given in a sentence, like for example, who is brother of whom.

Word Sense Disambiguation

Word sense disambiguation is the process of finding correct sense for a word from a collection of semantic tags (senses) in sentence.

Word Senses Disambiguation

Word sense disambiguation aims to identify the correct sense of an ambiguous word in a sentence [4]. The task comprises of two steps: first to determine all the possible senses for each word in a given context; second, automatically assigning accurate sense to each word.

For first step, to gather all the possible senses of a word we can include the following components:

- List of senses from everyday dictionary.
- A group of categories, features, associated words like synonyms and thesaurus [4].
- A dictionary for translation.
  For second step, there are further two major sources for assigning correct sense to ambiguous word.
- A broad source includes situation of context, information about word that appears in context and other linguistic information.
- External knowledge source includes lexical dictionary and also handmade knowledge sources [4].

WSD Features [2]

The word features can be collected from the sliding window in the sentence for example: if we set the size of sliding window to 5 then the value of the features will be wi-2, wi-1, wi, wi+1, wi+2 [5].

- Collocation: The set of words surrounding the ambiguous word [6].
- Bag of words features: In NLP, bag of word is used as simplified representation in which the corpus or sentence is illustrated as the bag (multiset) of its word, neglecting grammar and order of words, but keeping multiplicity.
- Domain features: To identify the most relevant domain in corpora using FarsNet (Persian WordNet).
- Co-occurrence: The set of words that have occurred frequently in a text is known as co-occurrence [6].

WSD Approaches

1) Knowledge based approach: Knowledge based approaches rely on the information contained in machine readable dictionaries, treasuries, WordNet [6].

2) Machine learning approach: Machine learning based approaches can be further classified into three categories [6]:

a) Supervised [6]: Supervised approaches use data-mining and machine learning techniques to train a classifier from already sense tagged corpora.
b) Unsupervised [6]: Unsupervised approaches are based on unlabeled corpora. These approaches do not make use of training corpus.
c) Semi-Supervised [6]: It is a hybrid approach of both supervised and unsupervised approaches.

Also, there is hybrid approach which uses knowledge based and corpus based approach to eliminate the challenge of knowledge acquisition problem.

WSD Applications

- Information retrieval [7],
- Machine translation [7],
- Speech recognition [1],
- Information extraction [1],
- Spelling correction [1],
- Speech processing [1],
- Text mining [1],
- Lexicography [1], and
- Text summarization [1].

Persian Language

Persian, also known as Parsi or Farsi belongs to the Indo-Iranian languages, a subfamily of the Indo-European languages [8]. There are approximately 110 million Persian speakers worldwide [9], with the language holding official status in Iran (Farsi), Afghanistan (officially known as Dari since 1958), and Tajikistan (referred to as Tajiki).

Corpus For Persian Language

In linguistics, a corpus or text corpus is a huge and structured set of texts. An annotating corpus is part-of-speech tagging, or POS-tagging within which info regarding every word is a part of speech (verb, noun, adjective, etc.), is added to the corpus in the form of tags. For Persian language, there are two mostly used corpuses, Dr. Bijankhan's Corpus and Hamshahri Corpus.

The Hamshahri Corpus (پیکره همشهری in Persian) is a sizable Persian corpus based on the Iranian newspaper [10]. The collection contains more than 160,000 articles with several categories. 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. There exist two versions of Hamshahri Corpus, Version 1.0 and Version 2.0 with more features like included images, bigger size and categorized news.

Performance Measures

Precision

Precision is defined as the ratio of the correctly disambiguated homographs to the total number of answered homographs [11].

Recall [7]

Recall can be defined as the ratio of the correctly disambiguated homographs to the total number of homographs [11].

F1-measure

F1 measure considers both the precision and recall and can be calculated as:

$$F1 - measure = \frac{2 * precision * recall}{precision + recall} \quad [5]$$

LITERATURE SURVEY

Sarrafzadeh *et al.* described that sense tagged corpora plays a crucial role in task of word sense disambiguation because it requires human efforts [7]. Such sense tagged corpora are not available for many languages including Persian. So, for enhancing supervised and semi-supervised approaches, they proposed sense tagged corpora in Persian. They used cross lingual word sense tagging to tag Persian words using English tagged words in parallel corpus. It has high accuracy but relatively low recall. Then they applied direct knowledge to tag other words. They performed tokenization, lemmatization, and POS tagging as pre-processing steps. To increase the number of tagged words in their corpus they used extended Lesk algorithm. In their evaluation they manually assigned the correct tag to the wrong tags given by their proposed technique. Their evaluation result gives good accuracy of 91%, but a relatively low recall of 46%.

Sarrafzadeh *et al.* proposed cross lingual approach for Persian word sense disambiguation [12]. The proposed method

overcomes the knowledge resource deficiency. They invented novel approach in terms of plausibility, feasibility and performance. They first applied tag sense to English words, then they translated the words into Persian and applied sense using cross-lingual approach. Then they compared result with supervised extended Lesk algorithm and showed that cross-lingual gives better results.

Hamidi *et al.* have gone through two classic approaches, Naïve Bayes and k-NN for Persian word sense disambiguation [4]. They also showed basic steps for eliminating ambiguity from context. They have shown important features like part of speech tagging, co-occurrence and collocation in WSD. They provided two major sources of information from which we can get different senses of ambiguous words. They have given comparative analysis of Naïve Bayes and k-NN approaches and evaluated on Hamshahri corpus. In Persian language, there is lack of sense tagged corpus, so they have manually labelled the corpus. They proved superiority of k-NN over Naïve Bayes approach in both stemmed and non-stemmed cases, and suggested that stemming can improve the performance along with other features. They found out accuracy of k-NN and Naïve Bayes algorithms that proved that k-NN classifier has better performance in all context sizes as compared to Naïve Bayes.

Miangah used mutual information statistics obtained from very large Persian corpus [13]. Mutual information values are calculated based on co-occurrence frequency and correlation of words. This mutual information presents degree of semantic relation between words. The two words having highest mutual information are considered to be most suitable sense of the polysemy words. They showed from their evaluation that the proposed work provides 91.24% accuracy, hence more suitable for cross-lingual information retrieval and machine translation.

Seraji proposed statistical part of speech tagger HunPOS trained on a Persian corpus [14]. HunPOS tagger is an extended version of TnT. Users can arrange tagger by different features based on language type. The tagger measures lexical probabilities based on current and previous tags. HunPOS embed morphological analyzer to precise the list of different possible text that the algorithm needs to deal with, as a result it speedup the process with highly improved precision. They achieved higher accuracy as compared to bigram models. The experimental results prove an overall 96.9% accuracy for Persian language. We can conclude that as compared to TnT part of speech tagger, HunPOS is a better option.

Kardan and Imani described difficulties for POS tagging such as disambiguating unknown words, multi category words, and Persian language is free order language with its own characteristics [5]. These difficulties can greatly affect the quality of POS tagging. They described some common challenges that POS tagging is facing for Persian language as compared to English language. The proposed method involves two steps: feature extraction and POS tagging. They used maximum entropy tagger as a classifier with finite set of input features and class labels. They used Bijankhan's tagged corpus for their experiment. They showed accuracy comparison between the proposed POS tagging method and previous methods like Memory based, Markov Model, TnT and Maximum likelihood. We can conclude that their proposed POS tagging approach outperforms other Persian taggers with accuracy of 97.53%.

Yaghoub-Zadeh *et al.* explained the importance of stop word removing as it is a primary task to reduce the index size in text processing systems [15]. They proposed an aggregate method for automatically building stop-word list for Persian information retrieval system. The proposed approach uses part of speech tagging and analyzing statistical features to enhance the accuracy of retrieval and minimize risk of removing informative words. The experimental result proved that the proposed approach enhances the average precision, decreases the index storage size and improves the overall response time.

Rezapour *et al.* classified WSD approaches into three categories: supervised approach,

unsupervised and semi-supervised approaches [6]. They proposed supervised WSD approach based on k-NN algorithm in which manually sense tagged data is used to train the classifier with two sets of features: collocation and co-occurrence. They proposed feature weighting strategy to improve the classification accuracy. They compared the accuracy of their proposed work with other WSD methods that are based on corpora. We can conclude that the proposed method has less accuracy as compared to Degan method.

Agirre and Lacalle proposed combination of k-NN with singular value decomposition (SVD) where each classifier learns from different set of features: local features (syntactic, collocation features), topical features (bag of words) and latent features based on SVD [16]. In their experiment, they tried to optimize parameters like number of neighbors, SVD dimensions and value of k. Because of slight overfitting on the training data, their all word system did not perform so well.

Jani and Pilevar proposed hybrid approach of supervised and unsupervised WSD. Their proposed work was based on conceptual categorization of context [17]. Therefore, they first categorized the ambiguous words. Then they constructed the related context discriminator and, since each word sense belongs to one category, the correct sense can be predicted. 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.

Riahi and Sedghi showed that supervised method needs large tagged corpus which is not available for some languages like Persian, so they used semi-supervised method with small tagged corpus and large untagged corpus to solve the problem [11]. They provided coarse grained WSD with supervised decision list to train the classifier and tri-training as semi-supervised method. They used three classifiers with different features in tri-training. They used manually extracted collocation features to train decision list. They used Hamshahri tagged and untagged corpus. They mentioned that when tagged data is large, semi-supervised method decreases the performance because of their uncertain nature. Result shows semi-supervised tri-training decision list result compared with supervised DLL and showed the superiority of supervised over semi-supervised approaches with accuracy of 95.87%.

Soltani and Faili proposed new approach for resolving lexical ambiguity using statistical information gained from corpus [18]. They proposed unsupervised graph-based approach using bilingual dictionary. They introduced new method of source corpora and target language corpora to measure semantic similarity. The experiments show that the unsupervised graph-based WSD which uses the proposed semantic similarity measured in the dependency graph outperforms all other methods on WSD for translating English to Persian words, significantly. Saedi and Shamsfard proposed hybrid of knowledge based and corpus based approach [19]. They applied WSD in between process of machine translation between English and Persian.

CONCLUSIONS

This survey presents an overall introduction of natural language processing (NLP), some natural language processing tasks (Syntax, semantics, discourse and speech), applications of NLP and most complex challenges in the era of natural language processing. Under NLP tasks, the word sense disambiguation is more focused. Word sense disambiguation (WSD) aims to identify the correct sense of an ambiguous word in a sentence [1]. We have gone through approaches in word sense disambiguation and some applications of WSD. Some details about Persian language have also given with grammatical points and information about sense tagged corpus that is available for Persian language (Hamshahri corpus).

REFERENCES

1. Bhala R, Abirami S. Trends in Word Sense Disambiguation. Springer; Mar 2012.

2. Pathak R, Thankachan B. Natural Language Processing Approaches, Application and Limitations. *International Journal of Engineering Research & Technology (IJERT)*. Sep 2012; Vol. 1.
3. Indurkhya Nitin, Damerau Fred J, editors. *Handbook of Natural Language Processing*. Chapman and Hall/CRC; 2010.
4. Hamidi M, Borji A, Ghidary S. Persian Word Sense Disambiguation. *15th ICEE Proceeding*. 2014.
5. Kardan A, Imani M. Improving Persian POS Tagging Using the Maximum Entropy Model. *Intelligent Systems (ICIS), 2014 Iranian Conference on*. IEEE; 2014.
6. Rezapour A, Fakhrahmad S, Sadreddini M. Applying Weighted KNN to Word Sense Disambiguation. *Proceedings of the World Congress on Engineering*, ResearchGate. 2011; Vol. 3.
7. Sarrafzadeh B, Yakovets N, Cercone N, *et al*. Towards Automatic Acquisition of a Fully Sense Tagged Corpus for Persian. *ResearchGate*. Jan 2017.
8. Shamsfard M, Hesabi A, Fadaei H, *et al*. Semi Automatic Development of FarsNet; The Persian WordNet. *ResearchGate*. Feb 2014.
9. Shamsfard M. Challenges and Open Problems in Persian Text Processing. Unpublished.
10. Rekabsaz N, Sabetghadam S, Lupu M, *et al*. Standard Test Collection for English-Persian Cross-Lingual Word Sense Disambiguation. Unpublished.
11. Riahi N, Sedghi F. A Semi-Supervised Method for Persian Homograph Disambiguation. *20th Iranian Conference on Electrical Engineering, (lCEE2012)*. IEEE; 2012.
12. Sarrafzadeh B, Yakovets N, Cercone N, *et al*. Cross Lingual Word Sense Disambiguation for Languages with Scarce Resources. *ResearchGate*. Oct 2017.
13. Miangah T. Solving Polysemy Problem of Persian Words Using Mutual Information Statistics. Unpublished.
14. Seraji M. A Statistical Part-of-Speech Tagger for Persian. *NODALIDA 2011 Conference Proceedings*. 2011; 340–343p.
15. Yaghoub-Zadeh-Fard M, Minaei-Bidgoli B, Rahmani S, *et al*. PSGW: An Automatic Stop-Word list Generator for Persian Information Retrieval Systems Based on Similarity Function and POS Information. IEEE; 2015.
16. Agirre E, Lacalle O. UBC-ALM: Combining k-NN with SVD for WSD. *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*. Jun 2007; 342–345p.
17. Jani F, Pilevar AH. Word Sense Disambiguation of Persian Homographs. In *Proceedings of the 7th International Conference on Software Paradigm Trends (ICSOFT-2012)*. 2012; 328–331p.
18. Soltani M, Faili H. A Statistical Approach on Persian Word Sense Disambiguation. Unpublished.
19. Saedi C, Shamsfard M. Translating Persian Documents into English Using Knowledge based WSD. *Digital Information Management, 2009. ICDIM 2009. Fourth International Conference on*. IEEE; 2009.

Cite this Article
Suliman Ameryani, Nirav M. Raja, Vishal Prajapati. A Survey on Word Sense Disambiguation for Persian Language. *Journal of Artificial Intelligence Research & Advances*. 2018; 5(1): 12–17p.