YORK
UNIVERSITÉ
UNIVERSITY

redefine THE POSSIBLE.

Cross Lingual Word Sense Disambiguation for Languages with Scarce Resources

Bahareh Sarrafzadeh, Nikolay Yakovets, Nick Cercone, Aijun An

Technical Report CSE-2011-01

January 24 2011

Department of Computer Science and Engineering
4700 Keele Street, Toronto, Ontario M3J 1P3 Canada

Cross Lingual Word Sense Disambiguation for Languages with Scarce Resources

Bahareh Sarrafzadeh, Nikolay Yakovets, Nick Cercone, and Aijun An

Department of Computer Science, York University, Canada
{bahar, hush, nick, aan}@cse.yorku.ca

Abstract. Word Sense Disambiguation (WSD) has long been a central problem in computational linguistics. WSD is the ability to identify the meaning of words in context in a computational manner. Statistical and supervised approaches require a large amount of labeled resources as training datasets. In contradistinction to English, the Persian language has neither any semantically tagged corpus to aid machine learning approaches for Persian texts, nor any suitable parallel corpora. Yet due to the ever-increasing development of Persian pages in Wikipedia, this resource can act as a comparable corpus for English-Persian texts.
In this paper, we propose a cross lingual approach to tagging the word senses in Persian texts. The new approach makes use of English sense disambiguators, the Wikipedia articles in both English and Persian, and a newly developed lexical ontology, FarsNet. It overcomes the lack of knowledge resources and NLP tools for the Persian language. We demonstrate the effectiveness of the proposed approach by comparing it to a direct sense disambiguation approach for Persian. The evaluation results indicate a comparable performance to the utilized English sense tagger.

Keywords: Word Sense Disambiguation, WordNet, Languages with Scarce Resources, Cross Lingual, Extended Lesk, FarsNet, Persian

1 Introduction

Human language is ambiguous, so that many words can be interpreted in multiple ways depending on the context in which they occur. While humans rarely think about the ambiguities of language, machines need to process unstructured textual information which must be analyzed in order to determine the underlying meaning.

WSD heavily relies on knowledge. Without knowledge, it would be impossible for both humans and machines to identify the words' meaning. Unfortunately, the manual creation of knowledge resources is an expensive and time consuming effort, which must be repeated every time the disambiguation scenario changes (e.g., in the presence of new domains, different languages, and even sense inventories) [1]. This is a fundamental problem which pervades approaches to WSD, and is called the *knowledge acquisition bottleneck*.

With the huge amounts of information on the Internet and the fact that this information is continuously growing in different languages, we are encouraged to investigate cross-lingual scenarios where WSD systems are also needed. Despite the large number of WSD systems for languages such as English, to date no large scale and highly accurate WSD system has been built for the Farsi language due to the lack of labeled corpora and monolingual and bilingual knowledge resources.

In this paper we propose a novel cross-lingual approach to WSD that takes advantage of available sense disambiguation systems and linguistic resources for the English language. Our approach demonstrates the capability to overcome the knowledge acquisition bottleneck for languages with scarce resources. This method also provides sense-tagged corpora to aid supervised and semi-supervised WSD systems. The rest of this paper is organized as follows: After reviewing related works in Section 2, we describe the proposed cross-lingual approach in Section 3, and a direct approach to WSD in Section 4; which is followed by evaluation results and a discussion in Section 5. In Section 6 our concluding remarks are presented and future extensions are proposed.

2 Related Work

We can distinguish different approaches to WSD based on the amount of supervision and knowledge they demand. Hence we can classify different methods into 4 groups [1]: Supervised, Unsupervised, Semi-supervised and Knowledge-based.

Generally, supervised approaches to WSD have obtained better results than unsupervised methods. However, obtaining labeled data is not usually easy for many languages, including Persian as there is no sense tagged corpus for this language.

The objective of Knowledge-based WSD is to exploit knowledge resources such as WordNet [2] to infer the senses of words in context. These methods usually have lower performance than their supervised alternatives, but they have the advantage of wider coverage, thanks to the use of large-scale knowledge resources.

The recent advancements in corpus linguistics technologies, as well as the availability of more and more textual data encourage many researchers to take advantage of comparable and parallel corpora to address different NLP tasks. The following subsection reviews some of the related works which address WSD using a cross-lingual approach.

2.1 Cross-Lingual Approaches

Parallel corpora present a new opportunity for combining the advantages of supervised and unsupervised approaches, as well as an opportunity for exploiting translation correspondences in the text. Cross lingual approaches to WSD disambiguate target words by labelling them with the appropriate translation. The main idea behind this approach is the plausible translations of a word in context restricts its possible senses to a subset [3].

In recent studies [4–7], it has been found that approaches that use cross-lingual evidence for WSD attain state-of-the-art performance in all-words disambiguation. However, the main problem of these approaches lies in the knowledge acquisition bottleneck: there is a lack of parallel and comparable corpora for several languages including Persian - which can potentially be relieved by collecting corpora on the Web. To overcome this problem, we utilized Wikipedia pages in both Persian and English. Before introducing our WSD system a brief survey of WSD systems for the Persian language follows.

2.2 Related Work for Persian

The shortcomings of efficient, reliable linguistic resources and fundamental text processing modules for the Persian language make it difficult for computer processing. In recent years there have been two branches of efforts to eliminate these shortcomings [8].

Some researchers are working to provide linguistic resources and fundamental processing units. FarsNet [9] is an ongoing project to develop a lexical ontology to cover Persian words and phrases. It is designed to contain a Persian WordNet in its first phase and grow to cover verbs' argument structures in its second phase. The included words and phrases are selected according to BalkaNet base concepts and the most frequent Persian words and phrases in utilized corpora. FarsNet 1.0 relates synsets in each POS category by the set of WordNet 2.1 relations. FarsNet also contains inter-lingual relations connecting Persian synsets to English synsets (in Princeton WordNet 3.0). [10] exploits an English-Persian parallel corpus which was manually aligned at the word level and sense-tagged a set of observations as a training dataset from which a decision tree classifier is learned. [8] devised a novel approach based on WordNet, eXtended WordNet and verb parts of FarsNet to extend the Lesk algorithm [11] and find the appropriate sense of a word in an English sentence. Since FarsNet was not released at the time of publishing this paper, they manually translated a portion of WordNet to perform WSD for the Persian side. [12] defined heuristic rules based on the grammatical role, POS tags and co-occurrence words of both the target word and its neighbors to find the best sense.

Others work on developing algorithms with less reliance on linguistic resources. We refer to statistical approaches [13–15] using monolingual corpora for solving the WSD problem in Farsi texts. Also conceptual categories in a Farsi thesaurus have been utilized to discriminate senses of Farsi homographs in [16].

Our proposed approach is unique from most cross-lingual approaches in the sense that we utilize a comparable corpus, automatically extracted from Wikipedia articles, which can be available for many language pairs even the languages with scarce resources and our approach is not limited to sense tagged parallel corpora only. Second, thanks to the availability of FarsNet, our method tags Persian words using sense tags in the same language instead of using either a sense inventory of another language or translations provided by a parallel corpus. Therefore the results of our work can be applied to many monolingual NLP tasks such as Information Retrieval, Text Classification as well as bilingual ones including Machine Translation and Cross-Lingual tasks. Moreover, the extended version of the Lesk algorithm has never been exploited to address WSD for Persian texts. Finally, taking advantage of available mappings between synsets in WordNet and FarsNet, we were able to utilize an English sense tagger which uses WordNet as a sense inventory to sense tag Persian words.

3 Introducing the Cross Lingual Approach: Persian WSD using Tagged English Words

This approach consists of two separate phases. In the first phase we utilize an English WSD system to assign sense tags to words appearing in English sentences. In the second phase we transfer these senses to corresponding Persian words. Since by design these two phases are distinct, the first phase can be considered as a black box and different English WSD systems can be employed. What is more, the corresponding Persian words can be Persian pages in Wikipedia or Persian sentences in the aligned corpus.

We created a comparable corpus by collecting Wikipedia pages which are available for both English and Persian languages and Persian articles are not shorter than 250 words. This corpus contains about 35000 words for the Persian side and 74000 words for English.

Therefore, the Cross Lingual system contains three main building blocks: English Sense Disambiguation, English to Persian Transfer and Persian Sense Disambiguation. These components are described in the following sections. Figure 1 indicates the system's architecture for the Cross Lingual section.

3.1 English Sense Disambiguation

As mentioned, different English Sense Disambiguation systems can be employed in this phase. In this system we utilized the Perl-based application SenseRelate [17] for the English WSD phase. SenseRelate uses WordNet to perform knowledge-based WSD.

This system allows a user to specify a range of settings to control the desired disambiguation. We determined that the relatedness measure that uses gloss overlaps (Extended Lesk) coupled with window size[1] of 5 led to the most accurate disambiguation.

As an input to SenseRelate we provided plain untagged text of English Wikipedia pages that was preprocessed according to application's preconditions. We also provided a tweaked stopword list that is more extensive than the one which came bundled with the application. SenseRelate will tag all ambiguous words in the input English texts using WordNet as a sense repository.

[1] The number of surrounding words for the target word

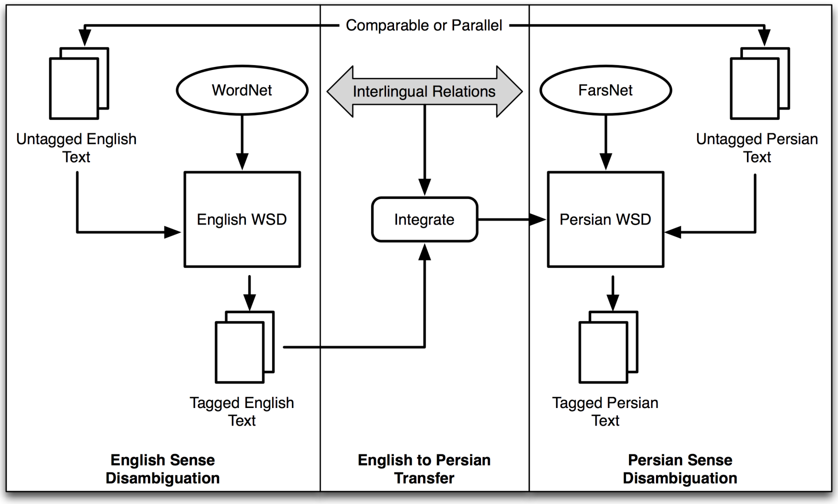

Fig. 1. Cross Lingual System Architecture

3.2 English to Persian Transfer

Running SenseRelate for input English sentences, we have English words tagged with sense labels. Each of these sense labels corresponds to a synset in WordNet containing that word in a particular sense. Most of these synsets have been mapped to their counterparts in FarsNet. In order to take advantage of these English tags for assigning appropriate senses to Persian words, first we transfer these synsets from English to Persian using interlingual relations provided by FarsNet.

Exploiting these mappings, we match each WordNet synset which is assigned to a word in an English sentence to its corresponding synset in FarsNet. For this part, we developed a Perl-based XML-Parser and integrated the results into the output provided by SenseRelate.

Along with transferring senses, we also need to transfer Wikipedia pages from English to Persian. Here, we choose the pages which are available in both languages. Hence we can work with the pages describing the same title in Persian.

3.3 Persian Sense Disambiguation

There are two different heuristics for disambiguating senses [1]:

- *one sense per collocation*: nearby words strongly and consistently contribute to determine the sense of a word, based on their relative distance, order, and syntactic relationship;
- *one sense per discourse*: a word is consistently referred with the same sense within any given discourse or document;

The first heuristic is applicable to any available parallel corpus for English Persian texts, and we can assign the same sense as the English word to its translation appearing in the aligned Persian sentence. In this case, we obtain a very high accuracy, although our system would be limited to this specific type of corpus.

Alternatively, since parallel corpora are not easy to obtain for many language pairs, we utilize Wikipedia pages which are available in both English and Farsi as a comparable corpus. We used these pages in order to investigate the performance of our system on such corpus which is easier to collect for languages with scarce resources.

Note that although Farsi pages are not the direct translation of English pages, the context is the same for all corresponding pages, which implies many common words appear in both pages. Consequently, we can assume domain-specific words appear with similar senses in both languages.

Based on the second hypothesis as the context of both texts is the same, for each matched synset in FarsNet which contains a set of Persian synonym words, we find all these words in the Persian text and we assign the same sense as the English label to them. Since there may be English words which occurred multiple times in the text and they could receive different sense tags from SenseRelate, we transfer the most common sense to Persian equivalences. Here we can use either the "most frequent" sense provided by WordNet as the "most common" sense or choose the most local frequent sense (i.e., in that particular context). Since the second heuristic is more plausible we opted to apply the most frequent sense of each English word in that text to its Persian translations. As an example consider SenseRelate assigned the second sense of the noun "*bank*" to this word in the following sentence: "*a bank is a financial institution licensed by a government.*" and this sense is the most frequent sense in this English article. The Persian equivalent noun (i.e., "*bank*") has six different senses. Among them we select the sense which is mapped to the second sense of word bank in WordNet and we assign this sense from FarsNet to "*bank*".

We consider 3 possible scenarios:

1. An English word has more than one sense, while the equivalent Persian word only has one sense. So, SenseRelate disambiguates the senses for this English word, and the equivalent Persian word does not need disambiguation. For example *"free"* in English is a polysemic word which can mean both *"able to act at will"* and *"costing nothing"*, while we have different words for these senses in Persian (*"azad"* and *"majani"* respectively). In this case we are confident that the transferred sense must be the correct sense for the Persian word.
2. Both the English and the Persian words are polysemous, so as their contexts are the same, the senses should be the same. In this case we use mappings between synsets in WordNet and FarsNet. For example, the word *"branch"* in English and its Persian equivalent *"shakheh"* both are polysemous with similar set of senses. For example, if SenseRelate assigned the 5th sense (i.e *"a stream or river connected to a larger one"*) of this word to its occurrence

in an English sentence, the mapped synset in FarsNet would also correspond to this sense of the Persian *"shakheh"*.

3. The worst case happens when an English word only has one sense, while the Persian equivalent has more than one. In this case, as the context of both texts are the same, the Persian word is more likely to occur with the same sense as the English word. For example the noun "Milk" in English has only one meaning, while its translation in Farsi (i.e., *"shir"*) has three distinct meanings: milk, lion and (water) tap. However, since SenseRelate assigns a synset with this gloss *"a white nutritious liquid secreted by mammals and used as food by human beings"* to this word, the first sense will be selected for *"shir"*.

In summary, for all 3 possible scenarios we utilize the mappings from WordNet synsets to FarsNet ones. However, according to our evaluation results, the first case usually leads to more accurate results and the third case results in the lowest accuracy. Nontheless, when it comes to domain-specific words, all three cases result in a high precision rate.

4 Direct Approach: Applying Extended Lesk for Persian WSD

Thanks to the newly developed FarsNet, the Lesk method (gloss overlap) is applicable to Persian texts as well. Since it is worthwhile to investigate the performance of this Knowledge based method - which has not as yet been employed for disambiguating Persian words - and compare the results of both Cross Lingual and Direct approaches, in the second part of this experiment, the Extended Lesk algorithm has been applied directly for Persian.

4.1 WSD using the Lesk Algorithm

The Lesk algorithm uses dictionary definitions (gloss) to disambiguate a polysemous word in a sentence context. The original algorithm counts the number of words that are shared between two glosses. The more overlapping the glosses are, the more related the senses are. To disambiguate a word, the gloss of each of its senses is compared to the glosses of every other word in a phrase. A word is assigned to the sense whose gloss shares the largest number of words in common with the glosses of the other words.

The major limitation to this algorithm is that dictionary glosses are often quite brief, and may not include sufficient vocabulary to identify related senses. An improved version of the Lesk Algorithm - Extended Lesk [18] - has been employed to overcome this limitation.

4.2 Extended Gloss Overlap

Extended Lesk algorithm extends the glosses of the concepts to include the glosses of other concepts to which they are related according to a given concept hierarchy.

Synsets are connected to each other through explicit semantic relations that are defined in WordNet. These relations only connect word senses that are used in the same part of speech. Noun synsets are connected to each other through hypernym, hyponym, meronym, and holonym relations. There are other types of relations between different part of speeches in WordNet, but we focused on these four types in this paper. These relations are also available for Persian synsets in FarsNet.

Thus, the extended gloss overlap measure combines the advantages of gloss overlaps with the structure of a concept hierarchy to create an extended view of relatedness between synsets.

4.3 Applying Extended Lesk to Persian WSD

In order to compare the results of Direct and Cross Lingual approaches, the output from the cross lingual phase is used as an input to the knowledge based (direct) phase. Each tagged word from the input is considered as a target word to receive the second sense tag based on the extended Lesk algorithm. We adopted the method described in [18] to perform WSD for the Persian language. Persian glosses were collected using the semantic relations implemented for FarsNet. STeP-1 [19] was used for tokenizing glosses and stemming the content words.

5 Evaluation

5.1 Cross-Lingual Approach

The results of this method have been evaluated on comparable English and Persian Wikipedia pages. Seven human experts were involved in the evaluation process; they evaluated each tagged word as “the best sense assigned”, “almost accurate” and “wrong sense assigned”. The second option considers cases in which the assigned sense is not the best available sense for a word in a particular context, but it is very close to the correct meaning (not a wrong sense) which is influenced by the evaluation metric proposed by Resnik and Yarowsky in [20]. Evaluation results indicate an error rate of 25% for these pages. Table 1 summarizes these results. Our results indicate that the domain-specific words which usually occur frequently in both English and Persian texts are highly probable to receive the correct sense tag.

Due to the relatively smaller size of Persian texts, this system suffers from a low recall of 35%. However, as Wikipedia covers more and more Persian pages every day, soon we will be able to overcome this bottleneck.

According to the evaluation results, our Cross Lingual method gained an F-score of 0.48 which is comparable to 0.54 F-score of SenseRelate using Extended Lesk [17]. This indicates the performance of our approach can reach the F-score of the utilized English tagger. Employing a more accurate English sense tagger thus improves the WSD results for Persian words by far.

This system can be further evaluated by comparing its output to the results of assigning either random senses or the first sense to words. Since the senses

Table 1. Evaluation Results

	Cross Lingual		Direct		Baseline	
	Precision	F-Score	Precision	F-Score	Precision	F-Score
Best Sense	68%		51%		39%	
Almost Accurate	7%	0.48	9%	0.44	8%	0.40
Wrong Sense	25%		40%		53%	

in FarsNet are not sorted based on their frequency of usage (as compared to WordNet), we decided to use the first sense appearing in FarsNet (for each POS). Assigning the first sense to all tagged Persian words, the performance decreased significantly in terms of accuracy. The results in Table 1 indicate that, applying our novel approach results in a 28% improvement in accuracy in comparison with this selected baseline. However, assigning the most frequent sense to Persian words would be a more realistic baseline which yields a better estimation for our system's performance. Thus by the time the frequency of usage is provided for FarsNet senses, we anticipate that this problem will be minimized.

5.2 Direct Knowledge-based approach

As mentioned, the output of the Cross-lingual method was tagged again using the Direct approach. Overall, 53% of the words received a different tag using the Direct approach. Table 1 indicates the evaluation results for this approach.

5.3 Comparison: Knowledge based vs. Cross Lingual

Both systems employ the Extended Lesk algorithm. While the Cross Lingual method applies Extended Lesk on the English side and transfers senses to Persian words, the Direct approach works with Persian text directly. In other words, the former considers the whole text as the context and assigns one sense per discourse and the latter considers surrounding words and assigns one sense per collocation. Furthermore, the Cross Lingual method exploits WordNet for extending the glosses which covers more words, senses and semantic relations than FarsNet which is employed by the Direct method.

The main advantage of the cross lingual method is that we can utilize any highly accurate English sense disambiguator for the first phase while the Persian side remains intact.

On the other hand, this approach assigns the same tag (the most common sense) to all occurrences of a word which sacrifices accuracy. Moreover, if there is no English text with the same context available for a Persian corpus, this method cannot be applied. However collecting comparable texts over the web is not difficult. Finally, when the bilingual texts are not the direct translation of one another the system coverage will be limited to *common* words in both English and Persian texts. So, Cross Lingual method mainly works well for domain words and not for all the words appearing in the Persian texts.

Although Persian WSD while working with Persian texts directly seems to be more promising the evaluation results indicate a better performance for the Cross Lingual system. The reasons for this observation have been investigated and are as follows:

1. *Lack of reliable NLP tools for the Persian language.* While STeP-1 has just been made available as a tokenizer and a stemmer, there is no POS tagger for Persian which complicated the disambiguation process.
2. *Lack of comprehensive linguistics resources for the Persian language.* FarsNet is a very valuable resource for the Persian language. However it is still at a preliminary stage of development and does not cover all words and senses in Persian. In terms of size it is significantly smaller (10000 synsets) than WordNet (more than 117000 synsets) and it covers roughly 9000 relations between both senses and synsets.
3. *More ambiguity for Farsi words.* Disambiguating a Farsi word is a big challenge. Due to the fact that the short vowels are not written in the Farsi prescription, one needs to consider all types of homographs including heteronyms and homonyms. Moreover, there is no POS tagger to disambiguate Farsi words which dramatically increases the ambiguity for many Farsi words.

6 Conclusion and Future Work

A large number of WSD systems for widespread languages such as English is available. However, to date no large scale and highly accurate WSD system has been built for the Farsi language due to the lack of labeled corpora and monolingual and bilingual knowledge resources.

In this paper we overcame this problem by taking advantage of English sense disambiguators, availability of articles in both languages in Wikipedia and the newly developed lexical ontology, FarsNet, in order to address WSD for Persian. The evaluation results of the Cross-lingual approach show a 28% improvement in accuracy in comparison with the first-sense baseline. The cross lingual approach performed better than the knowledge based approach which is directly applied to Persian sentences. However, one of the main reasons for this performance is that the lack of NLP tools and comprehensive knowledge resources for Persian introduces many challenges for systems investigating this language.

This paper in the first step examined a novel idea for cross-lingual WSD in terms of plausibility, feasibility and performance. The ultimate results of our approach demonstrate a comparable performance to the utilized English sense tagger. Therefore, in the next step we will replace SenseRelate with another English sense tagger with a higher F-score. Gaining higher accuracy and recall for the Persian WSD system we can exploit it as a part of a bootstrapping system to create the first sense tagged corpus to aid supervised WSD approaches for the Persian language. Finally, as the available tools and resources improve for the Persian language, the Direct approach can be employed to address WSD for Persian texts directly when there is no comparable English text is available.

References

1. Roberto Navigli. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41:10:1–10:69, February 2009.
2. George A. Miller. Wordnet: a lexical database for english. *Commun. ACM*, 38:39–41, November 1995.
3. Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. A statistical approach to sense disambiguation in machine translation. In *Proceedings of the workshop on Speech and Natural Language*, HLT '91, pages 146–151, Stroudsburg, PA, USA, 1991. Association for Computational Linguistics.
4. M. Diab and Ph. Resnik. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 255–262, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.
5. Mihltz M. and Pohl G. Exploiting parallel corpora for supervised word-sense disambiguation in english-hungarian machine translation. In *Proceedings of the 5th Conference on Language Resources and Evaluation*, pages 1294–1297, 2006.
6. Dan TufiŞ, Radu Ion, and Nancy Ide. Fine-grained word sense disambiguation based on parallel corpora, word alignment, word clustering and aligned wordnets. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.
7. Dan Tufiş and Svetla Koeva. Ontology-supported text classification based on cross-lingual word sense disambiguation. In *Proceedings of the 7th international workshop on Fuzzy Logic and Applications: Applications of Fuzzy Sets Theory*, WILF '07, pages 447–455, Berlin, Heidelberg, 2007. Springer-Verlag.
8. Y. Motazedi and M. Shamsfard. English to persian machine translation exploiting semantic word sense disambiguation. In *Computer Conference, 2009. CSICC 2009. 14th International CSI*, pages 253 –258, 2009.
9. M. Shamsfard, A. Hesabi, H. Fadaei, N. Mansoory, A. Famian, S. Bagherbeigi, E. Fekri, M. Monshizadeh, and S. M. Assi. Semi automatic development of farsnet; the persian wordnet. In *In Proceedings of 5th Global WordNet Conference*, 2010.
10. H. Faili. An experiment of word sense disambiguation in a machine translation system. In *Natural Language Processing and Knowledge Engineering, 2008. NLP-KE '08. International Conference on*, pages 1 –7, 2008.
11. M. Lesk. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, SIGDOC '86, pages 24–26, New York, NY, USA, 1986. ACM.
12. Ch. Saedi, M. Shamsfard, and Y. Motazedi. Automatic translation between english and persian texts. In *In Proceedings of the 3rd Workshop on Computational Approaches to Arabic-script based Languages*, 2009.
13. T. Mosavi Miangah and Ali. Delavar Khalafi. Word sense disambiguation using target language corpus in a machine translation system. *Literary and Linguistic Computing*, 20(2):237–249, June 2005.

14. M. Soltani and H. Faili. A statistical approach on persian word sense disambiguation. In *Informatics and Systems (INFOS), 2010 The 7th International Conference on*, pages 1 –6, 2010.
15. Mosavi Miangah T. Solving the polysemy problem of persian words using mutual information statistics. In *Proceedings of the Corpus Linguistics Conference(CL2007)*, 2007.
16. R. Makki and M. Homayounpour. Word sense disambiguation of farsi homographs using thesaurus and corpus. In Bengt Nordstrm and Aarne Ranta, editors, *Advances in Natural Language Processing*, volume 5221 of *Lecture Notes in Computer Science*, pages 315–323. Springer Berlin / Heidelberg, 2008.
17. Ted Pedersen and Varada Kolhatkar. Wordnet::senserelate::allwords: a broad coverage word sense tagger that maximizes semantic relatedness. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Demonstration Session*, NAACL-Demonstrations '09, pages 17–20, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.
18. Satanjeev Banerjee. Extended gloss overlaps as a measure of semantic relatedness. In *In Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pages 805–810, 2003.
19. M. Shamsfard, H. Sadat Jafari, and M. Ilbeygi. Step-1: A set of fundamental tools for persian text processing. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA).
20. Ph. Resnik and D. Yarowsky. Distinguishing systems and distinguishing senses: new evaluation methods for word sense disambiguation. *Nat. Lang. Eng.*, 5:113–133, June 1999.