Towards Automatic Acquisition of a Fully Sense Tagged Corpus for Persian

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

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

Abstract. Sense tagged corpora play a crucial role in Natural Language Processing, particularly in Word Sense Disambiguation and Natural Language Understanding. Since semantic annotations are usually performed by humans, such corpora are limited to a handful of tagged texts and are not available for many languages with scarce resources including Persian. The shortage of efficient, reliable linguistic resources and fundamental text processing modules for Persian have been a challenge for researchers investigating this language. We employ a newly-proposed cross-lingual sense disambiguation algorithm to automatically create large sense tagged corpora. The initial evaluation of the tagged corpus indicates promising results.

1 Introduction

Word Sense Disambiguation (WSD) is the task of selecting the most appropriate meaning for a polysemous word, based on the context in which it occurs. Recent advancements in corpus linguistics technologies and the greater availability of more and more textual data encourage researchers to employ comparable and parallel corpora to address various NLP tasks.

To exploit supervised WSD approaches for applications as Machine Translation (MT) and Information Retrieval (IR), a large amount of sense-tagged examples for each sense of a word is needed. Devising an automatic method to generate such corpora thus will be of great benefit for languages with scarce resources such as Persian.

Recently we proposed a novel cross-lingual WSD approach that takes advantage of available sense disambiguation systems and linguistic resources for English to identify the word sense in a Persian document based on a comparable English document of the same topic [1]. The method was evaluated on comparable corpora that consist of a set of pairwise articles of the same topic in English and Persian. The result was promising [1].

In this paper, we aim at creating sense-tagged corpora to aid supervised and semi-supervised WSD systems. For such a purpose, we apply our newly-proposed WSD method to a parallel corpus, which contains sentence-level translations between English and Persian. To improve performance, we also extend the cross-lingual WSD approach by adding a direct sense tagging phase and enhancing the sense transfer stage of the cross-lingual method. We evaluate the accuracy of our improved approach and report the results.

2 Related Work

The knowledge acquisition bottleneck is pervasive across approaches to WSD. The availability of large-scale sense tagged corpora is crucial for many NLP systems. There are two branches of efforts to overcome this bottleneck. Some aim at creating manually sense tagged corpora. Tagging is performed by lexicographers. Consequently, it is expensive, limiting the size of such corpora to a handful of tagged texts. To lower the cost and increase the coverage of the tagged corpus, some developers created manually tagged corpora (e.g. Open Mind Word Expert [2]) by distributing the annotation workload among millions of web users as potential human annotators. While most manually sense tagged corpora are developed for English [3], they are not limited to this language only [4].

Automatic creation of sense tagged corpora seeks to minimize the knowledge acquisition bottleneck inherent to supervised approaches. In [5] they acquire example sentences for senses of words automatically based on the information provided in WordNet and information gathered from the Internet using existing search engines. [6] uses an aligned English-French corpus. For each English word, the classification of contexts is done based on the different translations in French for the different word senses. A problem is that different senses of polysemous words often translate to the same word in French. For such words it is impossible to acquire examples with this method [5]. [7] uses a word-aligned English-Spanish parallel corpus, and independently applies WSD heuristics for each of the languages to obtain ranked lists of senses for each word and picks the best sense for the word based on the overlaps of these lists. [8] uses a word aligned English-Italian corpus obtained from the MultiSemCor[1] and the Italian component of MultiWordNet[2] which is aligned with WordNet to automatically acquire sense tagged data, exploiting the polisemic differential between two languages.

For Persian, there is no publicly available sense-tagged corpus to use. There have been different attempts to apply supervised approaches to WSD for which a set of manually tagged words were prepared [9], [10]. However, some researchers are working to provide linguistic resources and processing units for Persian. FarsNet 1.0 [11] is a lexical ontology that relates synsets in each POS category by the set of WordNet 2.1 relations and connects Farsi synsets to English ones (in WordNet 3.0) using inter-lingual relations.

Our approach is unique in the sense that there has been no attempt to create a sense tagged corpus using an automatic or semi-automatic approach for the Persian language. Second, thanks to the availability of FarsNet, as opposed to many cross lingual approaches, we tag 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 resulted corpus can be utilized for many monolingual NLP tasks such as IR, Text Classification as well as bilingual ones including MT and Cross-Lingual tasks. In comparison with most automatic approaches which use a bilingual parallel corpus to generate

[1] http://multisemcor.itc.it
[2] http://multiwordnet.itc.it

sense tagged corpora for a target corpus, we do not sense tag both languages independently, nor do we use translation correspondences to distinguish senses. Instead, taking advantage of available mappings between synsets in WordNet and FarsNet, we utilize an existing source language (English) sense tagger which uses WordNet as a sense inventory to sense tag the target language (Persian) words. Finally, in order to improve the recall of our system, we employ a direct sense tagging method called Extended Lesk which has never been exploited to address WSD for Persian texts.

3 Creating the Sense Tagged Corpus

A direct strategy for creating a sense tagged corpus for WSD is to use parallel corpora to identify correspondences between word pairs. We employ the cross-lingual word sense tagging method described in [1] which has a high accuracy, but a relatively low recall, to tag Persian words using corresponding English tagged words in the utilized parallel corpus. We then apply a direct knowledge based algorithm to sense tag the remaining words. We replaced the comparable corpus used in [1] with a parallel corpus. Since Persian sentences in this corpus are a direct translation of the English ones in addition to improvements we made to both English tagging and the sense transfer phases, we gain better accuracy and coverage for the tagging results.

Currently available Persian-English parallel corpora are Miangah's corpus [12][3] consisting of 4,860,000 words and Tehran (TEP) corpus [13] composed of 612,086 bilingual sentences extracted from movie subtitles. TEP is a larger corpus and freely available, but the sentences are short and informal. Miangah's is smaller in size and is not available for free, but the quality of data leads to more apropos results. The texts in the corpus include a variety of text types from different categories such as art, culture, literature and science.

Several steps of preprocessing were carried out. On the English side, tokenization, lemmatization and POS tagging were performed by the English tagger. At the Farsi side, however, we used STeP-1 [14] to perform tokenization and stemming. The other challenge with Persian text processing is that there can be identical characters with different encodings observed in different resources. These are unified during this step.

We exploited a cross lingual approach [1] to tag the word senses in Persian texts. We also applied a knowledge based method directly to the Persian sentences to improve the recall. A brief description of these two methods follows.

Cross Lingual Phase: Persian WSD using Tagged English Words This phase consists of two separate stages. First, we use an English WSD system to assign sense tags to English words. Next, we transfer these senses to corresponding Persian words. Since, by design, these two stages are distinct, different

[3] Available via European Language Resource Association (ELRA)

English WSD systems can be employed in the first stage. There are different factors affecting the performance of our system.

First the more accurate the English tagger is, the more accurate the Persian sense tags will be. Supervised systems proved to offer the highest accuracy for WSD. There are many supervised WSD systems developed for English. However, as supervised systems usually perform sense disambiguation for a small set of words, using such a system limits the coverage of our method. Therefore, currently, we utilized the unsupervised application SenseRelate [15] for the English WSD stage which performs all word sense tagging using WordNet. We selected the Extended Lesk algorithm [16] which leads to the most accurate disambiguation [15]. We evaluated and corrected the wrong tags assigned by SenseRelate in order to investigate the reliability of our cross lingual approach for assigning sense tags to Persian words assuming we have a perfectly sense tagged English side. SenseRelate tags all ambiguous words in the input English sentences. Each of these sense labels corresponds to a synset in WordNet containing that word in a particular sense. We transfer these synsets from English to Persian using interlingual relations provided by FarsNet and match each WordNet synset assigned to a word in an English sentence to its corresponding synset in FarsNet.

Second, we need to match Farsi words with their counterparts on the English side. When it is possible to apply an accurate word alignment method to the language pair under examination, the creation of the sense tagged corpus from parallel corpora can be simple. However, word alignment methods hardly present a satisfactory performance, especially in corpora of real translations, where correspondences are often not one to one [17]. Therefore, we do not employ word alignment methods, since they may convey serious errors to the tagged corpus. Instead, for each matched synset in FarsNet which contains a set of Persian synonym words, we find all these words and assign the same sense as the English label to its translations in the aligned Persian sentence.

Initial evaluation indicated some words cannot be matched at the Farsi side because Farsi synsets usually do not provide full lists of synonyms. Therefore we extended the synonym set for each Persian word, using an available English-Persian dictionary, such that, for each tagged English word from an English sentence, we find all Persian translations and add them to the Farsi synset. Although these words can convey different senses of the English word, we adjust it by giving higher priority to words which are provided by the FarsNet synset. Moreover, according to the one sense per discourse heuristic [6], it is not probable to observe same Farsi words with different senses in one sentence.

Direct Phase: Applying Extended Lesk for Persian WSD To increase the number of tagged words in our corpus, we applied a direct WSD algorithm to Persian sentences. Thanks to the availability of FarsNet, the Extended Lesk method is applicable to Persian texts as well. 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 [1]. Therefore,

we considered only the tags with a score higher than a predefined confidence threshold. This results in gaining a higher recall while the tags remain accurate.

4 Evaluation

The tagged corpus was evaluated on 480 words which were randomly selected from various domains such as Politics, Science, Culture, Art and had an average sense count of 2.17. Seven human experts were involved in the evaluation process. In the first step, the output from SenseRelate was revised manually and the wrong tags assigned were corrected. This led to fully accurate sense tagged English sentences. After these tags were transferred and assigned to Persian words on corresponding Persian sentences, the human experts 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 [18]. Evaluation results indicate an error rate of 9% for the selected Farsi words. Table 1 summarizes these results. Studying the output results revealed the content words describing the main concept of each sentence are highly probable to receive the correct sense tag.

This system demonstrates a good accuracy of 91%, but a relatively low recall of 46%. Note that the original English tagger has an average recall of 57%. This will act as an upper bound for our system's recall. The reason for a lower recall than the English tagger is that FarsNet 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. Another problem is tagging verbs in Persian sentences. Since verbs appear in their infinitive format in FarsNet while they are inflected in a particular tense and person, a better morphological analysis of Persian verbs is required to increase the number of matches. Moreover, structural differences between the English and Persian languages usually lead to observing single English words translating to Persian phrases or compound words. Since FarsNet does not contain all these words collocations, we might tag some part of a compound word and leave the rest untagged. Since our main goal is developing a cross-lingual, yet language independent, approach to create sense tagged corpora, we have not designed Persian-specific solutions to improve the recall at this time. Having an "ideal" aligned WordNet (a lexical resource such that all the sense distinctions in one language are reflected in the other, and all words and phrases are included) would minimize this issue.

Since the senses in FarsNet are not sorted based on their frequency of usage (as opposed to WordNet), we assigned the first sense appearing in FarsNet (for each POS) to words to create a baseline system. According to the results indicated in Table 1, applying our novel approach results in a 11% improvement in

Table 1: Evaluation Results

	Cross Lingual			Cross Lingual + Direct			Baseline		
	P	R	F-Score	P	R	F-Score	P	R	F-Score
Best Sense	80%			76%			45%		
Almost Accurate	11%	0.46	0.60	8%	0.57	0.67	11%	0.46	0.49
Wrong Sense	9%			16%			44%		

the F-score[4] in comparison with this selected baseline. However, assigning the most frequent sense to Persian words would be a more realistic baseline which we plan to employ once it is made available for FarsNet.

The untagged words remaining from Cross-lingual phase were sense tagged using the Direct approach. Since the final tagged corpus should be highly accurate, we did not sacrifice accuracy to gain a higher recall. Therefore, we considered a minimum score of 8[5], and approved the tags with an associate score of equal to or higher than this threshold. This results in an improvement of 11% in recall at a cost of 6% in accuracy. Due to the small size of FarsNet and the relatively higher error rate of the Direct approach, an improvement in the recall resulted in a decrease in accuracy. Hence, exploiting the Cross Lingual approach without passing the results through the Direct phase will result in obtaining a more accurate tagged corpus while the recall remains about 11% lower.

5 Conclusions and Future Work

We proposed an automatic approach for creating fully sense-tagged corpora for the Persian language which has an error rate of 9%. Although the resulted corpus might be noisy, it is still much easier and less time consuming to check already tagged data than to start tagging from scratch.

Since the accuracy of the tags assigned to the English words will affect that of Persian sense tags, a more accurate English tagger can improve the final results of our system. We are planning to replace SenseRelate with a more accurate English tagger such as WSDGate framework[6] to minimize the manual correction of English tags. Moreover, we are investigating linguistic based solutions to improve the matching desired Persian words during the Transfer phase. Finally, improvements in Word Alignment techniques For the English – Persian language pair can be of great benefit to maximize the coverage of our system.

Acknowledgements. This research is partially supported by Natural Sciences

[4] F-Score is calculated as $2\frac{(1-ErrorRate)\cdot Recall}{1-ErrorRate+Recall}$, where $ErrorRate$ is the percentage of words that have been assigned the wrong sense.

[5] This threshold is set based on experiments favouring precision over recall.

[6] http://wsdgate.sourceforge.net/

and Engineering Research Council of Canada (NSERC). We would like to thank Prof. Shamsfard from the Natural Language Processing Research laboratory of Shahid Beheshti University (SBU) for providing us with the FarsNet 1.0 package.

References

1. B. Sarrafzadeh, N. Yakovets, N. Cercone, and A. An, "Cross lingual word sense disambiguation for languages with scarce resources," in *Proc. of The 24th Canadian Conference on Artificial Intelligence*, 2011.
2. T. Chklovski and R. Mihalcea, "Building a sense tagged corpus with open mind word expert," in *Proc. of the ACL-02 workshop on Word sense disambiguation: recent successes and future directions - Volume 8*, 2002.
3. G. A. Miller and et al., "A semantic concordance," in *Proc. of the workshop on Human Language Technology*, 1993.
4. S. Koeva, S. Lesseva, and M. Todorova, "Bulgarian sense tagged corpus," in *In Proc. of the Fifth International Conference Formal Approaches to South Slavic and Balkan Languages*, 2006.
5. R. Mihalcea and D. I. Moldovan, "An automatic method for generating sense tagged corpora," in *Proc. of the 16th national conference on Artificial intelligence and the 11th Innovative applications of artificial intelligence conference*, 1999.
6. W. A. Gale, K. W. Church, and D. Yarowsky, "One sense per discourse," in *Proc. of the workshop on Speech and Natural Language*, 1992.
7. G. de Melo and G. Weikum, "Extracting sense-disambiguated example sentences from parallel corpora," in *Proc. of the 1st WDE*, 2009.
8. A. M. Gliozzo and M. Ranieri, "Crossing parallel corpora and multilingual lexical databases for wsd," in *Proc. of the 6th International Conference on Intelligent Text Processing and Computational Linguistics*, 2005.
9. R. Makki and M. Homayounpour, "Word sense disambiguation of farsi homographs using thesaurus and corpus," in *Proc. of the 6th international conference on Advances in Natural Language Processing*, 2008.
10. M. Soltani and H. Faili, "A statistical approach on persian word sense disambiguation," in *The 7th International Conference on INFOS*, 2010.
11. M. e. a. Shamsfard, "Semi automatic development of farsnet; the persian wordnet," in *Proc. of 5th Global WordNet Conference*, 2010.
12. T. Miangah, "Constructing a large-scale english-persian parallel corpus," in *Meta: Translators' Journal*, 2009.
13. T. Pilevar and H. Faili, "Persiansmt: A first attempt to english-persian statistical machine translation," in *JADT*, 2010.
14. e. a. M. Shamsfard, "Step-1: Standard text preparation for persian language," in *Proc. of Machine Translation Summit XII*, 2009.
15. T. Pedersen and V. Kolhatkar, "Wordnet::senserelate::allwords: a broad coverage word sense tagger that maximizes semantic relatedness," in *Proc. of Human Language Technologies: NAACL, Companion Volume: Demonstration Session*, 2009.
16. S. Banerjee, "Extended gloss overlaps as a measure of semantic relatedness," in *Proc. of the 18th International Joint Conference on Artificial Intelligence*, 2003.
17. L. Specia and et al., "An automatic approach to create a sense tagged corpus for word sense disambiguation in machine translation," in *Proc. of the 2nd Meaning Workshop*, 2005.
18. P. Resnik and D. Yarowsky, "Distinguishing systems and distinguishing senses: new evaluation methods for word sense disambiguation," *Nat. Lang. Eng.*, 1999.