Standard Test Collection for English-Persian Cross-Lingual Word Sense Disambiguation

Navid Rekabsaz, Serwah Sabetghadam, Mihai Lupu, Linda Andersson, Allan Hanbury
Vienna University of Technology
Favoritenstrasse 9 1040 Vienna Austria
surname@ifs.tuwien.ac.at

Abstract

In this paper, we address the shortage of evaluation benchmarks on Persian (Farsi) language by creating and making available a new benchmark for English to Persian Cross Lingual Word Sense Disambiguation (CL-WSD). In creating the benchmark, we follow the format of the SemEval 2013 CL-WSD task, such that the introduced tools of the task can also be applied on the benchmark. In fact, the new benchmark extends the SemEval-2013 CL-WSD task to Persian language.

Keywords: Persian, Farsi, Cross Lingual Word Sense Disambiguation, test collection

1. Introduction

Word Sense Disambiguation (WSD)—the task of automatically selecting the most related sense for a word occurring in a context—is considered as a main step in the course of approaching language understanding beyond the surface of the words. WSD has been intensively studied in Natural Language Processing (NLP) (Navigli, 2012), Machine Translation (Chan et al., 2007; Costa-Jussà and Farrús, 2014), and Information Retrieval (Zhong and Ng, 2012).

As a paradigm, Cross-lingual Word Sense Disambiguation (CL-WSD) focuses on lexical substitution in a target language such that it targets disambiguation of one word in a source language while translating to a target language. SemEval-2010 (Lefever and Hoste, 2010) and SemEval-2013 (Lefever and Hoste, 2013) provide an evaluation platform for term disambiguation from English to Dutch, German, Italian, Spanish, and French.

In recent years, several tools and libraries for Persian language have been developed. For example, Seraji et al. (2012) provides a set of tools for preprocessing (PrePer), sentence segmentation and tokenization (SeTPer), and POS tagging (TagPer). More recently, Samvelian et al. (2014) introduces PersPred, focusing on processing of compounding verbs. Finally, Feely et al. (2014) provides a front-end and new tools for language processing. While these tools mostly target the complexities of the language, there is no standard evaluation framework for Persian language in the WSD and CL-WSD domain.

In this paper, we address this shortage by creating and making available a new benchmark for English to Persian CL-WSD. In creating the benchmark, we follow the format of the SemEval 2013 CL-WSD task, such that the introduced tools of the task can also be applied on the benchmark. In fact, the new benchmark extends the SemEval-2013 CL-WSD task to Persian language.

In the following, we review the related work and resources in Persian language, followed by the description of the collection in Section 3..

2. Resources in Persian Language

Persian is a member of the Indo-European language family, and uses Arabic letters for writing. Seraji et al. (2012) provide a comprehensive overview on the main characteristics of the language. For instance, the diacritic signs are not written—it is expected that the reader can read the text in the absence of the short vowels. This characteristic causes a special kind of word ambiguity in writing, such that some words are pronounced differently while their written forms are the same e.g. كشتي means both "wrestling" and "ship".

Methods for approaching WSD and CL-WSD highly rely on knowledge and data resources in the language. In the following, we briefly review the main Persian language resources for addressing CL-WSD challenges.

The main knowledge resource in Persian is *FarsNet* (Shamsfard et al., 2010)—the Persian WordNet. Its structure is comparable to WordNet and goes by the same principles while containing significantly fewer words ($\sim$13K versus $\sim$147K). Also, most of its synsets are mapped to synsets in WordNet using equal or near-equal relations. Knowledge-based systems are limited and they are only at high cost extendable. Exploiting parallel corpora can be another effective method for CL-WSD. Existing parallel corpora (English-Persian) are as follows:

- Tehran English-Persian Parallel (TEP) (Pilevar et al., 2011): a free collection extracted from 1600 movie subtitles.
- Parallel English-Persian News (PEN) (Farajian, 2011): the collection aligns 30K sentences of news corpora but is not yet available.
- The collection provided by European Language Resource Association (ELRA) which is a commercial collection with approximately 100K aligned sentences.

In the absence of reliable and comprehensive resources, some CL-WSD methods exploit the use of monolingual corpora. The main available text collections in Persian are

Hamshahri (AleAhmad et al., 2009), *dotIR*[1], *Bigjekhan*[2], and Uppsala Persian Corpus (UPEC) (Seraji et al., 2012).
In terms of work on WSD and CL-WSD, Saedi et al. (2009) exploits the use of WSD in their English-Persian machine translator by first sense disambiguation in the source language and then translating it to the target language. For English-to-Persian translation, they use WordNet in combination with Lesk algorithm (Lesk, 1986), while for Persian-to-English, they consider the probability of the co-occurrence of the common senses in a context, learned from a monolingual corpus. More recently, Sarrafzadeh et al. (2011) follow a knowledge-based approach by exploiting FarsNet together with leveraging English sense disambiguation. Their model consists of three phases of: English sense disambiguation, utilizing WordNet and FarsNet to transfer the sense, and selecting the sense from FarsNet. As another method, they investigate direct WSD by applying extended Lesk algorithm for Persian WSD. They count the number of shared words between two glosses, the gloss of each sense of the target word with the gloss of other words in the phrase. The one with larger number of common words is chosen. They test on parallel page of Wikipedia in English and Persian evaluated by experts. Finally, they show that the first approach works better since they can use the state of the art disambiguater of English language and direct approach suffers from lack of NLP tools and ambiguity of Farsi words.
However, the mentioned methods evaluate their methods only internally and therefore the results are impossible to be compared with other possible approaches. In this work, we address this shortage by creating a new CL-WSD benchmark for Persian, based on the SemEval 2013 CL-WSD task.

3. Persian CL-WSD Evaluation Benchmark

In this section, we describe in detail the process of creating the CL-WSD evaluation benchmark from English to Persian. The created test collection completely matches the output format of the SemEval 2013 CL-WSD task (Lefever and Hoste, 2013) and adds a new language to this multilingual task. In addition, we tightly follow the methods in this task for the creation of the gold standard, with only minor alterations necessary in view of the available Persian language resources.

3.1. SemEval 2013 CL-WSD

The SemEval 2013 Cross-lingual Word Sense Disambiguation task aims to evaluate the viability of multilingual WSD on a benchmark lexical sample data set. Participants should provide contextually correct translations of English ambiguous nouns into five target languages: German, Dutch, French, Italian, and Spanish. The task contains a test set of 20 nouns, each with 50 sentences.
In order to create the golden standard as described in Lefever et al. (2014), in the first step a sense inventory was constructed based on the possible translations of the ambiguous terms. In order to find the target translations, Lefever et al. (2014) ran word alignment on aligned sentences of the Europarl Corpus (Koehn, 2005) and manually verified the results. In the next step, the resulting translations were clustered by meaning per focus words. Finally, annotators used this clustered sense inventory to select the correct translation for each word, for up to three translations per word.
Two different evaluation methods are used for the task: 1) *Best Result* evaluation, in which the system suggests any number of translations for each target word, and the final score is divided by the number of these translations. 2) *Out-of-five* (OOF) or more relaxed evaluation, in which the system provides up to five different translations, and the final score is that of the best of these five (more details in Lefever et al. (2013)).

3.2. New Persian Collection

Similar to Lefever et al. (2014), the creation of CL-WSD task for Persian consists of two parts: 1) Creating the sense inventory and 2) Annotation of the translations (i.e. the ground truth).
To create the sense inventory for the 20 nouns, due to the lack of a representative parallel corpora, we leveraged three main dictionaries of the Persian language—Aryanpour, Moein, and Farsidic.com—to obtain as large a coverage as possible for their translations. The translations themselves were added by a Persian linguist.
In order to provide a thorough set of translations, in addition to different meanings of nouns, their idiomatic meanings (in combinations) are also considered. For example, for the word "pot", a wide variety of direct translation (گلدان "vase", دیگ "container for cooking", قوري "container for holding drink", کتري "kettle") were selected. However, there is an expression like "melting pot", which is not in the dictionaries. These idiomatic meanings were added to the senses of this word as an expression or equal idiom, which in this case is جامعهي چند نژادي "multicultural (multiracial) society".
The linguist also divided the translations into different senses. The resulting clusters for nouns range from 2 (e.g., "education" to تحصیل "studying" or معرفت "wisdom") to 6 clusters (e.g., "post" to پست "mail", مقام "position", محل ماموریت "place where someone is on duty", تیر عمودي "vertical support/marker", اعلام کردن "announcing", معني اصطلاحي "idiom"). The number of translations ranged from 13 for the word "mood" to 42 for the word "ring"[3]. Table 1 shows details of these statistics.
In a second phase, this generated sense inventory is used to annotate the sentences in the test set (50 sentences for each ambiguous word). This phase was performed by three Persian native-speakers. Via a web-based application[4], annotators chose the appropriate translations in two steps: In the

[1] http://ece.ut.ac.ir/DBRG/webir/index.html
[2] http://ece.ut.ac.ir/dbrg/Bijankhan
[3] https://github.com/navid-rekabsaz/wsd_persian/tree/master/resources/sense-inventory
[4] https://github.com/navid-rekabsaz/wsd_persian/tree/master/software/webapplication

Table 1: Overview of annotators consensus for Persian language

word	# of clus.	# of trans.	avg. # clus./sent.	% clus. consensus
coach	4	18	1.00	98
education	2	15	1.02	98
execution	3	14	1.08	92
figure	5	33	1.08	92
job	3	21	1.02	98
letter	4	29	1.04	96
match	3	19	1.04	96
mission	3	19	1.02	98
mood	2	13	1.00	100
paper	3	32	1.02	98
post	6	38	1.00	100
pot	4	34	1.04	96
range	5	36	1.04	96
rest	4	40	1.00	100
ring	6	42	1.04	98
scene	4	25	1.02	98
side	3	32	1.00	96
soil	3	18	1.00	100
strain	4	39	1.02	98
test	2	13	1.00	100

first step, they chose the related sense (cluster). In the second step, the system showed the related translations for the sense, of which they chose up to three translations. In case of no related translation, they chose nothing and continued to the next question. The agreement between annotators is shown in Table 1 and is similar to that observed by Lefever et al. (2014).

Using the annotated data, we created the gold standard[5] in the same format as Lefever et al. (2013), such that all the evaluation scripts used in the SemEval 2013 CL-WSD task can also be used on this data. Example 1 and Example 2 show the ground truth for two sentences with the word "coach":

(1) SENTENCE 2: A branch line train took us to where a *coach* picked us up for the journey up to the camp.
coach.n.fa 2 :: اتوبوس 3; اتومبيل 2; واگن 1;

(2) SENTENCE 16: Agassi's *coach* came to me with the rackets.
coach.n.fa 16 :: مربي 3; مربي ورزش 2; سرمربي 2; معلم خصوصي 1;

4. Conclusion

In this paper, we create and provide a benchmark for English to Persian CL-WSD. The collection targets the ambiguity of 20 English words, each with 50 sentences. As the collection follows the format of the SemEval 2013 CL-WSD task, it in fact extends the task for Persian language.

[5] https://github.com/navid-rekabsaz/wsd_persian/tree/master/resources/golden/Persian

5. Acknowledge

This work is partly funded by two FWF projects MUCKE (I 1094-N23) and ADMIRE (P 25905-N23).

6. Bibliographical References

AleAhmad, A., Amiri, H., Darrudi, E., Rahgozar, M., and Oroumchian, F. (2009). Hamshahri: A standard persian text collection. *Knowledge-Based Systems*, 22(5):382–387.

Chan, Y. S., Ng, H. T., and Chiang, D. (2007). Word sense disambiguation improves statistical machine translation. In *Annual Meeting-Association for Computational Linguistics*, volume 45, page 33. Citeseer.

Costa-Jussà, M. R. and Farrús, M. (2014). Statistical machine translation enhancements through linguistic levels: A survey. *ACM Computing Surveys (CSUR)*, 46(3):42.

Farajian, M. A. (2011). Pen: parallel english-persian news corpus. In *Proceedings of the 2011th World Congress in Computer Science, Computer Engineering and Applied Computing*.

Feely, W., Manshadi, M., Frederking, R., and Levin, L. (2014). The cmu metal farsi nlp approach. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC14)*, pages 4052–4055.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT Summit*.

Lefever, E. and Hoste, V. (2010). Semeval-2010 task 3: Cross-lingual word sense disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 15–20. Association for Computational Linguistics.

Lefever, E. and Hoste, V. (2013). Semeval-2013 task 10: cross-lingual word sense disambiguation. In *7th International workshop on Semantic Evaluation (SemEval 2013)*, pages 158–166. Association for Computational Linguistics (ACL).

Lefever, E. and Hoste, V. (2014). Parallel corpora make sense: Bypassing the knowledge acquisition bottleneck for word sense disambiguation. *International Journal of Corpus Linguistics*, pages 333–367.

Lesk, M. (1986). 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*, pages 24–26. ACM.

Navigli, R. (2012). A quick tour of word sense disambiguation, induction and related approaches. In *SOFSEM 2012: Theory and practice of computer science*, pages 115–129. Springer.

Pilevar, M. T., Faili, H., and Pilevar, A. H. (2011). Tep: Tehran english-persian parallel corpus. In *Computational Linguistics and Intelligent Text Processing*, pages 68–79. Springer.

Saedi, C., Motazadi, Y., and Shamsfard, M. (2009). Automatic translation between english and persian texts. In *Proceedings of the Third Workshop on Computational Approaches to Arabic Script-based Languages, Ottawa, Ontario, Canada*.

Samvelian, P., Faghiri, P., and El Ayari, S. (2014). Extending the coverage of a mwe database for persian cps exploiting valency alternations. In *Language Resources and Evaluation Conference (LREC)*, pages 4023–4026.

Sarrafzadeh, B., Yakovets, N., Cercone, N., and An, A. (2011). Cross-lingual word sense disambiguation for languages with scarce resources. In *Advances in Artificial Intelligence: Proceedings of 24th Canadian Conference on Artificial Intelligence*, pages 347–358. Springer Berlin Heidelberg.

Seraji, M., Megyesi, B., and Nivre, J. (2012). A basic language resource kit for persian. In *Eight International Conference on Language Resources and Evaluation (LREC 2012), 23-25 May 2012, Istanbul, Turkey*, pages 2245–2252. European Language Resources Association.

Shamsfard, M., Hesabi, A., Fadaei, H., Mansoory, N., Famian, A., Bagherbeigi, S., Fekri, E., Monshizadeh, M., and Assi, S. M. (2010). Semi automatic development of farsnet; the persian wordnet. In *Proceedings of 5th Global WordNet Conference, Mumbai, India*.

Zhong, Z. and Ng, H. T. (2012). Word sense disambiguation improves information retrieval. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 273–282. Association for Computational Linguistics.