

# Addressing Cross-Lingual Word Sense Disambiguation on Low-Density Languages: Application to Persian

Navid Rekabsaz<sup>1</sup>, Mihai Lupu<sup>1</sup>, Allan Hanbury<sup>1</sup>, and Andres Duque<sup>2</sup>

<sup>1</sup> Institute of Software Technology and Interactive Systems  
TU WIEN

A-1040 Vienna, Austria  
[last\_name]@ifs.tuwien.ac.at

<sup>2</sup> Universidad Nacional de Educacin a Distancia (UNED)  
Madrid 28040, Spain  
adunque@lsi.uned.es

**Abstract.** We explore the use of unsupervised methods in Cross-Lingual Word Sense Disambiguation (CL-WSD) with the application of English to Persian. Our proposed approach targets the languages with scarce resources (low-density) by exploiting word embedding and semantic similarity of the words in context. We evaluate the approach on a recent evaluation benchmark and compare it with the state-of-the-art unsupervised system (CO-Graph). The results show that our approach outperforms both the standard baseline and the CO-Graph system in both of the task evaluation metrics (*Out-Of-Five* and *Best result*).

**Keywords:** Word Sense Disambiguation, cross-lingual, semantics, Word2Vec

## 1 Introduction

Word Sense Disambiguation (WSD) is the task of automatically selecting the most related sense for a word occurring in a context. It is considered as a main step in the course of approaching language understanding beyond the surface of the words.

Typically, WSD methods are classified into knowledge-based, supervised, and unsupervised. Knowledge-based approaches use available structured knowledge. Supervised approaches learn a computational model based on large amounts of annotated data. While these two approaches show competitive results in practice, they both have to face the knowledge acquisition bottleneck. This is a particular problem in specific domains or low-density languages. As an alternative, unsupervised approaches address WSD using only information extracted from existing corpora, such as various word co-occurrence indicators.

As a paradigm, multilingual and cross-lingual WSD methods focus on lexical substitution in a target language. Cross-lingual Word Sense Disambiguation (CL-WSD) targets disambiguation of one word in a source language while translating to a target language. SemEval-2010 [9] and SemEval-2013 [10] provide an evaluation platform for word disambiguation from English to Dutch, German, Italian, Spanish, and French. Recently, Rekabsaz et al. [14] added the Persian (Farsi) language to this set by following the CL-WSD SemEval format to create the test collection.

Many participating systems in the SemEval tasks exploit parallel corpora, mainly Europarl [8], to overcome the knowledge acquisition bottleneck [11, 15]. However, the approaches used in the tasks are not applicable for many languages and domains due to the scarcity of bilingual corpora. Persian, for instance, suffers from the lack of reliable

and comprehensive knowledge resources as well as parallel corpora. In such cases, unsupervised methods based on monolingual corpora (together with bilingual lexicon) are preferable, if not the only available option [18]. For example, Bungum et al. [2] find the probable translations of a context in the source language and identify the best translation using a language model of the target language. Duque et al. [4] build a co-occurrence graph in the target language, and test a variety of graph-based algorithms for identifying the best translation match.

In terms of combining WSD and word embedding, Chen et al. [3] use knowledge-based WSD to identify distinct representations for different senses of the same word. Our approach for CL-WSD is the opposite of this: starting from word embedding representations, it identifies the similarity of the potential translations to the words in its context and choose the one the highest semantic similarity to the context.

The contributions of the work are two-fold:

1. Providing a new state-of-the-art for unsupervised CL-WSD methods, based on the use of word embedding.
2. Evaluating the method on a recent benchmark, demonstrating its superior results in comparison to a state-of-the-art unsupervised approach as well as baselines.

In order to evaluate our approach, we used the collection of the Persian CL-WSD task [14], and compared our approach and the CO-Graph system [4], observing the advantages of using word embedding in WSD.

## 2 Resources in Persian Language

Based on our knowledge, the existing parallel corpora (English-Persian) are as follows: Tehran English-Persian Parallel (TEP) [13], Parallel English-Persian News (PEN) [6], and the European Language Resource Association (ELRA) corpus. Among the mentioned resources, TEP is the only freely available one, but it is only a small set of informal conversations in movie subtitles, and therefore it is not sufficient for a general representation of the language.

In the absence of reliable and comprehensive resources, our unsupervised CL-WSD method exploits the use of monolingual corpora. In this work, we use the *Hamshahri* [1] collection—a widely used monolingual resource for Persian.

In terms of related work addressing the CL-WSD problem in Persian, Sarrafzadeh et al. [16] follows a knowledge-based approach by exploiting *FarsNet* [17]—the Persian WordNet. However, since their evaluation collection is not available, the results are impossible to compare with other possible approaches.

## 3 Unsupervised CL-WSD Method

Our approach follows the main idea of the Lesk algorithm [12], namely that words in a given context tend to share a common topic. In the absence of external knowledge sources, we use word embedding to compute their semantic similarity. We measure the relatedness of each possible translation of the ambiguous word to all possible translations of the words in its context (the paragraph given by the task) and select the most similar one to the context.

To formulate our approach, let us define the list  $T$  of translation sets for the words in the context:  $T = \{T_1, T_2, \dots, T_n\}$  where  $n$  is the number of words in the context, and

$T_i$  is the set of translations for the  $i^{th}$  word in the context. For each possible translation  $t \in T_i$ , we also have  $P(t)$ —an indicator of how frequent this particular translation is.

In general, we compute the similarity of two translations  $t$  and  $\bar{t}$  as the cosine of their vectors, achieved from word embedding model. However, sometimes the translation  $t$  of one word in English may be two or more words in Persian, and thus we will have more than one vector. We thus define a general similarity between two translations:

$$Sim(t, \bar{t}) = \max_{w \in t, \bar{w} \in \bar{t}} (\cos(V_w, V_{\bar{w}})) \quad (1)$$

where  $V_w$  is the vector representation of the word translation  $w$ .

In what follows, in order to simplify the annotations, we will use the definition of similarity as  $Sim(t, \bar{t}) = \cos(V_t, V_{\bar{t}})$  rather than Eq. 1, thereby making the  $max$  function implicit.

Having a definition of similarity between two word translations, we now move to defining the similarity between a translation and a set of translations (i.e. the translation of the ambiguous word and the set of possible translations of the words in its context). In this work, we consider two ways to approach the problem: 1. generate one semantic vector for each possible translation of the context (by aggregating the vectors of the word translations that make up this context translation) and compare the translation candidate with the semantic vector of the context; 2. compare directly the vector of the translation candidate with all possible vectors of the words translations in its context. In both cases, the translation candidate with the highest score is chosen as the detected sense of the word.

We denote the first approach *RelAgg*. It generates the vector representation of the context using the  $contextVec(t, T)$  function defined in Algorithm 1, where the  $normalize(Vec)$  function normalizes the given vector using the Euclidean norm.

---

**Algorithm 1: contextVec Algorithm**


---

**Input:** translation candidate  $t$ , and the list of translation sets  $T$

**Output:** vector representation of the context

$sumVec \leftarrow []$ ;

**for**  $T_i \in T$  **do**

$maxVec \leftarrow []$ ;

$maxSim \leftarrow 0$ ;

**for**  $\bar{t} \in T_i$  **do**

$sim \leftarrow \cos(V_t, V_{\bar{t}})$ ;

**if**  $maxSim < sim$  **then**

$maxVec \leftarrow V_{\bar{t}}$ ;

$maxSim \leftarrow sim$ ;

$sumVec \leftarrow sumVec + maxVec$

**return**  $normalize(sumVec)$ ;

---

Given the vector representation of the context, the *RelAgg* approach is defined as the cosine function between the vectors representation of the translation candidate and the context. The final result of the approach is multiplied by the probability of the translation candidate:

$$RA(t, T) = \cos(V_t, contextVec(t, T))P(t) \quad (2)$$

Table 1: Results of F-measure on *Out-Of-Five (OOF)* and *Best Result (Best)* evaluations .

Eval.	Method	F-measure
OOF	RelAgg	<b>50.2</b>
	RelGreedy	49.3
	CO-Graph Dijkstra	44.1
	Standard Baseline	41.8
Best	RelAgg	<b>18.8</b>
	RelGreedy	18.3
	CO-Graph Dijkstra	17.4
	Standard Baseline	15.8

The second approach is denoted as *RelGreedy*:

$$RG(t, T) = \max_{T_i \in T} \left( \max_{\bar{t} \in T_i} (\cos(t, \bar{t})) \right) P(t) \quad (3)$$

In RelGreedy, among all the translations in the context, the value of the most similar one to the translation candidate is returned. Similar to RelAgg, the final score is multiplied by the probability of the translation candidate.

Finally, given the score of the relatedness of each translation to its context using either RelAgg or RelGreedy, we can select the best translation among the candidates:

$$Result = \arg \max_{t_i} (Rel^*(t_i, T)) \quad (4)$$

where  $t_i$  is a translation candidate for the word with ambiguity, and  $Rel^*$  is either RelAgg or RelGreedy.

## 4 Experiments and Results

*Data Preparation* As discussed, we selected the Hamshahri collection for the required monolingual corpora. Similar to Jadidinejad et al. [7], we use PerStem for stemming, together with TagPer as a state-of-the-art POS tagging tool. We created Word2Vec word embedding in the Skip-Gram model with sub-sampling at  $t = 10^{-4}$ , the context windows of 5 words, epochs of 25, words count threshold of 5, and dimension of 200.

Beside the monolingual corpus, a bilingual lexicon is required for our unsupervised CL-WSD approach. While using parallel corpora is considered as a more effective method for creating lexica [5], due to the lack of reliable parallel corpora, we have to use a simple English to Persian dictionary<sup>3</sup>. To have it in digital form, we used the online API of one of the existing translation services.

*Baselines* The first baseline—the *Standard* baseline—follows the method introduced in the SemEval 2013 CL-WSD task. Similar to the task, for the *Best Result* and *Out-Of-Five* evaluations, we selected the most common and the five most common translations respectively. Evaluating the baselines on the Persian CL-WSD task using F-measure (in the scale of 0 to 100) evaluation measure, we observed the value of 15.8 for the *Best Result* and 41.8 for the *Out-Of-Five* evaluation.

For the second baseline, we evaluate the Persian benchmark on the state-of-the-art unsupervised CL-WSD system called CO-Graph [4]. The CO-Graph system offers

<sup>3</sup> [https://github.com/navid-rekabsaz/wsd\\_persian/tree/master/resources/dictionary](https://github.com/navid-rekabsaz/wsd_persian/tree/master/resources/dictionary)

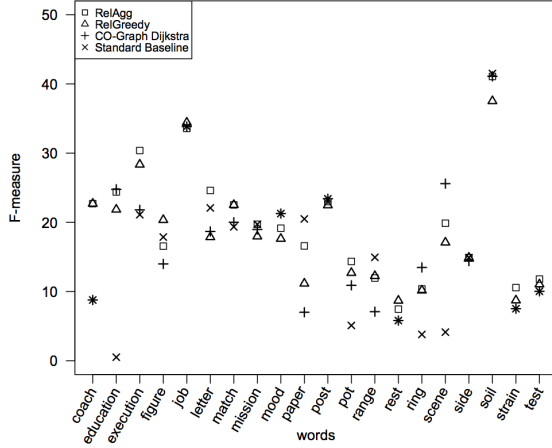


Fig. 1: Results of the *OOF Result* evaluation based on each word using F-measure

competitive results in the SemEval 2010 and SemEval 2013 CL-WSD tasks, for all the proposed languages. It outperforms all of the unsupervised participating systems using only monolingual corpora, and even most of the ones which use parallel corpora or knowledge resources (details in Duque et al. [4]). To evaluate the CO-Graph system on the Persian benchmark, we first created the graph using the articles of the Hamshahri collection, each as a document. In the construction of the graph, we only took into account the nouns by POS tagging. After evaluating various algorithms, we found the Dijkstra algorithm together with  $p\text{-value}=10^{-6}$  as the best performing approach with the F-measure metrics of 17.4 for the *Best Result* and 44.1 for the *Out-Of-Five* evaluation.

*Evaluation* We applied POS tagging on the sentences of the SemEval 2013 CL-WSD task and only selected the verbs and nouns as the context of the ambiguous words. We then lemmatized the words using WordNetLemmatizer of the NLTK toolkit and found their translations in the bilingual lexicon. Using the word embedding of the translated words, we calculated the relatedness score of each translation candidate to its context using RelAgg and RelGreedy (Section 3). The translation probability rate in our lexica was used as the  $P(t)$  value in Eq. 2 and Eq. 3. Table 1 shows the *Out-Of-Five* and *Best Result* evaluation results of RelAgg and RelGreedy relatedness approaches using the F-measure metrics.

The results for both the *Out-Of-Five* and *Best Result* evaluations show that our approach outperforms the standard and the CO-Graph baselines. Comparing the relatedness approaches, we observe similar results for the RelAgg and RelGreedy methods, while RelAgg has slightly better performance, specially in the *Best Result* evaluation.

We compared the effectiveness of our approach on different words with the standard and the CO-Graph baselines in Figure 1. The results show that while for most words our approach outperforms the standard baseline as well as the CO-Graph system, none of the systems could outperform the standard baseline for “mood” and “side”. Analyzing the evaluation results of these words, we observed that in some sentences, none of the nouns and verbs in the context share any common topic with senses of the ambiguous term. For example, using only the semantics of the nouns and verbs in the context, the correct sense of “mood” cannot be distinguished in either of the sentences: “it reflected the *mood* of the moment” (state of the feeling) and “a general *mood* in Whitehall”

(inclination, tendency) . Similar cases were observed for the word “side”: e.g., “both *sides* reaffirmed their commitment” (groups opposing each other) in comparison to “at the *side* of the cottage” (a position to the left or right of a place). While these examples show the limitations of the context-based methods, the overall results show the ability of word embedding and statistical-based approaches for the CL-WSD tasks, specially in the absence of reliable resources.

## 5 Conclusion

We study the opportunities of applying unsupervised approaches on Cross Language Word Sense Disambiguation (CL-WSD), focusing on its application in English to Persian language. The proposed method approaches CL-WSD using embedding of terms in context. We show that our approach outperforms both the CO-Graph system—a state-of-the-art system in unsupervised CL-WSD—as well as the standard baseline.

We however observed fundamental limitations of the methods based exclusively on context as bag of words. Despite this fact, the current work offers a possible solution for all languages/domains with scarce knowledge-based or parallel corpora resources, by exploiting the use of a monolingual corpus together with a simple bilingual lexicon.

## References

1. A. AleAhmad, H. Amiri, E. Darrudi, M. Rahgozar, and F. Oroumchian. Hamshahri: A standard persian text collection. *Knowledge-Based Systems Journal*, 2009.
2. L. Bungum, B. Gambäck, A. Lynam, and E. Marsi. Improving word translation disambiguation by capturing multiword expressions with dictionaries. In *Proc. of NAACL HLT*, 2013.
3. X. Chen, Z. Liu, and M. Sun. A unified model for word sense representation and disambiguation. In *Proc. of EMNLP*, 2014.
4. A. Duque, L. Araujo, and J. Martinez-Romo. Co-graph: A new graph-based technique for cross-lingual word sense disambiguation. *Natural Language Engineering Journal*, 2015.
5. A. Duque, J. Martinez-Romo, and L. Araujo. Choosing the best dictionary for cross-lingual word sense disambiguation. *Knowledge-Based Systems Journal*, 2015.
6. M. A. Farajian. Pen: parallel english-persian news corpus. In *Proc. of the 2011th World Congress in Computer Science, Computer Engineering and Applied Computing*, 2011.
7. A. H. Jadidinejad, F. Mahmoudi, and J. Dehdari. Evaluation of perstem: a simple and efficient stemming algorithm for persian. In *Multilingual Information Access Evaluation I. Text Retrieval Experiments*. 2010.
8. P. Koehn. Europarl: a parallel corpus for statistical machine translation. In *MT Summit*, 2005.
9. E. Lefever and V. Hoste. Semeval-2010 task 3: Cross-lingual word sense disambiguation. In *Proc. of the International Workshop on Semantic Evaluation*, 2010.
10. E. Lefever and V. Hoste. Semeval-2013 task 10: Cross-lingual word sense disambiguation. *Proc. of SemEval*, pages 158–166, 2013.
11. E. Lefever, V. Hoste, and M. De Cock. Parasense or how to use parallel corpora for word sense disambiguation. In *Proc. of ACL-HLT*, 2011.
12. M. Lesk. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proc. of Conference on Systems documentation*, 1986.
13. M. T. Pilevar, H. Faili, and A. H. Pilevar. Tep: Tehran english-persian parallel corpus. In *Computational Linguistics and Intelligent Text Processing*. 2011.
14. N. Rekabsaz, S. Sabetghadam, M. Lupu, L. Andersson, and A. Hanbury. Standard test collection for english-persian cross-lingual word sense disambiguation. In *Proc. of LREC*, 2016.

15. A. Rudnick, C. Liu, and M. Gasser. Hlti: cl-wsd using markov random fields for semeval-2013 task 10. In *Joint Conference on Lexical and Computational Semantics (\*SEM)*, 2013.
16. B. Sarrafzadeh, N. Yakovets, N. Cercone, and A. An. Cross-lingual word sense disambiguation for languages with scarce resources. In *Advances in Artificial Intelligence: Proc. of 24th Canadian Conference on Artificial Intelligence*. Springer Berlin Heidelberg, 2011.
17. 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 word-net. In *Proc. of Global WordNet Conference*, 2010.
18. S. Sofianopoulos, M. Vassiliou, and G. Tambouratzis. Implementing a language-independent mt methodology. In *Proc. of the First Workshop on Multilingual Modeling*. ACL, 2012.