Unsupervised Cross-Lingual Word Sense Disambiguation on Languages with Scarce Resources: Application to Persian

Abstract

We explore the use of unsupervised methods in Cross-Lingual Word Sense Disambiguation (CL-WSD) with the application of English to Persian (Farsi). We create a new test collection for CL-WSD in Persian, following the format of the SemEval 2013 CL-WSD task. We then evaluate our semantic vector word representation-based approach on the new evaluation benchmark and compare it with the standard baseline of the task as well as 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*).

1 Introduction

Word Sense Disambiguation (WSD) is the task of automatically selecting the most related sense for a word occurring in a context. WSD is considered as a main step in the course of approaching language understanding beyond the surface of the words and has been intensively studied in Natural Language Processing (NLP) (Navigli, 2012). WSD is used in Machine Translation (Chan et al., 2007; Costa-Jussà and Farrús, 2014), Information Retrieval (Zhong and Ng, 2012; Na and Ng, 2011), or Entity Linking (Moro et al., 2014).

Typically, the methods for approaching WSD are classified into knowledge-based, supervised, and unsupervised. Knowledge-based approaches use available structured knowledge (Agirre et al., 2014; Miller et al., 2012). Often, the systems exploit WordNet (Fellbaum, 1998) as the sense inventory, together with graph-based methods (Agirre et al., 2010; Guo and Diab, 2010). 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 scarce-resource languages (Pilehvar and Navigli, 2014). As an alternative, unsupervised approaches address WSD using only information extracted from existing corpora, such as various term co-occurrence indicators (Di Marco and Navigli, 2013).

As a paradigm, multilingual and cross-lingual WSD focus on lexical substitution in a target language. The creation of large knowledge resources such as BabelNet (Navigli and Ponzetto, 2010) and DBPedia (Auer et al., 2007) have opened up new possibilities for solving these tasks with multilingual data. For example, Navigli and Ponzetto (2012) exploit BabelNet to develop their graph-based monolingual and bilingual WSD across main European languages.

Cross-lingual Word Sense Disambiguation (CL-WSD) 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. As the task introduces the Europarl corpus (Koehn, 2005) as the main resource, many participating systems exploit these parallel corpora to overcome the knowledge acquisition bottleneck (Lefever et al., 2011; Rudnick et al., 2013). However, these methods 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 lexica) are preferable, if not the only available option (Sofianopoulos et al., 2012). For example, Bungum et al. (2013) find the probable translations of a context in the source language and iden-

tify the best using a language model of the target language. Duque et al. (2015a) 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 Semantic Vector Representations of words, Chen et al. (2014) use knowledge-based WSD to identify distinct representations for different senses of the same term. Our approach for CL-WSD is the opposite of this: starting from general vector representations of terms, it identifies their different senses in context. As vector representations we used two state-of-the-art methods: Word2Vec and GloVe. Word2Vec (Mikolov et al., 2013) introduces a highly incremental and scalable method such that when trained on large datasets, it makes it possible to capture many linguistic subtleties (e.g. similar relation between Italy and Rome in comparison to France and Paris). More recently, GloVe (Pennington et al., 2014) shows superior results in exploiting the implicit knowledge within corpora.

In order to evaluate our system, following the format of SemEval CL-WSD task (Lefever and Hoste, 2013), we contributed Persian as a new language to the framework of the task by providing a new test collection. We compared our approach and the CO-Graph system (Duque et al., 2015a) on this new evaluation collection, observing the advantages of using vector representations in WSD.

The contributions of the work are two-fold:

1. Creating a new CL-WSD benchmark for Persian.
2. Providing a new state-of-the-art for unsupervised CL-WSD methods, based on the use of vector representations of words.

The remainder of this paper is organized as follows: Section 2 investigates the available Persian language resources. Section 3 describes in detail the creation of the new Persian CL-WSD evaluation task. Then, Section 4 explains our unsupervised approach to solve the introduced task, followed by the outline of our experiments in Section 5.

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 "wrestling" and "ship".

In recent years, several tools and libraries were introduced, targeting the complexities of Persian language in NLP. For example, Seraji et al. (2012) provides a set of tools for preprocessing (Pre-Per), sentence segmentation and tokenization (SeTPer), and also POS tagging (TagPer). Dehdari et al. (2008) brings forward a stemmer and morphological analyzer, called PerStem. 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. In this work, similar to Jadidinejad et al. (2010), we use PerStem for stemming, together with TagPer as a state-of-the-art POS tagging tool.

In addition to the NLP tools, knowledge and data resources are an important part of WSD and CL-WSD solutions. 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.

While the knowledge-based systems are limited and only at high cost extendable to more specific domains, exploiting parallel corpora can be another effective method for CL-WSD. The existing parallel corpora (English-Persian) are as follows: Tehran English-Persian Parallel (TEP) (Pilevar et al., 2011) corpus—a free collection extracted from 1600 movie subtitles. The Parallel English-Persian News (PEN) (Farajian, 2011) corpus aligns 30K sentences of news corpora. However, to the extent of our knowledge, this collections is not yet available. Finally, the collection provided by European Language Resource Association (ELRA) which is a commercial collection with approximately 100K aligned sentences. Among the mentioned resources, TEP is the only publicly available one, but it only contains informal conversations, and therefore it does not provide a general representation of the language.

In the absence of reliable and comprehensive resources, our unsupervised CL-WSD method ex-

ploits the use of monolingual corpora. The available text collections in Persian are as follows: The *Hamshahri* collection (AleAhmad et al., 2009), a widely used Persian collection, containing approximately 318K news articles of the Hamshahri newspaper. The articles are of various subjects of economy, sport, politics, psychology, literature, or art from 1996 to 2007. The collection was introduced as the main resource for the Persian task of the CLEF Ad Hoc track (Ferro and Peters, 2010; Agirre et al., 2009) in 2008/2009. The *dotIR* collection[1], released in 2008, is created by crawling 1000K web pages in the .ir domain. Finally, *Bijekhan*[2] and Uppsala Persian Corpus (UPEC) (Seraji et al., 2012) are smaller collections with manually tagged POS data.

Between the Hamshahri and dotIR (as bigger collections), since the Hamshahri collection is more recent and has also been used more in the community of Information Retrieval (IR) and NLP, we selected it as the main resource for our experiments. In addition, in comparison to the dotIR collection, as the content of the Hamshahri collection contains revised newspaper articles, we assume that it is a better representation of the language.

In terms of related work addressing the CL-WSD problem in Persian, Sarrafzadeh et al. (Sarrafzadeh et al., 2011) follows a knowledge-based approach by exploiting FarsNet together with leveraging English sense disambiguation. However, since they evaluate their methods only internally, the results are impossible to compare 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. We then report the result of our unsupervised approach on the provided benchmark and compare it with a state-of-the-art CL-WSD system.

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 in fact 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, they 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”,

[1] http://ece.ut.ac.ir/DBRG/webir/index.html

[2] http://ece.ut.ac.ir/dbrg/Bijankhan

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

a wide variety of direct translation (گلدان, دیگ, قوري, کتري) 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.

The linguist also divided the translations into different senses.The resulting clusters for nouns range from 2 (e.g., "education" to تحصيل or معرفت) to 6 clusters (e.g., "post" to پست, مقام, محل ماموريت, تير عمودي, راي گيري, معني اصطلاحي). 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 choose the appropriate translations in two steps: In the first step, they choose the related sense (cluster). In the second step, the system showed the related translations for the sense, of which they choose up to three translations. In case of no related translation, they choose 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 :: واگن 2; اتومبيل 2; اتوبوس 3; 1;

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

4 Unsupervised CL-WSD Method

In this section, we introduce our unsupervised approach for Cross Lingual Word Sense Disambiguation. The approach follows the main idea of the Lesk algorithm (Lesk, 1986), namely that words in a given context tend to share a common topic. In the absence of external knowledge sources, we use the terms' vector representations 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 and select the one that is most similar to the context.

To formulate our approach, let us define the list T of translation sets for the words in the context: $T = \{T_1, T_2, .., T_n\}$ 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:

$$Sim(t, \bar{t}) = \cos(V_t, V_{\bar{t}}) \quad (1)$$

where V_t is the vector representation of the translation t. However, sometimes the translation t of one term in English may be two or more terms in Persian, and thus we will have more than one vector. We thus define a general similarity between two translations:

[3] Available at resources/sense-inventory
[4] Available at software/webapplication
[5] Available at resources/golden

$$Sim(t, \bar{t}) = \max_{w \in t, \bar{w} \in \bar{t}} (\cos(V_w, V_{\bar{w}})) \quad (2)$$

Obviously when both t and $\bar{t}$ consist of exactly one term, Eq. 2 is equivalent to Eq. 1.

As an example, given the words "railway" and "coach" and their translations "خط راه‌آهن" and "واگن درجه سه" with two and respectively three tokens, the Sim function returns cosine between the vectors of "راه‌آهن" and "واگن" (the highest cosine value among the 6 possible combinations).

In what follows, in order to simplify the annotations, we will use the definition of similarity given by Eq. 1 rather than Eq. 2, thus slightly abusing the notation and using interchangeably t and V_t.

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

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.

Algorithm 1: contextVec Algorithm

Input: candidate translation 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 candidate translation and the context. The final result is multiplied by the probability of the candidate translation:

$$RelAgg(t, T) = \cos(V_t, contextVec(t, T)) \times P(t) \quad (3)$$

The second approach is denoted as *RelGreedy*:

$$RelGreedy(t, T) = \max_{T_i \in T} \left(\max_{\bar{t} \in T_i} (\cos(t, \bar{t})) \right) \times P(t) \quad (4)$$

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

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

$$Result = \arg\max_{t_i} (Rel^*(t_i, T)) \quad (5)$$

where t_i is a translation candidate for the word with ambiguity, and Rel^* can be replaced by RelAgg or RelGreedy.

5 Experiments and Results

As mentioned before, the methods used in this work for facing the CL-WSD problem, exploit the use of a monolingual corpora together with a bilingual lexicon. In the following, first we describe the data acquisition and preparation process for these resources in Persian. Then, we evaluate two baselines on the created Persian CL-WSD benchmark, described in Section 3: the first is the standard baseline introduced in the SemEval 2013 CL-WSD task (Lefever and Hoste, 2013) and the second is a state-of-the-art system called CO-Graph (Duque et al., 2015a). Finally, we evaluate our approach, introduced in Section 4, and compare it with the baselines.

5.1 Data Preparation

As discussed in Section 2, we selected the Hamshahri collection for the required monolingual corpora. We first normalized the collection using Lucene's PersianNormalizationFilter. After normalization, by comparing the tokens in

Table 2: The confusion matrix of 500 annotated tokens after decompounding using AD Tree classifier.

		Predicted	
		Yes	No
Actual	Yes	254 (TP)	84 (FN)
	No	21 (FP)	141 (TN)
		275 (P)	225 (N)

the collection with a comprehensive list of Persian words[6], we observed that from approximately 587K distinct tokens in the collection only 94K(16%) are in the list of the known words. Tracing the collection's tokens, we observed many incorrect tokens, create by concatenation of two words. For example "درکتاب" ("inbook") should have been split in two words: "در کتاب" ("in book").

In order to mitigate this problem, we implemented a greedy decompounder which splits a token in at most two parts. The decompounder uses an existing word-frequency list to find the best splitting alternative. We create this word-frequency list from the Hamshahri collection with the assumption that the words with the higher frequencies in the collection are more probable to be correct. The decompounder first finds all the possible alternatives for splitting the word and uses the list to compute the mean value of the frequencies of the two new tokens. Then, it splits the token if the mean value is higher than the frequency of the original token.

In order to increase the accuracy of the decompounder and avoid incorrectly splitting the words, we randomly selected 500 cases from all the split ones and checked whether they had correctly been decompounded. After the evaluation, we observed an error rate of 32%. In order to decrease this error rate, we applied an AD Tree classifier together with 10-Fold Cross Validation while using the geometric mean, harmonic mean, and standard variation of the frequencies of the split words as the features. The classifier achieved a precision of 0.81. Table 2 shows that the error rate of the system has been reduced to 7.6% (FP/P) as it only decompounds 55% (P/All) of the introduced candidates[7].

Having the designed decompounder, we applied it on the all 587K unique tokens of the collection. While the decompounder split about 132K tokens, the number of unique words in the collection decreased from 587K to 485K, of which 213K(43%) were known by the Persian's words list. Therefore, using the decompounder, we increased the number of the known words 269% (from 94K to 213K) while introducing approximately 7.6% error in the split tokens[8].

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 (Duque et al., 2015b), due to the lack of reliable parallel corpora, we directly extracted the data from Google Translate. Beside the provided translation, the online platform also provides the rate indicating how often the translation is used regarding to its corresponding word in the source language. In our lexicon, this translation probability rate can have one of the three values of 0.25 (rare), 0.5 (uncommon), or 0.75 (common)[9].

5.2 Baselines

In this section, we explain two baselines for the created Persian CL-WSD benchmark. These baselines are further compared with the results of our approach.

The first one—the *Standard* baseline—follows the method introduced in the SemEval 2013 CL-WSD task for creating the baseline. 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 evaluation measure, we observed the value of 15.8 for the *Best Result* and 41.8 for the *Out-Of-Five* evaluation[10].

Since the standard baseline considers only the most common translations, it cannot provide a realistic view on the effectiveness of the CL-WSD systems. Therefore, we evaluate the created Persian benchmark on the state-of-the-art unsupervised CL-WSD system called CO-Graph (Duque et al., 2015a). The CO-Graph system offers competitive results in the SemEval 2010 and SemEval 2013 CL-WSD tasks, for all the proposed languages, namely Spanish, French, Italian, Dutch

[6] http://github.com/reza1615/PersianOcr

[7] The decompounder source code is available at software/decompounder

[8] The Zipf's distribution of the processed Hamshahri and ANC collections is available at supplementary-materials/zipf

[9] Available at resources/dictionary

[10] Available at resources/baseline

and German. It is able to outperform all of the unsupervised participating systems using only monolingual corpora, and even most of the ones which use parallel corpora or knowledge resources.

The initial hypothesis for the CO-Graph system relies on the idea that words in a document tend to (statistically) adopt a related sense. The system first creates a graph of connections between the words, using the documents in the collection, and then applies different algorithms (Dijkstra, Community-based, Static PageRank, Personalized PageRank) to disambiguate the words based on their contexts. The construction of the graph is based on the statistical significance (p-value) of the co-occurrences of the words in the same documents (more details in Duque et al. (2015a)).

In order 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 and parsing the collection using TagPer (Seraji et al., 2012) and PerStem (Dehdari and Lonsdale, 2008) tools. The process of creating the graph took approximately 34 hours. We then evaluated the benchmark using the described lexica and all four algorithms with the p-value range of 10^{-5} to 10^{-17}. Finally, we found Dijkstra algorithm together with p-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[11].

5.3 Evaluation

In this section, we report the evaluation of our approach on the Persian benchmark and compare it with the baselines. In the first step, given the corpus of the processed Hamshahri collection, we created semantic vector representations of the words in Persian using two state-of-the-art methods: Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014). We trained the Word2Vec model using its toolkit[12] by applying the Skip-Gram approach with sub-sampling at $t = 10^{-4}$. The GloVe word representation was also constructed by its toolkit[13] while using its default parameter settings. Regarding the common parameters, we selected the context windows of 5 words, epochs of 25, and words count threshold of 5 (the words with collection frequency less than five are considered as noise and filtered out) for both methods. Using 40 threads, the construction of each model took approximately 45 minutes for Word2Vec, and less than 20 minutes for GloVe. For each method, we prepared a set of three models with vector dimensionalities of 50, 100, and 200.

In the next step, 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 vector representations of the translated words, we calculated the relatedness score of each translation candidate to its context using RelAgg and RelGreedy (Section 4). The translation probability rate in our lexica was used as the $P(t)$ value in Eq. 3 and Eq. 4. Given the

Table 3: Results of F-measure on *Out-Of-Five (OOF)* and *Best Result (Best)* evaluations based on RelAgg and RelGreedy approaches. The vector representations of words were created by the Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) methods with 50 and 200 dimensions, using the corpus of the Hamshahri collection

Eval.	Method	Vector Repres.	F-measure
OOF	RelAgg	W2V-200	50.2
		W2V-50	49.9
		GloVe-200	**50.3**
		GloVe-50	49.7
	RelGreedy	W2V-200	49.3
		W2V-50	49.5
		GloVe-200	50.2
		GloVe-50	50.0
	CO-Graph Dijkstra		44.1
	Standard Baseline		41.8
Best	RelAgg	W2V-200	18.8
		W2V-50	18.3
		GloVe-200	**19.8**
		GloVe-50	18.2
	RelGreedy	W2V-200	18.3
		W2V-50	18.4
		GloVe-200	18.5
		GloVe-50	18.4
	CO-Graph Dijkstra		17.4
	Standard Baseline		15.8

[11]The detailed results are available at supplementary-materials/co-graph

[12]https://code.google.com/p/word2vec

[13]http://nlp.stanford.edu/projects/glove

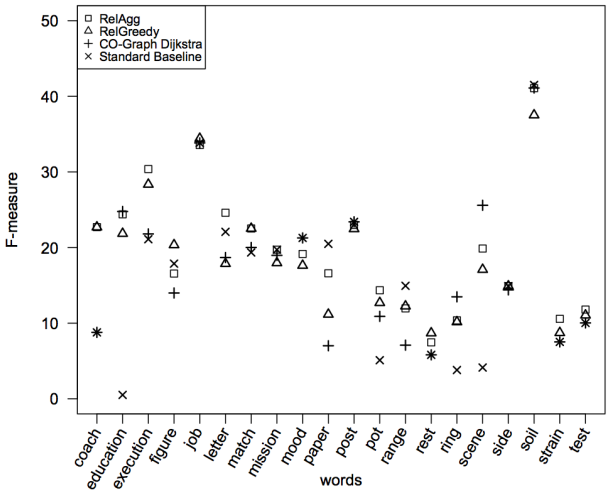

Figure 1: Results of the *OOF Result* evaluation based on each word using F-measure

score of the translation candidates, we created the run files for the *Out-Of-Five* evaluation by selecting the 5 best translations, and the top one for the runs of the *Best Result* evaluation. Table 3 shows the *Out-Of-Five* and *Best Result* evaluation results of RelAgg and RelGreedy relatedness approaches together with Word2Vec's and GloVe's vector representations of words each with 50 and 200 dimensions on the Persian CL-WSD benchmark using the F-measure metrics. The results for dimensionality 100 were very similar to those of 50 and are not reported here for space considerations.

The results for both the *Out-Of-Five* and *Best Result* evaluations show that our approach based on vector representation of the words outperforms the standard as well as 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. Regarding the different vector representations of the words with common evaluation and relatedness methods, we also see very similar results, while the GloVe method with 200 dimensions shows overall better performance.

In order to observe the effectiveness of the systems on different words, we selected the best performing settings in both RelAgg and RelGreedy (GloVe with 200 dimensions), and compared them with the standard and the CO-Graph baselines. The results for the *Out-Of-Five* evaluation are shown in Figure 1[14].

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). These examples show the limitations of the context-based methods. In addition to the context, the probability of occurrence of the words in a specific order (language modeling), potentially including terms of closed POS classes, is probably the missing piece here.

6 Conclusion and Future Work

We study the opportunities of applying unsupervised approaches on CL-WSD, focusing on its application in English to Persian language. The proposed method addresses the CL-WSD problem using semantic vector representations of the words in the context. In addition, addressing the problem of lack of standard evaluation framework for Persian language in the NLP domain, we create and make available a new benchmark for English to Persian CL-WSD following the format of the SemEval 2013 CL-WSD task. Finally, evaluating our approach on the new benchmark, we show that it outperforms both the CO-Graph system–a state-of-the-art system in unsupervised CL-WSD– as well as the standard baseline, for both the *Best* and *Out-of-five* evaluation metrics of the task.

We have however also observed fundamental limitations of the methods based exclusively on context as bag of words. The combination of language modeling and deep learning methods for CL-WSD will probably address these current limitations. Nevertheless, the current work offers a possible solution for all languages or domains with scarce knowledge-based or parallel corpora resources, by exploiting the use of a monolingual corpus together with a bilingual lexicon.

[14]The corresponding plot for the *Best Result* is available at supplementary-materials/evaluation

References

Eneko Agirre, Giorgio Maria Di Nunzio, Nicola Ferro, Thomas Mandl, and Carol Peters. 2009. Clef 2008: ad hoc track overview. In *Evaluating systems for multilingual and multimodal information access*, pages 15–37. Springer.

Eneko Agirre, Aitor Soroa, and Mark Stevenson. 2010. Graph-based word sense disambiguation of biomedical documents. *Bioinformatics*, 26(22):2889–2896.

Eneko Agirre, Oier Lopez de Lacalle, and Aitor Soroa. 2014. Random walks for knowledge-based word sense disambiguation. *Computational Linguistics*, 40(1):57–84.

Abolfazl AleAhmad, Hadi Amiri, Ehsan Darrudi, Masoud Rahgozar, and Farhad Oroumchian. 2009. Hamshahri: A standard persian text collection. *Knowledge-Based Systems*, 22(5):382–387.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. *Dbpedia: A nucleus for a web of open data*. Springer.

Lars Bungum, Björn Gambäck, André Lynum, and Erwin Marsi. 2013. Improving word translation disambiguation by capturing multiword expressions with dictionaries. *NAACL HLT 2013*, 13:21.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *Annual Meeting-Association for Computational Linguistics*, volume 45, page 33. Citeseer.

Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. 2014. A unified model for word sense representation and disambiguation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1025–1035.

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

J Dehdari and D Lonsdale. 2008. A link grammar parser for persian. aspects of iranian linguistics.

Antonio Di Marco and Roberto Navigli. 2013. Clustering and diversifying web search results with graph-based word sense induction. *Computational Linguistics*, 39(3):709–754.

Andres Duque, Lourdes Araujo, and Juan Martinez-Romo. 2015a. Co-graph: A new graph-based technique for cross-lingual word sense disambiguation. *Natural Language Engineering*, FirstView:1–30, 5.

Andres Duque, Juan Martinez-Romo, and Lourdes Araujo. 2015b. Choosing the best dictionary for cross-lingual word sense disambiguation. *Knowledge-Based Systems*, 81:65–75.

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

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

Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.

Nicola Ferro and Carol Peters. 2010. Clef 2009 ad hoc track overview: Tel and persian tasks. In *Multilingual Information Access Evaluation I. Text Retrieval Experiments*, pages 13–35. Springer.

Weiwei Guo and Mona Diab. 2010. Combining orthogonal monolingual and multilingual sources of evidence for all words wsd. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1542–1551. Association for Computational Linguistics.

Amir Hossein Jadidinejad, Fariborz Mahmoudi, and Jon Dehdari. 2010. Evaluation of perstem: a simple and efficient stemming algorithm for persian. In *Multilingual Information Access Evaluation I. Text Retrieval Experiments*, pages 98–101. Springer.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*.

Els Lefever and Veronique Hoste. 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.

Els Lefever and Véronique Hoste. 2013. Semeval-2013 task 10: Cross-lingual word sense disambiguation. *Proc. of SemEval*, pages 158–166.

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

Els Lefever, Véronique Hoste, and Martine De Cock. 2011. Parasense or how to use parallel corpora for word sense disambiguation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 317–322. Association for Computational Linguistics.

Michael Lesk. 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.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tristan Miller, Chris Biemann, Torsten Zesch, and Iryna Gurevych. 2012. Using distributional similarity for lexical expansion in knowledge-based word sense disambiguation. In *COLING*, pages 1781–1796.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244.

Seung-Hoon Na and Hwee Tou Ng. 2011. Enriching document representation via translation for improved monolingual information retrieval. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 853–862. ACM.

Roberto Navigli and Simone Paolo Ponzetto. 2010. Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 216–225. Association for Computational Linguistics.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Joining forces pays off: Multilingual joint word sense disambiguation. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 1399–1410. Association for Computational Linguistics.

Roberto Navigli. 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.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 2014)*, 12.

Mohammad Taher Pilehvar and Roberto Navigli. 2014. A large-scale pseudoword-based evaluation framework for state-of-the-art word sense disambiguation. *Computational Linguistics*, 40(4):837–881.

Mohammad Taher Pilevar, Heshaam Faili, and Abdol Hamid Pilevar. 2011. Tep: Tehran english-persian parallel corpus. In *Computational Linguistics and Intelligent Text Processing*, pages 68–79. Springer.

Alex Rudnick, Can Liu, and Michael Gasser. 2013. Hltdi: Cl-wsd using markov random fields for semeval-2013 task 10. In *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, volume 2, pages 171–177.

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

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

Mojgan Seraji, Beáta Megyesi, and Joakim Nivre. 2012. A basic language resource kit for persian. In *LREC*, pages 2245–2252. Citeseer.

Mehrnoush Shamsfard, Akbar Hesabi, Hakimeh Fadaei, Niloofar Mansoory, Ali Famian, Somayeh Bagherbeigi, Elham Fekri, Maliheh Monshizadeh, and S Mostafa Assi. 2010. Semi automatic development of farsnet; the persian wordnet. In *Proceedings of 5th Global WordNet Conference, Mumbai, India*.

Sokratis Sofianopoulos, Marina Vassiliou, and George Tambouratzis. 2012. Implementing a language-independent mt methodology. In *Proceedings of the First Workshop on Multilingual Modeling*, pages 1–10. Association for Computational Linguistics.

Zhi Zhong and Hwee Tou Ng. 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.