Persian SemCor:
A Bag of Words Sense Annotated Corpus for the Persian Language

Hossein Rouhizadeh , Mehrnoush Shamsfard , Mahdi Dehghan , and Masoud Rouhizadeh

Shahid Beheshti University, Tehran, Iran
hrouhizadeh@gmail.com, m-shams@sbu.ac.ir
mahdi.dehghan551@gmail.com, mrouhizadeh@gmail.com

Abstract

Supervised approaches usually achieve the best performance in the Word Sense Disambiguation problem. However, the unavailability of large sense annotated corpora for many low-resource languages make these approaches inapplicable for them in practice. In this paper, we mitigate this issue for the Persian language by proposing a fully automatic approach for obtaining Persian SemCor (PerSemCor), as a Persian Bag-of-Word (BoW) sense-annotated corpus. We evaluated PerSemCor both intrinsically and extrinsically and showed that it can be effectively used as training sets for Persian supervised WSD systems. To encourage future research on Persian Word Sense Disambiguation, we release the PerSemCor in $nlp.sbu.ac.ir$.

1 Introduction

Word Sense Disambiguation (WSD) is the task of associating ambiguous context words with their most suitable meanings in a pre-defined sense inventory. WSD can be mentioned as a key area in Natural Language Processing (NLP) since it plays a crucial rule in multiple down-stream tasks such as Machine Translation (Neale et al., 2016). The main approaches of WSD can be grouped into two categories i.e., Knowledge-based and Supervised WSD (Raganato et al., 2017b). Knowledge-based WSD systems tend to exploit information from structure or content of lexical resources such as WordNet (Miller et al., 1990) and BabelNet (Ponzetto and Navigli, 2010). On the other hand, the latter approach utilizes machine learning techniques to train a model for automatic sense annotation (Zhong and Ng, 2010),(Raganato et al., 2017a),(Chaplot and Salakhutdinov, 2018). Thanks to the training phase, supervised systems usually outperform the knowledge-based alternatives (Raganato et al., 2017b). In fact, the main reason for the high performance of the supervised systems is the utilization of large manually sense annotated corpus through the training process. Unfortunately, obtaining manually sense annotated corpora such as SemCor (Miller et al., 1993) (i.e. the largest and the most predominant manually sense annotated corpus developed for English) is extremely hard and time-consuming and as a result only a limited number of languages can perform supervised WSD. To tackle this issue, in recent years, a line of research has focused on developing automatic or semi-automatic methodologies capable of producing annotated corpora (Pasini and Navigli, 2017) (Pasini et al., 2018) (Scarlini et al., 2019) (Scarlini et al., 2020a) (Scarlini et al., 2020a) (Barba et al., 2020). Although the developed annotated corpora are multi-lingual and lead the supervised systems to achieve a big improvement in WSD, as mentioned in Scarlini et al. (2019), they suffer from some limitations such as (1): strict dependency on the structure of the knowledge graph, (2): requiring huge parallel corpora. In addition, almost all developed corpora are only limited to nouns and provide no annotated instances for other parts-of-speech (POS) i.e., verbs, adjectives, and adverbs.

In this paper, we focus on developing a fully automatic approach for creating a sense annotated corpus for the Persian language. A key part of the former developed approaches for construction of automatic sense-annotated corpora is the use of high performance pre-processing tools i.e., lemmatizer, tokenizer and POS tagger. However, to the best of our knowledge, developed Persian pre-processing tools can not perform as well as their counterparts for English or other European languages. It could be problematic especially when we need to tokenize multi-words and obtain their lemma for exploiting sense candidates from FarsNet (the Persian WordNet) synsets. To deal with this, we designed our method in such a way that it requires no auto-

matic lemmatizer, tokenizer or PoS tagger.
Our proposed system takes English sense annotated corpora (SemCor, for instance) as input and utilizes inter-language semantic relations between English and Persian to obtain a Bag-of-Words (BOW) sense annotated corpus for Persian. It can be a step forward towards the development of sentence-level sense-annotated corpora for the Persian language. The main contributions of our proposed system are as follows:

- **Obtaining state-of-the-art performance on the SBU-WSD-Corpus**
  Our experiments on the standard Persian All-Words WSD test set, developed by Rouhizadeh et al. (2020), indicates that the supervised baselines, trained on PerSemCor, outperform knowledge-based alternatives and achieve state-of-the-art performance in Persian All-words WSD.

- **Providing sense tagged instances for words with different POS**
  In contrast to the almost all recent automatically developed sense-annotated corpora, PerSemCor is not limited to nominal instances and provide sense annotated samples for all parts-of-speech, i.e. nouns, verbs, adjectives, and adverbs.

- **Low dependency on the structure of knowledge-resource**
  We reduced the dependency on the structure of knowledge-resources by only utilizing one inter-language relation between FarsNet and WordNet (i.e, 'Equal-to relation')

- **No dependency to the performance of Persian pre-processing tools**
  In order to ignore the possible lexical or syntax-based errors in PerSemCor , i.e. the errors that can be generated by Persian tokenizers, lemmatizers or Pos taggers, we designed our approach in such a way that include no dependency on the Persian pre-processing tools.

2 Data and Resources

SemCor: English SemCor (Miller et al., 1993) is a subset of the English Brown corpus and include 352 documents with more than 220K sense annotations. The whole corpus is manually tagged with senses from WordNet 1.6. SemCor can be mentioned as the most widely used sense annotated corpus in the English WSD literature (Scarlini et al., 2020b), (Huang et al., 2019), (Luo et al., 2018b), (Luo et al., 2018a). In this paper, we used SemCor 3.0 which includes mapped sense annotations from WordNet 1.6 to WordNet 3.0. [1]

WordNet: WordNet (Miller et al., 1990) is one of the most widely used lexical resources in the WSD literature. It was initially developed for English at Princeton University. WordNet organizes words and phrases into synsets (sets of synonym words with the same POS) and provides a $gloss$ (descriptive definition of the synset words) and possibly an $example$ (a practical example of synset words) for each of them. WordNet synsets are linked via several lexical and semantic relationships. WordNet 3.0, covers around $155K$ English words and phrases organized in around $117K$ synsets.

FarsNet (The Persian WordNet): FarsNet (Shamsfard et al., 2010) is the first Persian lexical ontology which has been developed in the NLP lab of Shahid Beheshti University[2]. Over the past 12 years, numerous studies have been conducted to develop FarsNet (Rouhizadeh et al., 2007)(Rouhizadeh et al., 2010)(Mansoory et al., 2012) (Yarmohammadi et al., 2008) (Khalghani and Shamsfard, 2018). FarsNet 3.0, the last version of FarsNet, covers more than $100K$ Persian words and phrases. Similar to English WordNet, the basic components of FarsNet are synsets that are interlinked via several types of relations. FarsNet relations can be classified to two main classes: Inner-language and Inter-language relations.
Inner-Language Relations connect pairs of word senses and synsets of FarsNet. More in details, Inner-language relations of FarsNet include two major classes, i.e, Semantic and Lexical relations which are defined between FarsNet senses and synsets, respectively. The Inner-Language relations of FarsNet include all the WordNet 2.1 relations as well as some other relationships like 'patient-of', 'salient', and 'agent-of'.
On the other hand, Inter-Language Relations are held between FarsNet 3.0 and WordNet 3.0 synsets. 'Equal-to' and 'Near-equal-to' are two main classes of this kind of relation. 'Equal-to' indicates that words of two synsets (One in FarsNet and another one in WordNet) have the exactly same meaning and PoS. Whereas, the latter one is representative of the similar (not the same) meaning between two

[1] web.eecs.umich.edu/ mihalcea/downloads.html
[2] http://farsnet.nlp.sbu.ac.ir/Site3/Modules/Public/Default.jsp

synsets. It is worth noting that Inter-Language relations between FarsNet and WordNet are not necessarily pair-wise. In other words, one WordNet synset can be linked to one or more FarsNet synsets via 'Equal-to' relation.

Persian-news Corpus: A key component of our system is leveraging a large Persian raw corpus. Our main objectives to utilize such a corpus is to train word embedding models. Although Wikipedia dumps have shown to be useful for training such models [3] in a variety of languages, Persian Wikipedia articles are often short and are not the best choice for this end. To deal with this, we crawled around 1 M documents from several Iranian news agencies web sites [4] to train the word embedding models on that.

Google Translate: Google Translate is a neural machine translation, developed by Google, which provides both word-level and sentence-level translation tool for more than 100 languages. For each input word w of the source language, the word-level translation tool of Google Translate provides a translation table, consisting of three columns: 1) translation candidates, 2) synonyms of the input word with the same translation and 3) frequency of the translation candidates in the public documents [5].

Figure 1 shows the output of the English-Persian tool of Google Translate for the word 'research'. As can be seen, Google Translate suggests 9 translation candidates for the word 'research' in Persian. Additionally, according to the third column of the output schema, it can be concluded that the Persian words appeared in the first and the fifth row of the figure are the most common translations of 'research' in Persian.

In this paper, we used word-level English-Persian tool of Google Translate in the construction pipeline of the PerSemCor.

Persian all-words WSD test set: For evaluating the supervised systems, trained on PerSemCor, we use SBU-WSD-Corpus (Rouhizadeh et al., 2020) as the only available all-word WSD test set for the Persian language. SBU-WSD-Corpus include 19 documents (16 documents for training and 3 for tuning) covering different domains such as Sports, Medical, Science, Technology, etc. It is annotated with senses from FarsNet 3.0 sense inventory and includes 2784 sense-annotated instances (1764 nouns, 494 verbs, 515 adjectives ,and 111 adverbs).

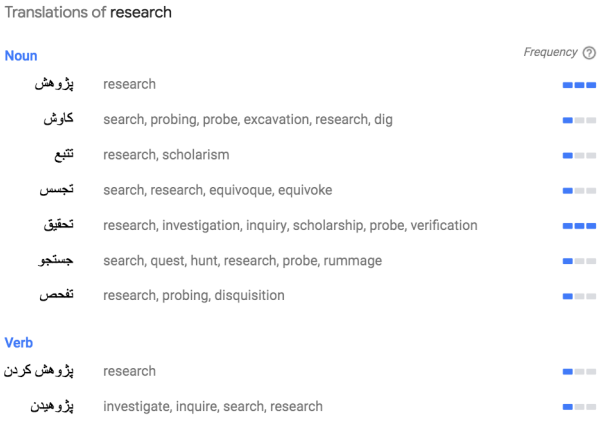

Figure 1: Output of English-Persian Google Translate tool for the word 'research'.

3 Construction of PerSemCor

In this section, we present our proposed approach which aims at automatic construction of PerSemCor, a BoW sense-annotated corpus for the Persian language. The main idea of our proposed approach is inspired by the assumption presented in Bentivogli et al. (2004), i.e, sense annotations of a source language can be transferred to a target language. Given a sense annotated corpus (SemCor, in our case) as input, our proposed system utilizes inter-language semantic relations between English and Persian lexical graphs (WordNet and FarsNet) to obtain a Bag-of-Words(BoW) sense annotated corpus for Persian.

In the following, we first introduce a set of notations which have been used in our proposed approach and then provide details on the way we used the relations between WordNet and FarsNet to create PerSemCor.

3.1 Formal description of notations used in the proposed system

- $S = \{w_{en_1}, ..., w_{en_N}\}$: An English sentence including N English words $(w_{en_1}, ..., w_{en_N})$
- $S\prime = \{w_{p_1}, ..., w_{p_M}\}$: BoW translation of S in Persian including M Persian words $(w_{p_1}, ..., w_{p_M})$
- WN_{key}: Synset key in WordNet[6].

[3] www.dumps.wikimedia.org

[4] we only crawled the news-agencies websites that cover multiple news categories

[5] The length of the blue bar indicates the prevalence of each translation in Persian (see Figure1

[6] Each synset of WorNet is specified with a unique ID (key)

- FN_{key}: Synset key in FarsNet[7]
- $WnSyn_{key}$: The WordNet synset which is identified with the unique ID: key
- $FnSyn_{key}$ The FarsNet synset which is identified with the unique ID: key

3.2 Proposed Approach

Given the English sentence $s = \{w_1, ...w_n\}$ from SemCor, we first remove the stop words and divide the content words into three groups, i.e. C_1, C_2 and C_3. Next, we transfer the words and annotations of C_1, C_2 and C_3 into Persian, respectively.

C_1: The sense-annotated words with one connection with FarsNet

The words of C_1 only include one connection ('equal-to' relation) with FarsNet. For each $w_{en} \in C_1$ which is sense-labeled with WN_{key} (i.e, key of $WnSyn_{key}$), we first retrieve the FarsNet synset $FnSyn_{key}$ which is connected to $WnSyn_{key}$ via 'Equal-to' relation. Although all the present words in $FnSyn_{key}$ share the same meaning, we aim to choose the most suitable one, i.e, $w_p \in FnSyn_{key}$, to make PerSemCor approach to the real Persian texts. Among the synset words, we choose the most frequent one as the best one. To this end, we utilize Google Translate which provides frequency information about the translations of w_{en} (see section 2 for more details) and choose the word w_p with the highest frequency in translation candidates as the best translation.

The proposed approach can be considered as a hybrid approach as it uses semantic and statistical information to transfer (w_{en}, WN_{key}) to (w_p, FN_{key}). More in detail, the approach makes use of 'Equal-to' relations between FarsNet and WordNet which transfer lexical-semantic information from English to Persian. In addition, we employ Google Translate to obtain statistical information of translation candidates and choose the most frequent word one as the final choice.

C_2: The sense-annotated words with at least two connections with FarsNet

As mentioned in section 2, inter-language relations between FarsNet and WordNet are not necessarily pair-wise. Therefore, one annotation key of an English word may have more than one connection to FarsNet. It is worth noting that the FarsNet synsets with the same connection with one WordNet synset share the same meaning. Similar to the former hybrid approach, applied on C_1 words, the aim is to find the best synset which includes the best translation of w_{en} in Persian. To this end, for each $w_{en} \in C_2$, we utilize Google Translate and extract all the possible translations of w_{en} in Persian. Considering $T = \{t_1, ..., t_k\}$ as the possible translations of w_{en} in Persian, we extract the most frequent one $(tj, 0 \leqslant j \leqslant k)$ and choose the synset which include tj as the most suitable synset.

C_3: The words with no connection with FarsNet

These words either do not have a sense label in SemCor or their label does not have a connection to FarsNet. As a result, unlike the words of the former groups, we can not obtain any FarsNet synset to exploit translation candidates. In other words, no semantic information is available via lexical graph connections.

To deal with this, we first utilize the vector representation of former translated words of s (i.e. Persian translation of C_1 and C_2 words) to represent the Persian sentence in semantic space ($V_{s\prime}$). More formally, if the former Persian translated words in $s\prime$ are $\{w_{p_1}, ..., w_{p_k}\}$, $V_{s\prime}$ will be computed as follows:

$$V_{s\prime} = \frac{1}{k} \sum_{i=1}^{k} V(w_{p_i}) \quad (1)$$

where $V(w_{p_i})$ is the vector representation of w_{p_i}.

Next, for each $w_{en} \in C_3$, we utilize Google Translate and extract $T = \{t_1, ..., t_m\}$ as the translation candidates of w_{en} in Persian. Then we compute the cosine similarity between vector representation of each $t_i \in T$ and $V_{s\prime}$ (Formula 2) and choose t_j $(0 \leqslant j \leqslant m)$ with highest similarity as the best translation of w_{en} in Persian.

$$t_j = \arg\max_{t \in \mathcal{T}} Cos(V(t), V_{s\prime}) \quad (2)$$

The result of the above steps is a Persian BoW sentence which is POS tagged, lemmatized, tokenized, and semantically-annotated. We perform the above steps for all the sentences of SemCor and provide PerSemCor as a BoW sense-annotated corpus for Persian. We also provide the general statistics of PerSemcor and compare them with English SemCor in Table 1. The statistics include the number of documents together with the number of sentences, number of annotations (divided per POS), number of distinct senses, number of distinct lemmas, and average polysemy of both PerSemCor and English SemCor.

[7] A unique ID (key) is assigned to each FarsNet synset

	Docs	Sentences	Noun Tags	Verb Tags	Adj Tags	Adv Tags	All tags	Distinct Senses	Distinct lemmas	Average polysemy
En SemCor	352	31176	87002	88334	31784	14787	226036	33,362	22436	6.8
Per SemCor	352	31176	56955	55972	19985	9078	141819	10381	7122	3.5

Table 1: General statistics of English and Persian SemCor.

POS	Noun	Verb	Adjective	Adverb
Coverage	74.0	76.0	82.3	84.7

Table 2: Coverage of PerSemCor on SBU-WSD-Corpus

4 Evaluation

We carried out a number of experiments on PerSemCor to evaluate it both intrinsically and extrinsically. More in detail, in our intrinsic evaluations, we assessed the quality of sense annotations of PerSemCor. In addition, we utilized PerSemCor for training a set of supervised WSD baselines to extrinsically evaluate it.

4.1 Intrinsic Evaluation

In order to assess the intrinsic quality of PerSemCor, i.e. evaluating the generated annotations, we created a golden standard by randomly sampling 100 sentences from English SemCor. As the next step, we translated the sentences into Persian and asked an Iranian linguist to semantically annotate them with FarsNet 3.0 senses. The result of our evaluation, i.e. comparison between manual and automatic sense tags, indicates that our strategy for transferring sense tags from English to Persian seems promising as more than 95% of automatic tags were the same with the manual counterparts. The high quality of the transmitted sense labels can be explained by the fact that all inter-language relationships between FarsNet and WordNet synsets are determined by expert linguists and therefore are very accurate and reliable.

4.2 Extrinsic Evaluation

We exploited the Word Sense Disambiguation task to assess the quality of our automatically- generated corpus. Therefore, we trained a reference WSD model on the data generated by OneSeC and compared the results against those achieved by the same model trained on other re- sources. In order to extrinsically asses the quality of PerSemCor, we employ it as training set for obtaining supervised WSD models. It is worth noting that since no other Persian WSD training set is available, we only compare the obtained results against knowledge-based alternatives. To this end, we make use of knowledge-based benchmarks presented by Rouhizadeh et al. (2020). The WSD approaches include:

- **Most Frequent Sense approach (MFS)**: We used Most Frequent Sense (MFS) approach as our baseline. The approach is context-independent and always choose the most frequent sense of each word in PerSemCor, as the most suitable one.
- **Part-of-Speech based approaches**: These models represent each target word by PoS tags of its surrounding words. For instance, consider the word w_i in a context C including 6 nouns, 2 verbs, 2 adjectives and 1 adverb. We represent w_i with the feature vector $[4, 2, 3, 1]$, where the features are representative of the number of nouns, verbs, adjectives, and adverbs in C, respectively.
- **Word embedding based approaches**: Word embedding models leverage contextual information of raw data to represent words and phrases in a semantic space. They have shown to be useful in many NLP tasks including WSD(Iacobacci et al., 2016). Following Saeed et al. (2019), we carried out several experiments to demonstrate the benefit of using such models in the training phase of WSD models. In addition, we were interested to check the impact of different word embedding models on the performance of WSD models. To this end, we trained two word embedding models, i.e. word2vec (Mikolov et al., 2013) and Glove (Pennington et al., 2014) on Persian-news corpus (see section 2) and carried out the same experiments with them. For each target word w_i in a context $C = \{w_1, ..., w_m\}$, we represent w_i with a n-dimensional vector (n is the size of embedding vectors) which is the average of word vectors of C.

		Noun			Verb			Adj			Adv			All		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
FN 1st sense		48.4	48.4	48.4	43.5	43.5	43.5	81.1	81.1	81.1	90.0	90.0	90.0	55.0	55.0	55.0
MFS		58.0	51.8	54.8	70.0	56.8	62.7	84.8	74.6	79.3	93.6	79.3	85.9	66.1	68.2	61.7
MLP	POS	61.0	55.2	58.0	77.2	65.2	70.7	89.6	79.3	84.2	90.1	81.0	85.7	72.3	63.0	67.3
	W2V	64.0	58.0	60.8	77.9	65.8	71.3	90.1	79.8	84.7	90.1	81.0	85.7	74.3	64.8	69.2
	Glove	64.1	58.0	61.0	78.4	66.2	71.8	89.7	79.4	84.2	90.1	81.0	85.7	74.4	64.8	69.3
DT	POS	58.3	52.8	55.4	76.5	64.6	70.0	88.9	78.6	83.4	90.1	81.0	85.7	70.4	61.4	65.6
	W2V	61.5	55.8	58.5	75.5	63.7	69.2	90.5	80.2	85.0	90.1	81.0	85.7	72.5	63.2	67.5
	Glove	61.7	55.8	58.6	70.3	59.3	64.3	89.3	79.0	83.8	90.1	81.0	85.7	71.5	62.1	66.5
KNN	POS	58.8	53.2	55.8	70.1	64.7	67.3	90.3	80.0	84.9	90.1	81.0	85.7	69.8	61.9	65.6
	W2V	62.9	57.0	59.8	71.4	65.8	71.4	90.6	80.2	85.1	90.1	81.0	85.7	72.7	64.3	68.2
	Glove	62.1	56.2	59.0	71.8	66.2	71.7	**91.2**	**80.8**	**85.7**	90.1	81.0	85.7	72.4	64.0	68.0
SVM	POS	62.4	56.5	59.3	77.2	65.3	70.7	90.0	79.6	84.4	90.1	81.0	85.7	73.2	63.8	68.21
	W2V	**64.2**	**58.2**	**61.1**	78.6	66.3	72.0	90.4	80.0	84.9	90.1	81.0	85.7	**74.6**	**69.5**	**69.5**
	Glove	62.8	56.9	59.7	**78.8**	**66.6**	**72.3**	91.0	80.6	85.5	90.1	81.0	85.7	71.9	65.5	68.5

Table 3: Precision (P), Recall (R) and F-1 (F) performance of supervised WSD systems on SBU-WSD-Corpus

	Noun	Verb	Adjective	Adverb	All
MFS	54.8 — 59.2	62.7 — 65.0	79.3 — 84.2	85.9 — 90.1	61.7 — 65.8
MLP	61.0 — 64.9	71.8 — 73.1	84.2 — 89.5	85.7 — 90.1	69.3 — 72.4
DT	58.5 — 63.2	69.2 — 71.5	85.0 — 90.1	85.7 — 90.1	67.5 — 70.6
KNN	59.8 — 64.8	71.4 — 73.7	85.1 — **90.2**	85.7 — 90.1	68.2 — 71.4
SVM	61.1 — **65.0**	72.0 — **74.3**	84.9 — 90.0	85.7 — 90.1	69.5 — **72.7**

Table 4: Comparison between performance of the supervised WSD systems when the MFS back-off strategy is disabled (the number to the left of each cell) or enabled (the number to the right of each cell).

Machine learning algorithms: Following (Saeed et al., 2019), we employed four machine learning techniques, i.e. Support Vector Machine (SVM) (Cortes and Vapnik, 1995), K-Nearest Neighbor (KNN) (Altman, 1992) , Decision Tree (DT) (Black, 1988), and Multilayer Perceptron (MLP) (McCulloch and Pitts, 1943), which utilize the feature vectors, obtained by mentioned approaches, to train WSD models. We also compare the performance of the supervised models with MFS as the baseline of the supervised systems. In addition, we compare the results with FarsNet first sense approach[8] as the former baseline of Persian WSD (Rouhizadeh et al., 2020).

Results and analysis: In Table 3, we compare the performance of different machine learning algorithms when trained by different approaches. It is worth noting that PerSemCor is capable of covering most context words of SBU-WSD-Corpus (see Table 2). In order to clearly show the effect of PerSemCor in the final performance WSD systems, we report the precision (P), recall (R), and the harmonic mean (F1) of different systems, broken by PoS, when no back-off strategy was used.

As expected, the F1-performance of all systems on nouns is lower than other parts-of-speech. This can be explained by the ambiguity level of nouns in the SBU-WSD-Corpus as it is greater than all the other parts-of-speech. As can be seen, MFS can outperform the FarsNet 1st sense approach on disambiguating nouns and verbs by a large margin (10% on nouns and 18% on verbs). It clearly shows the potential of PerSemCor in providing information about sense distribution of Persian words.

Comparing different approaches, the results show that all machine learning algorithms achieve the highest performance when they use word embedding approaches as feature vectors for training. It clearly shows the great impact of using embedding vectors in a WSD pipeline. However, as can be seen, the use of different word embedding models does not greatly affect the final performance of the systems. Comparing machine learning algorithms, SVM outperforms all the other ones in almost all cases. In addition, the best results obtained when SVM trained with the word embedding based feature vectors.

Additional experiments:

1. **Applying Back-off strategy**: A back-off strategy is an alternative method that is used when our system is unable to decide the meaning of the

[8]The approach simply chooses the first sense of FarsNet as the best meaning of each word

		Noun	Verb	Adjective	Adverb	All
Supervised Systems	MFS	59.2	65.0	84.2	90.1	65.8
	MLP	64.9	73.1	89.5	90.1	72.4
	DT	63.2	71.5	90.1	90.1	70.6
	KNN	64.8	73.7	**90.2**	90.1	71.4
	SVM	**65.0**	**74.3**	90.0	90.1	**72.7**
Knowledge Based Systems	FarsNet 1st sense	48.4	43.5	81.1	90.0	55.0
	Basile14	62.7	66.3	83.6	82.9	67.8
	UKB (ppr)	58.4	70.5	82.4	83.6	65.7
	UKB (ppr-w2w)	58.3	71.5	84.4	84.5	66.2

Table 5: F-1 performance of different supervised and knowledge-based models on SBU-WSD-Corpus

input word. For instance, for the words occurring only with one meaning in the training data, we can use MFS as the back-off strategy[9]. This technique has shown to be helpful in several developed WSD systems (Raganato et al., 2017b). To test the effect of using a back-off strategy, we, therefore, decided to perform additional experiments on PerSemCor when the MFS back-off strategy is used[10]. As can be seen in Table 4, all the WSD models achieve higher performance when MFS back-off is used. It is indicative of the usefulness of applying this technique in multiple WSD pipelines.

2. **Comparison with knowledge-based systems**

In Table 5, we compared the F1 performance of supervised models against knowledge-based benchmarks (Rouhizadeh et al., 2020), including Basile14 (Basile et al., 2014), UKB (Agirre et al., 2018) and FarsNet 1st sense (baseline of knowledge-based models). The results show that supervised systems outperform knowledge-based models on all parts-of-speech. It clearly shows the high ability of PerSemCor on training WSD models as it leads simple supervised baselines to state-of-the-art performance when compared against the most recent knowledge-based models. More interestingly, the simplest supervised approach, i.e. MFS approach, is able to achieve competitive results with state-of-the-art knowledge-based systems. It will be more impressive considering that PerSemCor generated without any human intervention.

[9]Note that for the words which never occur in the training data, we consider the first sense of FarsNet as the most predominant one (Raganato et al., 2017b)(Rouhizadeh et al., 2019)

[10]For each machine learning technique, we only report the result of best performing setting

5 Related Work

Knowledge acquisition bottleneck i.e, producing a large amount of lexical-semantic data, can be mentioned as one of the most important problems in WSD. It is more crucial when it comes to supervised WSD as these types of systems need sense annotated data for training a machine learning model. Over recent decades, a variety of approaches have been proposed to mitigate this issue. They can be grouped into two main categories:

Manual annotation, where all the sense tags of the corpora are provided by human efforts. SemCor is one of the first manually annotated corpora for English, developed by the WordNet Project research team at Princeton University. It was initially tagged with senses for WordNet 2.1 and contains more than 200k sense annotated instances. Although SemCor has lead the supervised systems to achieve state-of-the-art performance in English WSD, obtaining such corpora is hard and time-consuming. To reduce or eliminate human intervention for obtaining semi-automatically or fully automatically sense-annotated corpora, a range of approaches have been proposed

Automatic annotation, where a semi-automatic or fully automatic approach is used to generate sense tags.

OMSTI (One Million Sense-Tagged Instances) (Taghipour and Ng, 2015) can be mentioned as one the largest and most predominant sense-tagged corpora for English, created in a semi-automatically manner. The authors of the paper leveraged a large English-Chinese parallel corpus and manual translations of senses to obtain one million training instances. Another group of systems make use of formerly annotated corpora in English, SemCor for instance, to create a new sense-tagged corpora for a second language. Bentivogli et al. (2004) and

Bond et al. (2012) used a parallel corpus (a subset of the SemCor) to create a sense-annotated corpus for the Italian and Japanese languages, respectively. Both approaches utilized word level alignments between the sentences of the parallel corpora to semantically annotate the target instances. Bovi et al. (2017) utilized Babelfy (Moro et al., 2014) as a language independent WSD system and NASARI (Camacho-Collados et al., 2016) as a vector representation of concepts to develop a parallel sense-annotated corpus for four European languages. Pasini and Navigli (2017) and Pasini and Navigli (2020) eliminated the requirement of parallel corpora by proposing Train-O-Matic, which makes use of structural-semantic information from a lexical network to automatically annotate the context words. Scarlini et al. (2019), also proposed a system which leverages Wikipedia categories and semantic vector of concepts to perform automatic sense annotation. The most similar method to our work is proposed by Barba et al. (2020). They make use of multi-lingual BERT and BabelNet to project senses from SemCor to the sentences in low-resource languages. However, the proposed system relies on high-performance pre-processing tools which are not available for Persian. In addition, the only available All-Words WSD test set for Persian is SBU-WSD corpus which is tagged based on FarsNet 3.0 senses, and as a result, the proposed approach can not be evaluated on Persian.
Considering the unavailability of key components of the formerly developed approaches for Persian (English-Persian word alignment tool: (Bond et al., 2012), (Bentivogli et al., 2004)), large English-Persian parallel corpora: (Bovi et al., 2017), high-performance tokenizer, and lemmatizer: (Pasini and Navigli, 2017) ,(Pasini and Navigli, 2020), (Scarlini et al., 2019), (Barba et al., 2020)), we propose a fully automatic approach to obtain a sense annotated corpus for the Persian language. In contrast to the most aforementioned approaches, which only provide sense-annotated nominal instances, our approach provides sense-annotated samples for all parts-of-speech (nouns, verbs, adjectives, and adverbs).

6 Conclusion

In this paper, we presented PerSemCor, a fully-automatic constructed sense-annotated corpus for the Persian language. Our approach for building PerSemCor includes no human intervention as it uses semantic inter-language relations to annotate the Persian words. Moreover, we eliminated the burden of high-performance pre-processing tools, i.e. tokenizer and lemmatizer, as they can be a source of error in constructing training data sets for the Persian Language. We evaluated the built corpus, PerSemCor, both intrinsically and extrinsically, and proved that it can count as a high-quality sense-annotated corpus for training supervised Persian WSD models. As the future work, we plan to create a Persian sentence-level sense-annotated corpus by employing a 'BoW2seq' approach, i.e. an approach which takes a set of shuffled words of a sentence as input and reorder them like a real sentence.

References

Eneko Agirre, Oier López de Lacalle, and Aitor Soroa. 2018. The risk of sub-optimal use of open source NLP software: UKB is inadvertently state-of-the-art in knowledge-based WSD. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, Melbourne, Australia. Association for Computational Linguistics.

Naomi S Altman. 1992. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185.

Edoardo Barba, Luigi Procopio, Niccolo Campolungo, Tommaso Pasini, and Roberto Navigli. 2020. Mulan: Multilingual label propagation for word sense disambiguation. In *Proc. of IJCAI*, pages 3837–3844.

Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. 2014. An enhanced lesk word sense disambiguation algorithm through a distributional semantic model. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1591–1600.

Luisa Bentivogli, Pamela Forner, and Emanuele Pianta. 2004. Evaluating cross-language annotation transfer in the multisemcor corpus. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 364–370.

Ezra Black. 1988. An experiment in computational discrimination of english word senses. *IBM Journal of research and development*, 32(2):185–194.

Francis Bond, Timothy Baldwin, Richard Fothergill, and Kiyotaka Uchimoto. 2012. Japanese semcor: A sense-tagged corpus of japanese. In *Proceedings of the 6th global WordNet conference (GWC 2012)*, pages 56–63. Citeseer.

Claudio Delli Bovi, Jose Camacho-Collados, Alessandro Raganato, and Roberto Navigli. 2017. Eurosense: Automatic harvesting of multilingual sense

annotations from parallel text. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 594–600.

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, 240:36–64.

Devendra Singh Chaplot and Ruslan Salakhutdinov. 2018. Knowledge-based word sense disambiguation using topic models. In *AAAI*.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. Glossbert: Bert for word sense disambiguation with gloss knowledge. *arXiv preprint arXiv:1908.07245*.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Embeddings for word sense disambiguation: An evaluation study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 897–907.

Fatemeh Khalghani and Mehrnoush Shamsfard. 2018. Extraction of verbal synsets and relations for farsnet. In *Proceedings of the 9th Global WordNet Conference (GWC 2018)*, page 424.

Fuli Luo, Tianyu Liu, Zexue He, Qiaolin Xia, Zhifang Sui, and Baobao Chang. 2018a. Leveraging gloss knowledge in neural word sense disambiguation by hierarchical co-attention. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1402–1411.

Fuli Luo, Tianyu Liu, Qiaolin Xia, Baobao Chang, and Zhifang Sui. 2018b. Incorporating glosses into neural word sense disambiguation. *arXiv preprint arXiv:1805.08028*.

Niloofar Mansoory, Mehrnoush Shamsfard, and Masoud Rouhizadeh. 2012. Compound verbs in persian wordnet. *International Journal of Lexicography*, 25(1):50–67.

Warren S McCulloch and Walter Pitts. 1943. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. 1990. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244.

George A Miller, Claudia Leacock, Randee Tengi, and Ross T Bunker. 1993. A semantic concordance. In *HUMAN LANGUAGE TECHNOLOGY: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

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.

Steven Neale, Luís Gomes, Eneko Agirre, Oier Lopez de Lacalle, and António Branco. 2016. Word sense-aware machine translation: Including senses as contextual features for improved translation models. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2777–2783.

Tommaso Pasini, Francesco Maria Elia, and Roberto Navigli. 2018. Huge automatically extracted training sets for multilingual word sense disambiguation. *arXiv preprint arXiv:1805.04685*.

Tommaso Pasini and Roberto Navigli. 2017. Train-o-matic: Large-scale supervised word sense disambiguation in multiple languages without manual training data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 78–88.

Tommaso Pasini and Roberto Navigli. 2020. Train-o-matic: Supervised word sense disambiguation with no (manual) effort. *Artificial Intelligence*, 279:103215.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Simone Paolo Ponzetto and Roberto Navigli. 2010. Knowledge-rich word sense disambiguation rivaling supervised systems. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 1522–1531.

Alessandro Raganato, Claudio Delli Bovi, and Roberto Navigli. 2017a. Neural sequence learning models for word sense disambiguation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1156–1167.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017b. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110.

Hossein Rouhizadeh, Mehrnoush Shamsfard, and Masoud Rouhizadeh. 2019. Knowledge-based word sense disambiguation with distributional semantic expansion. In *Proceedings of the 2019 Workshop on Widening NLP*.

Hossein Rouhizadeh, Mehrnoush Shamsfard, and Masoud Rouhizadeh. 2020. Knowledge-based word sense disambiguation with distributional semantic expansion for the persian language. In *2020 10th International Conference on Computer and Knowledge Engineering (ICCKE)*. IEEE.

Masoud Rouhizadeh, Mehrnoush Shamsfard, and Mahsa A Yarmohammadi. 2007. Building a wordnet for persian verbs. *GWC 2008*, page 406.

Masoud Rouhizadeh, A Yarmohammadi, and Mehrnoush Shamsfard. 2010. Developing the persian wordnet of verbs: Issues of compound verbs and building the editor. In *Proceedings of 5th Global WordNet Conference*.

Ali Saeed, Rao Muhammad Adeel Nawab, Mark Stevenson, and Paul Rayson. 2019. A word sense disambiguation corpus for urdu. *Language Resources and Evaluation*, 53(3):397–418.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2019. Just "onesec" for producing multilingual sense-annotated data. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 699–709.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020a. Sense-annotated corpora for word sense disambiguation in multiple languages and domains. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5905–5911.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020b. Sensembert: Context-enhanced sense embeddings for multilingual word sense disambiguation. In *AAAI*, pages 8758–8765.

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*, volume 29.

Kaveh Taghipour and Hwee Tou Ng. 2015. One million sense-tagged instances for word sense disambiguation and induction. In *Proceedings of the nineteenth conference on computational natural language learning*, pages 338–344.

Mahsa A Yarmohammadi, Mehrnoush Shamsfard, Mahshid A Yarmohammadi, and Masoud Rouhizadeh. 2008. Sbuqa question answering system. In *Computer Society of Iran Computer Conference*, pages 316–323. Springer.

Zhi Zhong and Hwee Tou Ng. 2010. It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 system demonstrations*, pages 78–83. Association for Computational Linguistics.