IJWR

SBU-WSD-Corpus: A Sense Annotated Corpus for Persian All-words Word Sense Disambiguation

Hossein Rouhizadeh[a], Mehrnoush Shamsfard*[b], Vahide Tajalli[c]
[a,b] Faculty of Computer Science and Engineering, Shahid Beheshti University, Tehran, Iran; hrouhizadeh@gmail.com[a], m-shams@sbu.ac.ir[b]
[c] University of Tehran, Tehran, Iran; vtajalli@gmail.com

ABSTRACT

Word Sense Disambiguation (WSD) is a long standing task in Natural Language Processing (NLP) that aims to automatically identify the most relevant meaning of the words in a given context. Developing standard WSD test collections can be mentioned as an important prerequisite for developing and evaluating different WSD systems in the language of interest. Although many WSD test collections have been developed for a variety of languages, no standard All-words WSD benchmark is available for Persian. In this paper, we address this shortage for the Persian language by introducing SBU-WSD-Corpus, as the first standard test set for the Persian All-words WSD task. SBU-WSD-Corpus is manually annotated with senses from the Persian WordNet (FarsNet) sense inventory. To this end, three annotators used SAMP (a tool for sense annotation based on FarsNet lexical graph) to perform the annotation task. SBU-WSD-Corpus consists of 19 Persian documents in different domains such as Sports, Science, Arts, etc. It includes 5892 content words of Persian running text and 3371 manually sense annotated words (2073 nouns, 566 verbs, 610 adjectives, and 122 adverbs). Providing baselines for future studies on the Persian All-words WSD task, we evaluate several WSD models on SBU-WSD-Corpus.

Keywords*— *Word Sense Disambiguation; WSD Corpus; All-words WSD; Persian Language Processing

1. Introduction

Word Sense Disambiguation (WSD) is an open problem in Natural Language Processing (NLP) which aims to automatically recognize the correct meaning of ambiguous words in a particular context. For instance, consider the sentence "The bank will lend us money." where we want to disambiguate the word bank. Retrieving all possible meanings of bank from a pre-defined sense inventory (WordNet, for instance), a WSD algorithm should be ideally able to associate the word bank with its financial institute meaning.

WSD has applications in other NLP tasks such as Machine Translation [7], Information Retrieval and Extraction [51], Question Answering [36], etc. WSD tasks can be distinguished into two generic categories: (1) Lexical Sample WSD and (2) All-words WSD. Developed Lexical Sample WSD systems aim to disambiguate a set of restricted predefined words. Whereas the goal of developing All-words WSD systems is disambiguating all occurring words in a particular context. Generally, All-words WSD approaches are useful for downstream NLP applications [42]. Compared to the Lexical Sample approaches, developing such systems seems to be more challenging. This is mainly because the developed All-words WSD systems should ideally be able to cover a wide range of open-class words in the language of interest. Whereas, the Lexical Sample systems only require disambiguating a limited number of words. In this paper, we focus on All-Words WSD for the Persian language.

WSD approaches can be grouped into two main approaches: (1): knowledge-based and (2): supervised. Knowledge-based WSD approaches exploit information from a lexical resource such as machine-readable dictionaries, thesauri, and ontology to perform WSD. On the other hand, supervised systems apply machine learning techniques on a sense-annotated corpus to train WSD models. Thanks to the training phase, supervised systems generally outperform knowledge based alternatives. It worth noting that, due to the unavailability of sense-annotated corpora for many languages, performing supervised WSD is not possible. .While, knowledge-based approaches only require lexical resources (as sense repositories) that are available for a wide range of languages and can be used as an appropriate alternative. On the other hand, the sense repositories may be absent or may have some gaps and some words or senses be missed in them, in these cases, WSI (word sense induction) methods are used to identify word senses. WSI uses various methods such as clustering to induce the approximate meaning of the word. Evaluating all of these tasks including WSI and supervised and knowledge based WSD can be done on the same test sets; sense-annotated corpora. To the best of our knowledge, the only available All-words sense-annotated corpus for the Persian language is Persian SemCor [49], which have been developed automatically. Previous studies on All-words WSD have focused on a variety of languages such as English, Dutch, Italian, etc [30], [32], [34] However, many low-resource languages such as Persian have not been studied as well. In this paper, we introduce and discuss the creation pipeline of SBU-

http://dx.doi.org/10.22133/ijwr.2023.354098.1128
Citation H. Rouhizadeh, M. Shamsfard and V. Tajalli, " SBU-WSD-Corpus: A Sense Annotated Corpus for Persian All-words Word Sense Disambiguation," *International Journal of Web Research*, vol.5, no.2,pp.77 -85, 2022, doi: http://dx.doi.org/10.22133/ ijwr.2023.354098.1128.

*Coressponding Author
Article History: Received: 8 August 2022; Revised: 27 December 2022; Accepted: 30 December 2022

WSD-Corpus as the first developed test set for the Persian All-words WSD.

Persian (also known as Farsi) is an Indo-European (IE) language that is currently spoken by more than 110 million people in several countries such as Iran, Afghanistan, and Tajikistan. Persian language uses a modified Arabic script and is written from right to left. Millions of Persian texts are available via online web pages, newspapers, books, etc. As a result, there is no doubt in the necessity of developing computational models for Persian as a low-resource language [44]. Similar to other fields of study, standard test sets are required for evaluating WSD approaches. However, none is available for Persian. The main objective of this research is to address the lack of an All-words WSD test set for the Persian language.

SBU-WSD-Corpus contains 5892 content words of Persian running text. The corpus includes 3371 instances (2073 nouns, 566 verbs, 610 adjectives, and 122 adverbs) which are manually annotated by three annotators. We benchmark SBU-WSD-Corpus with several supervised and knowledge-based WSD models, providing baseline results for future research on All-Words WSD for the Persian language. The main contributions of this research are as follows:

Creating a standard All-words WSD dataset:

With the goal of developing a standard All-Words WSD data set, we followed all guidelines, suggested by SensEval-2 [9]. To the best of our knowledge, this is the first available test set for Persian All-words WSD task. With the introduction of SBU-WSD-Corpus, we hope to open avenues for future WSD research in Persian. Additionally, we provide details of our corpus creation pipeline, which can be useful for researchers of other low resource languages to develop similar useful resources.

Presenting benchmarks for future research in Persian All-Words WSD:

To provide baseline for evaluation of Persian All-Words WSD systems, a set of best performing supervised (trained on Persain SemCor) and knowledge-based WSD systems are carried out on SBU-WSD-Corpus. In addition detailed analysis and comparison between different systems are provided.

Usefulness of SBU-WSD-Corpus for evaluation of other Persian NLP tasks:

The whole documents of SBU-WSD-Corpus has been manually tokenized, Pos-tagged and lemmatized by an expert linguist. As a result, it can be also used as a test set for evaluating a range of basic Persian preprocessing tools such as PoS-taggers, lemmatizers, tokenizers and sentence segmentations, etc.

Free access to the developed data set

To encourage future research on Persian All-words WSD, SBU-WSD-Corpus will be freely available for the research community.

The rest of the paper is structured as follows. Section 2 surveys a range of related works. Section 3 describes the different steps of creating the corpus. Section 4 introduces the WSD experiments applied to the corpus. In section 5 the results and the analysis about the performance of the evaluated benchmarks are presented. Finally, the conclusions and further possible works are found in section 6.

2. Related Work

Over recent decades, a variety of sense annotated corpora have been developed for both All-words and Lexical-Sample WSD tasks. Generally, sense annotated corpora can be divided into two main groups: (a)WSD Training corpora and (b)WSD Test Set corpora.

WSD Training corpora, which includes a variety of sense annotated samples in the language of interest. Sense annotated corpora for lexical sample task only include annotated samples for the limited number of predefined words. However, All-words sense annotated corpora should ideally cover multiple instances for a wide range of open-class words. Among the developed training WSD datasets, we briefly introduce SemCor [22] and its different versions, OMSTI [44], the Italian Syntactic-Semantic Treebank [24] CLE Urdu Sense Tagged corpus [50] as All-words WSD datasets and DSO corpus [29], Line-hard-Serve corpus [23] and the Interest corpus [6] as Lexical Sample WSD datasets in the following.

SemCor, is the first and most prominent All-words Sense annotated corpora for English. SemCor contains 352 manually tagged documents (Taken from Brown corpus [11]) and includes 226040 sense annotations. It was initially tagged with senses from WordNet 2.1. Sense tags of the current version of SemCor, are mapped to WordNet 3.0 senses. Different versions of SemCor are also available for some other languages. [5], Eusemcor [1], Bsemcor [16] and Spsemcor [14] are developed versions of Semcor for Japanese, Basque, Bulgarian and Spanish languages respectively. OMSTI (One Million Sense-Tagged Instances) is another widely used All-words sense annotated corpora for English. It was semi-automatically annotated with senses from WordNet 3.0 and includes 911134 sense annotations in 813798 sentences. An English-Chinese parallel corpus [10] is used for the construction of OMSTI. The Italian Syntactic Semantic Treebank (ISST) is an Italian manually All-words sense-annotated corpus. The corpus consists of 305547 tokens including 81236 manually sense tagged words, annotated with Italian WordNet [41]. CLE Urdu Digest corpus is the Urdu All-words WSD corpus which contains 17006 sense annotated nouns, tagged with senses from CLE Urdu WordNet [50]. The Lexical sample sense annotated corpora surveyed in this section include DSO, Line-hard-Serve and the Interest corpus. DSO corpus is a manually sense-annotated corpus including 192800 sentences drawn from Brown corpus and Wall Street Journal. 121 nouns and 70 verbs have been tagged with senses from WordNet 1.5. Line-hard-Serve is another predominant English lexical-sample corpus. It includes 12000 instances from the American Printing House for the Blind, and the San Jose Mercury of the words line (noun), hard (adjective), and serve (verb).

WSD Test Set corpora, which are not as large as training corpora and as a result, are not appropriate for use as training sets in supervised approaches.

The major part of developed WSD benchmark corpora for both Lexical Sample and an All-words WSD tasks belongs to SensEval (The International Workshop on Evaluating Word Sense Disambiguation Systems) and SemEval (The

International Workshop on Semantic Evaluation) competitions. SemEval (the new name of Senseval) is an ongoing series of evaluations of computational semantics systems in several languages. The main focus of Senseval-1 through Senseval-3 was on both All-words and Lexical Sample WSD tasks. The fourth series of Senseval (renamed Semeval) has been expanded to the evaluation of computational semantic analysis systems not necessarily related to WSD. The outcomes of the competitions are standard WSD frameworks for multiple languages [9], [26] For instance, the main benchmark for English All-words WSD (presented by [34]) is the unified version of different Senseval and Semeval English All-words WSD tasks [9], [47], [33], [28]. It contains 7253 sense annotated instances (4200 nouns, 1652 verbs, 955 adjectives and 346 adverbs) annotated with senses from WordNet 3.0 sense inventory. A variety of European languages such as English, French, Dutch, Italian and Spanish are covered by different series of theses competitions.

Recently, [42] and [43] developed a Lexical Sample and All-words test sets for the Urdu language. The Lexical Sample corpus includes sense annotated samples for 50 target words (30 nouns, 11 adjectives, and 9 adverbs) and the All-words corpus contains 5042 words of Urdu running text and 466 sense annotated words. Ambiguous words within both corpora were manually tagged with senses from the Urdu Lughat dictionary.

There are also some researches on Persian WSD and WSI ([49], [12], [13], [14], [25], [18], [46] and [19]) and some small-sized corpora are developed.

The only Persian all-words corpus available so far is PerSemCor (Persian SemCor) [50] which is developed automatically from English SemCor. PerSemCor consists of 352 documents containing 31176 un-ordered sentences, 141819 sense tags over 10381 dictinct senses from FarsNet 3.0 sense repository.

Other developed corpora are small-sized some-words corpora focusing on a small number (4-20) of target words. For example focused on 20 English ambiguous words and added 50 sentences for each to a corpus. Then they translated the corpus to Persian to create a some-words Persian corpus suitable for cross-lingual WSD.

Makki and Homayounpour [19] applied an unsupervised, thesaurus based method on 15 homograph target words.

Masoudi and ghouchani [18] worked on 15 target words and developed a test corpus of 10623 sentences annotated for 32 senses of the 15 target words,

Mahmoudvand and Hourali [17] proposed a Semi-supervised approach for some-words Persian WSD and tested their approach on a corpus with 5368 sentences annotated for three target words.

Moradi and colleagues [25] proposed an unsupervised disambiguation method using trained word2vec model of the second language. They selected 200 sample sentences for each 4 target words as their testset.

Ghayoomi [12] and [23] proposed a WSI method for 20 target words. He gathered 100 sentences for each word and made a corpus of 2000 sentences which are manually annotated according to SemEval 2010 standards.

As seen, although several WSD training and test corpora have been developed for a variety of languages, no WSD golden standard corpus is available for Persian. To address the lack of a standard WSD benchmark for Persian, we put forward SBU-WSD Corpus as the first Persian All-words WSD test set. We also carried out a set of best performing WSD systems on SBU-WSD-Corpus as baseline for future researches in Persian All-words WSD.

3. Building the Sbu-Wsd-Corpus

To create a standard All-words WSD test set, we followed the suggestions made by the Senseval- 2 [9] competition. For the All-words task, the Senseval-2 guidelines suggest that (1) a standard test set should contain at least 5000 words of the running text, and (2) all context words should be tagged. The creation of SBU-WSD-Corpus can be thought of as a pipeline of four steps(i.e, Data Collection , Choosing sense inventory, Annotation process and Corpus Format), described in the following sections (Sections 3.1 to 3.4).The statistics of SBU-WSD-Corpus are presented in Section 3.5.

3.1. Data Collection

The documents selected for the SBU-WSD-Corpus are taken from in-house news corpora which include one million news documents crawled from different Iranian news websites. The news corpora contain documents from a variety of domains including sports, politics, science, culture, etc. The process of collecting documents for SBU-WSD-Corpus includes two steps:

We first extracted 100 documents from our news corpora and then computed the average ambiguity of the context words of each document. Second, in order to make the task more challenging, we chose19 documents with highest average ambiguity for construction of SBU-WSD-Corpus. As preprocessing step, we first manually tokenized, lemmatized, and PoS-tagged the documents to make them ready for the sense annotation phase

3.2. Sense Inventory

WordNet [21] is one of the most widely used lexical resources in many areas of NLP including WSD. It was originally designed for English at Princeton University. The basic components of WordNet are synsets, each expressing a unique concept by a set of words with the same meaning and PoS, a gloss (i.e, a brief definition of the synset words), and possibly an example (i.e, a usage example of synsets words). WordNet entries are represented by different synsets, denoting the different meanings they can take. For instance the word phone has four synsets in WordNet, denoting four possible meanings of phone in multiple contexts[1]. The current version of WordNet (WordNet 3.1) covers 155, 327 English words and phrases organized in 117, 979 synsets.

[1] In table, each synset are shown in the W#N#i or W#V#i format which correspond to the ith nominal or verbal ynsets of the target word W in the WordNet, respectively

WordNet synsets are interlinked via Lexical or Semantic relations which are held between pairs of word senses and synsets, respectively. WordNet can also be viewed as a semantic network in which nodes correspond to the synsets and edges to the lexical or semantic relations. Instances of Lexical and Semantic relations are shown in Table 1.

FarsNet: The Persian WordNet

Currently, WordNet is developed for many languages including Persian. The Persian WordNet, FarsNet [45], is the first lexical ontology for the Persian language which has been developed in the NLP lab of Shahid Beheshti University. The FarsNet project developed for more than 12 years. Over two past decades, a range of development have been done on FarsNet [38], [39], [15]. The current version of FarsNet (FarsNet 3.0) covers more than 100,000 Persian words and phrases and 40,000 synsets[2]. Similar to other WordNets, FarsNet groups words (nouns, verbs, adjectives, and adverbs) into synsets and connect them via different kinds of relations. FarsNet also provides a gloss and an example for each synset. FarsNet relations can be classified into two major groups: inner-language and inter-language relations.

The inner-language relations are defined between FarsNet senses and synsets while the inter-language relations align the FarsNet and WordNet synsets. The inner language relations of FarsNet include all WordNet 2.1 relations (i.e. hypernymy, hyponymy, holonymy, antonymy, etc.) as well as some extra relations such as agent-of, patient-of, salient, etc. Additionally, as FarsNet 3.0 is mapped to WordNet 3.0, the inter-language relations (equal-to and near-equal-to) are defined between FarsNet and WordNet 3.0 synsets. In this research, we used FarsNet as the sense inventory to annotate the context words of the documents.

3.3. Annotation Process

The whole SBU-WSD-Corpus is manually annotated by three Persian native speakers. All the annotators were familiar with FarsNet and WSD. To achieve a high-quality sense-annotated corpus, we followed the annotation procedure, suggested by [42]. The annotation process consists of two steps. In the first step, two taggers used SAMP (a tool for sense annotation with senses from FarsNet 3.0) to annotate 3 documents of the corpus. An expert linguist together with both annotators then discussed the annotations specifically the conflicting ones. Taggers then annotated the rested documents.

As the final phase, the expert linguist checked all annotations and re-annotated the words with different sense labels. The Inter-Annotator Agreement (IAA) and Cohen's kappa score obtained from the first step were 90.3 and 0.83 respectively.

3.4. Corpus Format

The main corpus is released in a standard XML format (from [35]). The annotated corpus includes two files: a single XML file in which all the documents are stored in, and a text file for mapping context words to sense numbers showing the annotations. A part of the corpus is shown in Figure 1. In the following, we describe the XML tags of the corpus.

- <corpus>: The tag indicates the beginning of the whole corpus.
- <text id>: The <text id> tag is representative of the beginning of a new document each specified with a unique identifier attribute (id).
- <sentence>:similar to <text id>, <sentence> tag shows the start of a particular sentence specified with a unique id attribute.
- < instance >: The tag represents context words with a relevant sense in FarsNet and specifies unique id, lemmas (Lemma), and PoS tags.
- <wf>: <wf> tag shows a context word with no corresponding sense in FarsNet, specified with a lemma (Lemma) and PoS tag.

To make the corpus independent to the sense repository, the mapping between the context words (instances) and their corresponding sense numbers is done in another file (text file). In the text file we have pairs of (instance id, FarsNet sense number) to make the annotation complete.

3.5. Corpus Statistics

SBU-WSD-Corpus consists of 19 documents obtained from in-house news corpora. The documents cover different domains including sports, religion, and culture. In Table 2 we show the general statistics of the dataset. For both test and tuning set, we report the number of words of running text together with the number of annotated words and ambiguity level per PoS. Following WSD literature, we computed ambiguity level of as total number sense candidates of words, divided by the number of annotated words. It worth noting that monosemous instances have been considered in the process. We also show the sense distribution of the test set words per PoS in Figure 2. As it can be seen, nouns are the most ambiguous part-of-speech followed by verbs, adjectives and adverbs which shows the least ambiguity in their meaning. In addition, more than 25 percent of nouns and 23 percent of verbs have more than 5 different meanings in FarsNet, indicating the task hardness on disambiguating nouns and verbs of the developed corpus. On the other hand, adjectives and adverbs seem easier to disambiguate, as most of them have only one or two senses in FarsNet.

As it can be seen in Figure 1 and Table 2, not all tokens of the corpus are annotated by sense numbers. The contexts words with are not annotated are shown by <wf> tag. These are either words out of noun, verb, adjective and adverb categories (ex. prepositions) or words with no corresponding senses in FarsNet. For the latter category if the word is a proper-noun and its hypernym exists in FarsNet, the lemma and the sense number will be set to the hypernym's lemma and sense number respectively.

4. Experimental Setup

In this section, we present several supervised and knowledge-based systems as baselines of Persian All-words WSD task. The systems are introduced in section 4.1, the evaluation measures are explained in section 4.3 and the results and analysis about the performance of the systems are shown in section 5.

[2] FarsNet web service is freely available at farsnet.nlp.sbu.ac.ir

Table 1. Example of some WordNet relations.

Relation name	*Relation Type*	*Definition*	*Example in WorNet*
Hypernymy	Semantic	A is hypernym of B, if every B is a kind of A	A Portable Computer is Hypernym ofLaptop
Hyponymy	Semantic	A is hyponym of B, if every A is a kind of B	Olive is Hyponym of Fruit
Antonymy	Lexical	A is Antonym of B, if A is polar opposite of B	Hot is Antonym of Cold

Table 2. General statistics of SBU-WSD-Corpus

		Test Set	*Tuning Set*	*All*
# Docs		13	3	16
# Tokens		5045	847	5892
Number of Instances per PoS	Nouns	1764	307	2071
	Verbs	494	70	564
	Adjectives	515	95	610
	Adverbs	111	11	122
Mean Sense per PoS	Nouns	4.0	3.9	4.0
	Verbs	3.4	2.9	3.3
	Adjectives	1.6	1.7	1.6
	Adverbs	1.2	1.3	1.2

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your journal for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

4.1. Comparison Ssytems

In this section, we briefly describe the All-words WSD systems used in our experiments. We include 10 systems (five supervised and five knowledge-based) in our empirical comparison.

a) FarsNet 1st sense

As mentioned in [45], FarsNet word senses are ranked by their usage frequency in Persian by expert linguists. We consider the FarsNet 1st sense approach as the baseline of knowledge-based systems. The approach is context-independent and always chooses the first sense of FarsNet as the most probable sense of each context word.

b) Lesk and Extended Lesk

Lesk is one of the most traditional WSD algorithms based on the overlap between the definition of senses and the context words. The algorithm counts the mutual words between the gloss a given sense and the context of the target word and chooses the sense with the highest count as the proper one. The pipeline of the proposed algorithm (named as Extended Lesk) is highly similar to the Lesk algorithm. The only difference is that Extended Lesk expands the definition of a given sense by including the definitions of its semantically related concepts from WordNet (e.g. hypernyms, hyponyms, etc).

c) Basile14

Extending two aforementioned variations of Lesk algorithms (Lesk and Extended Lesk) [4] developed an unsupervised language-independent WSD system. Instead of counting mutual words between context and sense glosses of the target word, the system uses distributional semantic space to compute the similarity between context and sense glosses. They also utilize sense frequency information from SemCor to give higher priority to most frequent senses.

d) UKB

Agirre and colleagues [2, 3] proposed a graph-based WSD system which applies PageRank algorithm over a semantic graph, constructed by WordNet. In the constructed graph, the nodes and edges are WordNet synsets and relations, respectively. The algorithm assigns a PageRank value to the nodes and chose the node with the highest value as the best meaning of each target word. Two main variants of the algorithm are ppr and pprw2w. The first approach performs random walk on a graph personalized on the word context and disambiguates all the context words in one go. However, the latter performs the disambiguation process for each word separately.

Our comparison also includes best supervised systems, reported in [49] which utilized Persian SemCor as training set. Rouhizadeh and colleagues [49] employed four machine learning algorithms, i.e. Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Decision Tree (DT), and Multilayer Perceptron (MLP) to train supervised Persian WSD models on Persian SemCor. All the systems make use of word embedding models as feature vector. Following this reference we consider MFS as the baseline of supervised approaches. For each target word, the approach selects the most occurring sense in Persian SemCor as the most probable one.

4.2. Parameters Settings

To carry out the experiments in a fair setup, we first optimized the parameters of the systems on the tuning set of SBU-WSD-Corpus.

Among Knowledge-based systems, the pipeline of both Extended Lesk and Basile14 WSD systems, only include one parameter to tune, i.e. context size. We used the available implementations to evaluate both systems with context sizes 3, 5, 10, 20, and the whole text. Interestingly, for both systems, the best results (reported in Table 3) obtained by context size 34. As mentioned in [2], UKB include no parameter to tune. To evaluate the system, we used the last implementation, provided by the authors of the original paper. For supervised systems, we report the best results reported in [49].

All supervised systems and also Basil14 system make use of a word embedding model to represent the target texts in a semantic space.

As an unsupervised machine learning technique, word embedding models (e.g, word2vec [20], Glove [31], BERT [8]) make use of large collections of unlabeled data to specify

```
<?xml version="1.0" encoding="UTF-8">
<corpus lang="fa" CorpusName="SBU-WSD-Corpus">
<text id="d000">
<sentence id="d000.s000">
<instance id="d000.s000.t001" lemma="سرپرست" pos="Noun">سرپرست</instance>
<instance id="d000.s000.t002" lemma="اداره" pos="Noun">اداره</instance>
<instance id="d000.s000.t003" lemma="کل" pos="ADJ">کل</instance>
<instance id="d000.s000.t004" lemma="هواشناسی" pos="Noun">هواشناسی</instance>
<instance id="d000.s000.t005" lemma="خوزستان" pos="Noun">خوزستان</instance>
<wf lemma="از" pos="P">از</wf>
<instance id="d000.s000.t006" lemma="افزایش" pos="Noun">افزایش</instance>
<instance id="d000.s000.t007" lemma="دما" pos="Noun">دمای</instance>
<instance id="d000.s000.t008" lemma="هوا" pos="Noun">هوا</instance>
<wf lemma="به" pos="P">به</wf>
<instance id="d000.s000.t009" lemma="بیشتر" pos="ADJ">بیشتر</instance>
<wf lemma="از" pos="P">از</wf>
<instance id="d000.s000.t010" lemma="پنجاه" pos="ADJ">پنجاه</instance>
<instance id="d000.s000.t011" lemma="درجه" pos="Noun">درجه</instance>
<wf lemma="در" pos="P">در</wf>
<instance id="d000.s000.t012" lemma="ساعت" pos="Noun">ساعات</instance>
<instance id="d000.s000.t013" lemma="آینده" pos="ADJ">آینده</instance>
<instance id="d000.s000.t014" lemma="خبردادن" pos="Verb">خبرداد</instance>
<wf lemma="." pos=".">.</wf>
</sentence>
```

Figure. 1. An example sentence drawn from SBU-WSD-Corpus

Table 3. F-1 performance of different supervised and knowledge-based models on SBU-WSD-Corpus [49].

Approach	*System*	*Noun*	*Verb*	*Adjective*	*Adverb*	*All*
Supervised Sytems	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
Knowledge Based Systems	SVM	65.0	65.0	90.0	90.1	72.7
	FN 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

similar n-dimensional vectors to the semantically similar words. In order the systems with Persian, we used Gensim software package [37] to train a 300-dimensional word2vec model [20] on our in-house Persian news corpora.

4.3. Evaluation Measure

As mentioned in [27] the performance of WSD systems can be evaluated by four standard metrics, described in the following:

a) Coverage

The coverage (C) of a WSD system is defined as the number of sense assignments provided by the system over the number of words in the test corpus (Equ.(1)).

$$C = \frac{\#sense\ assignments}{\#total\ instances\ of\ the\ testset} \quad (1)$$

b) Precision

The precision (P) is defined as the number of correctly disambiguated words over the total number of disambiguated words returned by the system as Equ.(2) shows.

$$P = \frac{\#correctly\ disambiguated\ words}{\#disambiguated\ words} \quad (2)$$

c) Recall

The recall (R) of a WSD system is the number of. the correct answers provided by the system divided by the number of expected answers (Equ.(3)).

$$R = \frac{\#correctly\ disambiguated\ words}{\#total\ instances\ of\ the\ test\ set} \quad (3)$$

d) F-measure

F-measure is defined as harmonic mean of P and R and is computed as Equ.(4) shows.

$$F = \frac{2*P*R}{P+R} \qquad (4)$$

Note that F-measrue = R = P, when a system provides an answer for each word in the test set. Following Senseval-2 guidelines, we evaluate the performance of the systems with F-measure.

5. Results And Analysis

Table 2 shows the F-Measure performance of all comparison systems on the SBU-WSD-Corpus dataset. We additionally, report the performance of each system, divided by PoS tags. As it can be seen, supervised systems, trained on Persian SemCor consistently outperform knowledge-based systems across the dataset. It clearly shows the high ability of Persian SemCor on training WSD models for Persian. It is also interesting to note the performance of the MFS approach, which is considered as the baseline of supervised systems, can achieve competitive results with the best performing knowledge-based systems. One of the main conclusions that can be taken from the evaluation is the positive effect of word embedding models in disambiguating Persian words. As discussed in section4.2, all the supervised models and also Basile14 utilize word embedding models in their disambiguation pipelines. We provide a detailed analysis of the performance of Basile14, as the best performing knowledge-based model, to clearly show the effect of the word embedding model in its default pipeline.

As mentioned in section 4.1, the disambiguation pipeline of Extended Lesk and Basile14 systems are highly similar. A comparison between the results obtained by these systems indicates that the use of word embedding can have a significant impact on the performance of the system. As it can be seen in Table 3, the performance of Basile14 improved by a large margin (12 percent), compared to the Extended Lesk. As discussed in section 4.1, the pipeline of Basile14 includes two key components: (1) Word Embedding model and (2) gloss definitions of the sense inventory, both are available for Persian. Existing mutual words in the gloss of different senses which result in similarity in their semantic vectors can be mentioned as the most important bottleneck of the system. To deal with this, the system expands the gloss of each sense by including the glosses of the semantically related concepts (i.e. the concepts which have a direct relation to the synset) (see more details on section 4.1).

We also reported the performance of the systems, divided by PoS tags. As it can be seen from Table 3 the performance of most systems on disambiguating nouns is lower than other PoS tags. It can be explained by the ambiguity level of different PoS tags, shown in Table 2. As shown in Table 2, the average ambiguity of the present nouns in SBU-WSD-Corpus is 4.0 which is greater than all the other PoS tags. Additionally, in Figure 2, we showed that more than 25 percent of nouns have more than 6 senses, indicating the difficulty of noun disambiguation in the developed data set. On the other hand, adjectives and adverbs seem easier to disambiguate, as their ambiguity level is 1.6 and 1.2 respectively.

6. CONCLUSION

In this paper, we presented a standard evaluation corpus for Persian All-words WSD. The corpus contains 19 Persian documents, manually tokenized, lemmatized, PoS-tagged, and

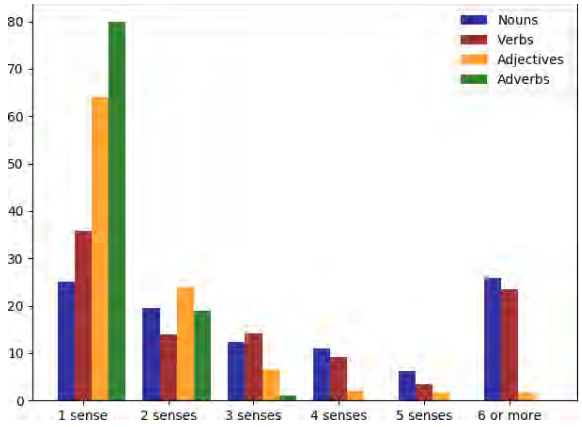

Figure. 2. Sense distribution of the annotated words of the SBU-WSD-Corpus, divided by PoS tags.

senses-tagged. It contains 5892 words of running text and covers different domains including Economics, Sports, etc.

Additionally, we applied several supervised and knowledge-based WSD systems on the corpus. The results show that the supervised systems can outperform the knowledge-based alternatives. We evaluated several benchmark All-words WSD models on SBU-WSD-Corpus, providing baselines for future improvements on the Persian All-words WSD task. In addition, to encourage future research on Persian All-words WSD, we have made SBU-WSD-Corpus freely available. A possible extension to this work will include applying other knowledge-based WSD methods which are applicable to low-resource languages.

Declarations

Funding

This research did not receive any grant from funding agencies in the public, commercial, or non-profit sectors.

Authors' contributions

HR: Study design, acquisition of data, statistical analysis, drafting the manuscript; revision of the manuscript MS: Supervision, design, interpretation of the results, drafting the manuscript, revision of the manuscript; VT: acquisition of data, drafting the manuscript.

Conflict of interest

The authors declare that there is no conflict of interest.

References

[1] E Agirre, I Aldezabal, J Etxeberria, E Izagirre, K Mendizabal, E Pociello, and M Quintian. 2005. Eusem-cor: euskarako corpusa semantikoki etiketatzeko eskuliburua; editatze-, etiketatze-eta epaitze-lanak. Technical report, Internal report.

[2] Eneko Agirre, Oier Lo´pez 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), pages 29–33, Melbourne, Australia. Association for Computational Linguistics.

[3] Eneko Agirre and Aitor Soroa. 2009. Personalizing pagerank for word sense disambiguation. In Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), pages 33–41.

[4] 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.

[5] 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.

[6] Rebecca Bruce and Janyce Wiebe. 1994. Word-sense disambiguation using decomposable models. In Proceedings of the 32nd annual meeting on Association for Computational Linguistics, pages 139–146. Association for Computational Linguistics.

[7] Marine Carpuat and Dekai Wu. 2007. Improving statistical machine translation using word sense disambiguation. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL).

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805.

[9] Philip Edmonds and Scott Cotton. 2001. Senseval-2: overview. In The Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems, pages 1–5. Association for Computational Linguistics.

[10] Andreas Eisele and Yu Chen. 2010. Multiun: A multilingual corpus from united nation documents. In LREC.

[11] W Nelson Francis and Henry Kucera. 1979. Brown corpus manual. Letters to the Editor, 5(2):7.

[12] Masood Ghayoomi, Identifying Persian words' senses automatically by utilizing the word embedding method," Iranian Journal of Information Processing & Management, vol. 35, no. 1, pp. 25–50, 2019.

[13] Masood Ghayoomi, Word Sense Induction in Persian and English: A Comparative Study, Journal of Information Systems and Telecommunication (JIST), 2021, Vol 9(36), pp. 263-274

[14] Rube´n Izquierdo-Bevia´, Lorenza Moreno-Monteagudo, Borja Navarro, and Armando Sua´rez. 2006. Spanish all-words semantic class disambiguation using cast3lb corpus. In Mexican International Conference on Artificial Intelligence, pages 879–888. Springer.

[15] 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.

[16] Svetla Koeva, Svetlozara Leseva, Ekaterina Tarpomanova, Borislav Rizov, Tsvetana Dimitrova, and Hristina Kukova. 2010. Bulgarian sense-annotated corpus–results and achievements. FASSBL7, page 41.

[17] M. Mahmoodvand and M. Hourali, "Semi-supervised approach for Persian word sense disambiguation," *2017 7th International Conference on Computer and Knowledge Engineering (ICCKE)*, 2017, pp. 104-110.

[18] Babak Masoudi, Saeed Rahati Ghouchani, 2016. A LDA topic Model for Farsi Word Sense Disambiguation, Signal and Data Processing, 12(4), 117-125.

[19] Raheleh Makki, Mohammad Mahdi Homayounpour, 2008. Word Sense Disambiguation of Farsi Homographs Using Thesaurus and Corpus. In: Nordström, B., Ranta, A. (eds) Advances in Natural Language Processing. GoTAL 2008. Lecture Notes in Computer Science(), vol 5221. Springer, Berlin, Heidelberg.

[20] 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.

[21] 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.

[22] George A Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G Thomas. 1994. Using a semantic concordance for sense identification. In Proceedings of the workshop on Human Language Technology, pages 240–243. Association for Computational Linguistics.

[23] George A Miller, Claudia Leacock, Randee Tengi, and Ross T Bunker. 1993. A semantic concordance. In Proceedings of the workshop on Human Language Technology, pages 303–308. Association for Computational Linguistics.

[24] Simonetta Montemagni, Francesco Barsotti, Marco Battista, Nicoletta Calzolari, Ornella Corazzari, Alessandro Lenci, Antonio Zampolli, Francesca Fanciulli, Maria Massetani, Remo Raffaelli, et al. 2003. Building the italian syntactic-semantic treebank. In Treebanks, pages 189–210. Springer.

[25] Moradi, E. Ansari and Z. Žabokrtský, "Unsupervised Word Sense Disambiguation Using Word Embeddings," *2019 25th Conference of Open Innovations Association (FRUCT)*, 2019, pp. 228-233,

[26] Andrea Moro and Roberto Navigli. 2015. Semeval- 2015 task 13: Multilingual all-words sense disambiguation and entity linking. In Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015), pages 288–297.

[27] Roberto Navigli. 2009. Word sense disambiguation: A survey. ACM computing surveys (CSUR), 41(2):1– 69.

[28] Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. Semeval-2013 task 12: Multilingual word sense disambiguation. In Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), pages 222–231.

[29] Hwee Tou Ng, Chung Yong Lim, and Shou King Foo. 1999. A case study on inter-annotator agreement for word sense disambiguation. In SIGLEX99: Standardizing Lexical Resources.

[30] Dieke Oele and Gertjan Van Noord. 2017. Distributional lesk: Effective knowledge-based word sense disambiguation. In IWCS 2017—12th International Conference on Computational Semantics—Short papers.

[31] 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.

[32] Alexander Popov, Kiril Simov, and Petya Osenova. 2019. Know your graph. state-of-the-art knowledge-based wsd. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), pages 949–958.

[33] Sameer Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. Semeval-2007 task-17: English lexical sample, srl and all words. In Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007), pages 87–92.

[34] 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.

[35] 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.

[36] Ganesh Ramakrishnan, Apurva Jadhav, Ashutosh Joshi, Soumen Chakrabarti, and Pushpak Bhattacharyya. 2003. Question answering via bayesian inference on lexical relations. In Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering-Volume 12, pages 1–10. Association for Computational Linguistics.

[37] Radim Rehurek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pages 45– 50, Valletta, Malta. ELRA. http://is.muni.cz/ publication/884893/en.

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

[39] 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.

[40] Hossein Rouhizadeh, Mehrnoush Shamsfard and Masoud Rouhizadeh, "Knowledge Based Word Sense Disambiguation with Distributional Semantic Expansion for the Persian Language," *2020 10th International Conference on Computer and Knowledge Engineering (ICCKE)*, 2020, pp. 329-335

[41] Adriana Roventini, Alone Antonietta, Francesca Bertagna, Nicoletta Calzolari, Cacila Jessica, Christian Girardi, Bernardo Magnini, R Marinelli, Manuela Speranza, and A Zampolli. 2003. Italwordnet: building a large semantic database for the automatic treatment of italian.

[42] Ali Saeed, Rao Muhammad Adeel Nawab, Mark Stevenson, and Paul Rayson. 2019a. A sense annotated corpus for all-words urdu word sense

disambiguation. ACM Transactions on Asian and LowResource Language Information Processing (TALLIP), 18(4):1–14.

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

[44] Mehrnoush Shamsfard. 2011. Challenges and open problems in persian text processing. Proceedings of LTC, 11.

[45] 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.

[46] Mehdi Soltani, Heshaam Faili, (2010), A statistical approach on Persian word sense disambiguation, The 7th International Conference on Informatics and Systems (INFOS), pp.1–6.

[47] Benjamin Snyder and Martha Palmer. 2004. The english all-words task. In Proceedings of SENSEVAL3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, pages 41–43.

[48] 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

[49] Hossein Rouhizadeh, Mehrnoush Shamsfard, Mahdi Dehghan, and Masoud Rouhizadeh. 2021. Persian SemCor: A bag of word sense annotated corpus for the Persian language. In Proceedings of the 11th Global Wordnet Conference, pages 147–156, University of South Africa (UNISA). Global Wordnet Association.

[50] Saba Urooj, Sana Shams, Sarmad Hussain, and Farah Adeeba. 2014. Sense tagged cle urdu digest corpus. Centre for Language Engineering, Al-Khawarizmi Institute of Compute Science, University of Engineering and Technology, Lahore.

[51] 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., 1989.

Hossein Rouhizadeh is Ph.D. student at the University of Geneva. His research interests include using machine learning and natural language processing methods to analyse texts in the biomedical domain.

Dr. Mehrnoush Shamsfard has been with Shahid Beheshti University from 2004. She is currently associate professor of Faculty of computer science and engineering, and also the head of NLP research Laboratory of this faculty. Her main fields of interest are natural language processing, knowledge and ontology engineering, text mining and semantic and intelligent web.

Dr. Vahide Tajalli is a Ph.D. graduate in linguistics from University of Tehran with eight years of experience in computational linguistics. She is cooperating with the NLP research lab of Shahid Beheshti University.