A Synset Relation-enhanced Framework with a Try-again Mechanism for Word Sense Disambiguation

Ming Wang[1] and **Yinglin Wang**[2, *]
School of Information Management and Engineering
Shanghai University of Finance and Economics, Shanghai, China
[1]wangming@163.sufe.edu.cn, [2]wang.yinglin@shufe.edu.cn

Abstract

Contextual embeddings are proved to be overwhelmingly effective to the task of Word Sense Disambiguation (WSD) compared with other sense representation techniques. However, these embeddings fail to embed sense knowledge in semantic networks. In this paper, we propose a Synset Relation-Enhanced Framework (SREF) that leverages sense relations for both sense embedding enhancement and a try-again mechanism that implements WSD again, after obtaining basic sense embeddings from augmented WordNet glosses. Experiments on all-words and lexical sample datasets show that the proposed system achieves new state-of-the-art results, defeating previous knowledge-based systems by at least 5.5 F1 measure. When the system utilizes sense embeddings learned from SemCor, it outperforms all previous supervised systems with only 20% SemCor data.

1 Introduction

Word Sense Disambiguation (WSD) is an ongoing research area in Natural Language Processing community. It is aimed at determining the correct meaning (sense) of a word in its context given a list of potential or competing senses in a sense inventory. According to Navigli (2009), most of WSD solutions can be categorized into supervised and knowledge-based approaches. For supervised systems, they rely on sense-annotated data to train either word experts (Zhong and Ng, 2010) or a neural language model (Raganato et al., 2017a) for disambiguation and thus perform better than their knowledge-based counterparts (Banerjee and Pedersen, 2002; Basile et al., 2014; Agirre et al., 2014), which merely utilize sense knowledge in a sense inventory. However, knowledge-based approaches can better scale to a multilingual scenario or a specific domain where sense annotation is limited.

Contextual representations learned from neural language models (Peters et al., 2018) are proved to be beneficial to the task of WSD. Many recent systems (Loureiro and Jorge, 2019; Vial et al., 2019; Scarlini et al., 2020) utilize language models, especially BERT (Devlin et al., 2019), as a feature extraction tool to obtain contextual sense representations and outperform previous approaches by large margins. There are also systems (Luo et al., 2018; Kumar et al., 2019) that incorporate sense definitions into language models and achieve state-of-the-art performance. However, most of the systems are implemented in a supervised manner using a widely exploited sense-annotated corpus, SemCor (Miller et al., 1994), and merging knowledge from the sense inventory as a supplement. There is much space to explore regarding how to better exploit knowledge in a sense inventory such as different WordNet relations and super-sense that categorizes WordNet senses into 45 clusters.

In this paper, we present SREF, a knowledge-enhanced WSD approach that effectively exploits the sense definitions and relations in an inventory. First, we design a gloss augmentation for those synsets that have a short definition in WordNet so that each synset can learn a reliable sense embedding with features from BERT. Then, based on these embeddings, we explore the contribution of different synset relations in WordNet (Miller, 1995) to learn relation-enhanced sense embeddings. After the first WSD is conducted with a nearest neighbor approach against an

* corresponding author

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 6229–6240,
November 16–20, 2020.

ambiguous word's context embedding and the relation-enhanced embeddings of the word's potential senses, we implement a try-again mechanism to the top 2 competing senses using synset relations and super-sense category. When applying the proposed strategy to tackle WSD, our system achieves state-of-the-art performance among knowledge-based systems. When we concatenate our sense embeddings with those learned from SemCor, new state-of-the-art performance in supervised category is achieved. We thus summarize our contributions as follows:

(1) We propose a fine-grained utilization of short WordNet sense glosses to retrieve web mentions to supplement sense embedding learning, and a method to create sense embeddings in a bag-of-sense manner by utilizing WordNet sense relations.

(2) We design a try-again mechanism that employs both synset relations and super-sense connections. To the best of our knowledge, this is the first attempt on employing WordNet relations to implement WSD again with sense relation knowledge.

(3) State-of-the-art performance is achieved in both all-words and lexical sample WSD datasets, surpassing previous systems by 5.5 F1 measure in knowledge-based all-words WSD. The supervised version of our system achieves state-of-the-art performance with only 20% SemCor data. The source code is available at: github.com/lwmlyy/SREF.

2 Related Work

In order to tackle WSD, approaches in two streams have been well developed over the last few decades, namely supervised and knowledge-based approaches. Their major difference is whether a sense-annotated corpus is employed.

2.1 Supervised Systems

Supervised systems originally regard WSD as a sense classification problem, building one classifier for each target word. IMS (Zhong and Ng, 2010), among others (Tsatsaronis et al., 2007, Iacobacci et al., 2016, Papandrea et al., 2017), is the most widespread system that leverages SVM to classify senses. In recent years, a more efficient supervised scheme has been proposed. Rather than training a few classifiers, it constructs a single neural architecture (Raganato et al., 2017a) with an annotated corpus and disambiguates words based on the output of the last layer. These methods have not outperformed traditional counterparts until sense definitions were incorporated (Luo et al., 2018). It has also become a trend that newly proposed systems (Kumar et al., 2019; Huang et al., 2019; Loureiro and Jorge, 2019; Vial et al., 2019; Scarlini et al., 2020) tend to exploit WordNet sense knowledge one way or another.

Despite the employment of sense knowledge, many systems still require a Most Frequent Sense (MFS) fallback since SemCor only covers a small proportion of WordNet lemmas. To address this issue, LMMS (Loureiro and Jorge, 2019) takes into account the synset and hypernymy relation in WordNet to extend sense embeddings to full coverage, utilizing BERT to contextualize the annotated senses in SemCor as a starting point. This approach achieves an unprecedented improvement in WSD tasks, although the synset relations are not adequately explored.

The recent development in contextual embeddings has injected much power into supervised WSD systems. Many of them rely on WordNet gloss to embed contextual information regarding a particular sense. However, a simple fact seems to be overlooked that many synset glosses are excessively short to deliver sufficient information. We thus propose a gloss augmentation method to relieve this issue. This is different from the previous gloss expanding methods (Ponzetto and Navigli, 2010; Miller et al., 2012), which expand glosses with either separate words or Wikipedia documents, rather than selected short sentences.

2.2 Knowledge-based Systems

Knowledge-based systems typically design some algorithms with which to operate on the semantic networks for disambiguation. One major branch is to consider the similarity between potential senses and the ambiguous word, including Lesk (Lesk, 1986) and other following researches (Banerjee and Pedersen, 2002; Basile et al., 2014). Another branch is to run graph algorithms (Agirre et al., 2014, Moro et al., 2014) on the semantic network and disambiguate based on sense connections in the network. There are also studies (McCarthy et al., 2007; Bhingardive et al., 2015) that focus on exploring how to learn or manipulate MFS given the fact that MFS is a highly competitive strategy.

Seeking language transferability, many knowledge-based methods (Basile et al., 2014; Camacho-Collados et al., 2016) rely on multilingual resources such as Wikipedia and BabelNet. A recent work (Scarlini, et al., 2020) follows the same idea by using BERT to learn contextual sense representations from retrieved mentions in both resources. Its supervised version is capable of beating many latest systems in noun disambiguation. However, both knowledge resources are constructed from a perspective of noun or entity relation, limiting the system's capability of disambiguating words in other part-of-speech (POS). In this paper, we augment the synset gloss (regardless of synset POS) of short length with retrieved mentions from the web so that the contextual representations can be more comprehensive for senses in all POS.

Although many previous similarity-based methods have explored the value of synset relations in WordNet, most of them utilize related synsets in a bag-of-word manner. For example, in enhanced Lesk (Basile et al., 2014), gloss words of related synsets are first merged into the gloss word set of a potential sense. Using the word set, a sense embedding is learned by summing all its word embeddings. This approach naturally neglects the word order in a sense gloss and weakens the difference between senses. In our approach, the sense embedding learning process is implemented in a bag-of-sense perspective so the weaknesses are relieved. Also, we propose a novel relation exploitation scheme to disambiguate again with not only the potential sense itself but also its related senses in WordNet. This is distinct from the methods in previous researches where relations are exploited to compress or cluster senses into coarse-grained senses (Miller and Iryna, 2015; Vial et al., 2019).

3 Preliminaries

In this section, we introduce WordNet and BERT, the contextual representation learning model.

3.1 WordNet

WordNet is a commonly used sense inventory for English WSD and it covers 117,659 synsets and 206,978 senses in its 3.0 version. A synset contains a set of senses that share the same meaning. For each synset, a definition (gloss) is provided to show what it means, or in some cases, to explain the synset less ambiguously. For example, *intend.v.01* (*intend* as its lemma), *mean.v.04* and *think.v.07* convey an identical meaning of *have in mind as a purpose* while *think.v.05* is defined as *imagine or visualize*. Also, many synsets are contextualized with one or more example sentences, e.g. *I mean no harm* for *mean.v.04*.

	N	V	A	R
gloss length	11.5	6.2	7.2	5.0
ambiguity	1.4	2.6	1.6	1.3

Table 1: Wordnet Synset Gloss Length (Number of Gloss Words per Synset) and Lemma Ambiguity (Number of Synsets per Lemma) in Different POS.

The synsets are organized into four groups according to their POS, namely noun (N), verb (V), adjective (A) and adverb (R). Synsets in each POS are connected by different relations separately in most cases. There are over 15 relations for synsets but many of them are defined for synsets in a particular POS. For instance, hypernymy and hyponymy relations are only available for nouns and verbs while entailment relation is valid for verbs alone. There is also a cross-POS relation in WordNet, defined as 'derivationally related form'. As an example, *intend.v.01* and *intention.n.03* are derivationally related.

WordNet defines a coarse-grained sense category named super-sense, which arranges senses into 45 clusters including *noun.person*, *noun.artifact* and others, 26 of which are for nouns, 15 for verbs, 3 for adjectives and 1 for adverbs. Senses in the same category have a weak connection to each other.

Despite the notable contribution of synset gloss to many WSD systems, synset relations are more valuable since they provide possibilities that machines could recognize synset connections. Here, we utilize WordNet relations for sense embedding enhancement (section 4.2) and a try-again mechanism (section 4.3).

3.2 BERT Utilization

BERT, a transformer-based language model, has attracted much attention from researchers of many NLP applications. In our research, we utilize BERT as a feature extraction model to learn a sense embedding for each WordNet sense using its gloss.

However, directly using synset gloss to learn a

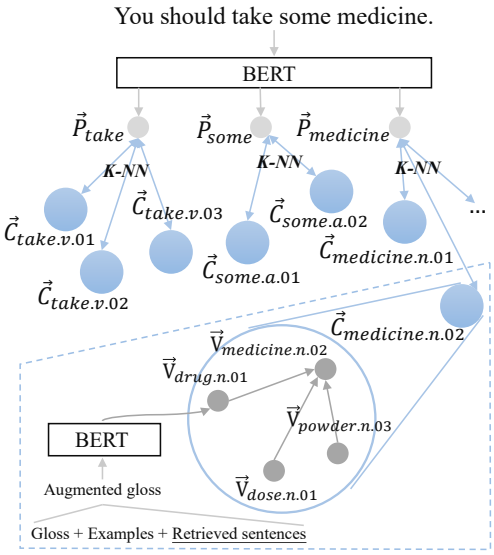

Figure 1: Knowledge-enhanced WSD Framework without the Try-again Mechanism

sense embedding is problematic since many synset glosses contain insufficient context for representation learning. Among others, the gloss for *think.v.05* is *imagine or visualize*, which is too short to carry adequate information. Table 1 presents the average synset gloss length and ambiguity of lemmas in four POS. It shows a relatively short gloss length for verb, adjective, and adverb synsets.

To address the above issue, we propose a gloss augmentation method (section 4.1) to bring in more context information regarding those poorly contextualized synsets.

In our final proposal, for each sense, we use BERT to learn its basic sense embedding from the concatenation of its gloss (and lemmas), example sentences and retrieved sentences from the web. In detail, we use $BERT_{LARGE\_CASED}$ as our feature extraction model and sum the output of the last 4 layers (a typical setting in previous researches such as LMMS, Loureiro and Jorge, 2019) at all output positions.

4 Method

Figure 1 demonstrates the overall concept of the framework without the try-again mechanism using an example. It relies on a K-NN algorithm to predict the correct sense of each word under disambiguation. The algorithm is implemented against a context representation ($\vec{P}$, lighter grey circle) directly from BERT at the position of the word under disambiguation and a knowledge-enhanced representation ($\vec{C}$, smaller blue circle) from BERT and WordNet knowledge. The big blue circle briefly illustrates how related senses are merged into one specific sense (section 4.2). In this big circle, the grey circles are basic sense embeddings ($\vec{V}$, grey circle) learned from the synset's augmented gloss (section 4.1) via BERT.

synset	gloss	queries
crash.v.05	break violently or noisily; smash	break violently, break noisily, smash
force.n.01	a powerful effect or influence	a powerful effect, a powerful influence

Table 2: Query Examples for Some Glosses.

4.1 WordNet Gloss Augmentation

In order to relieve the under-contextualization issue of many synsets, we propose a gloss augmentation approach to draw in more contextual information. Precisely, we simply use the short-length glosses as queries (words or phrases) to retrieve sequences from the web and combine the sequences with the original gloss and example sentences to learn a contextual representation from BERT. The whole process is built upon two hypotheses as follows.

(1) The words in the linguistic explanation of a synset tend to be less ambiguous and are often skewed to MFS/WordNet 1st sense. This is supported by the fact that more than 75% of the WordNet gloss words are labeled as MFS in the Princeton WordNet Gloss Corpus (Mihalcea and Moldovan, 2001).

(2) Word phrases in a synset gloss are even less ambiguous. Also, we calculate the proportion of polysemous phrase lemma in all phrase lemmas in WordNet. It shows a small proportion of those ambiguous phrase lemmas, 13.9% (4,922 out of 46,470).

Inspired by the above two hypotheses, we design a gloss augmentation method to retrieve sequences that contain gloss mentions. This is only operated on those synsets whose gloss has less than 6 words, which are easier to apply rules on. We detail the procedures as follows:

(1) For synsets whose gloss length is smaller than 6 words, cut each gloss or compose gloss words into one or more phrases under heuristic rules (split the gloss sentence with ';' into spans; segment each span based on the location of 'or'), see Table 2 for some

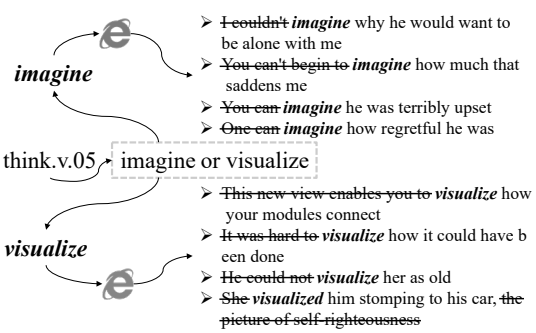

Figure 2: Gloss Augmentation for *think.v.05*

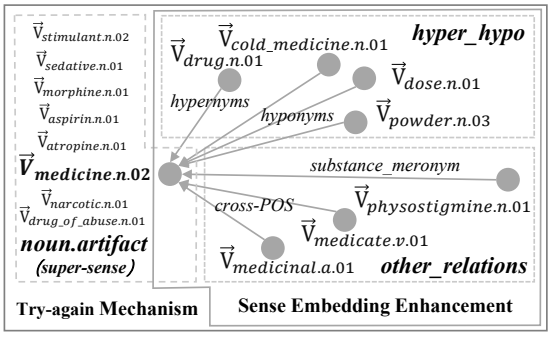

Figure 3: Synset Relation Exploitation for Sense Embedding Enhancement and Try-again Mechanism.

examples; For each query, retrieve sentences from *Baidu* translation website in its bilingual example section when each sentence contains the exact query;

(2) Filter out those sentences where query's POS is not the same as the synset's if the query is a word; extract the sub-sentence which includes the query but filters out the words before the query to reduce noise; Filter out those sentences that occur in more than one retrieved sentence sets of competing synsets (e.g. *think.v.01*, *think.v.02*) of a lemma to avoid overlap.

After the sequences (cf. Figure 2) are obtained, we combine them with each corresponding synset's gloss to learn a basic contextual representation.

4.2 Sense Embedding Enhancement

In this section, we introduce how to exploit WordNet relations for learning relation-enhanced sense embeddings. After each basic sense embedding is learned from its augmented gloss via BERT, it is further enhanced with a weighted sum of all its directly connected senses' basic sense embeddings. Here, we use all the relations except *verb_group* because this relation connects competing senses in many cases, weakening the difference between each other. The right proportion of Figure 3 reveals the process of sense embedding enhancement for *medicine.n.02*.

The relations are categorized into two classes named *hyper_hypo* (hypernymy and hyponymy) and *other_relations*. This is because the former class covers most of the connections in WordNet. We experiment on how the utilization of these two classes of relations benefit the task of WSD later.

Formula (1) details the sense embedding enhancement. Given all basic sense embeddings ($\vec{V}$), we enhance the embedding of sense s with the basic sense embedding ($\vec{V}_k$) of all its directly connected senses (R_s, including sense s) obtained with different WordNet relations. $d_{s,k}$ is the shortest path distance between sense s and sense k.

$$\vec{C}_s = \sum_k^{R_s} \frac{1}{1+d_{s,k}} * \vec{V}_k \quad (1)$$

Given the above enhanced sense embeddings, we calculate the similarity (dot product) between an ambiguous word's context embedding $\vec{P}_w$ and the potential senses' enhanced embeddings $\vec{C}_s$ after normalization. The disambiguation at the first attempt (1st WSD) is completed by selecting the potential sense with the highest similarity. The lemma and POS are utilized when retrieving the potential senses from WordNet.

Algorithm 1: Try-again Mechanism

Input: context embedding $\vec{P}_w$ of an ambiguous word w and enhanced sense embedding $\vec{C}_{s_{w,i}}$ of w's ranked i^{th} potential senses $s_{w,i}$ (i = 1, 2)

Output: $sim(w, s_{w,i})$ (i = 1, 2)

```
for i = 1 to 2 do
    R_{s_{w,i}} = [], sim = [];
    for relation in WordNet.all_relations do
        R_{s_{w,i}}.extend(s_{w,i}.relation);
    if super-sense(s_{w,1}) != super-sense(s_{w,2}) then
        R_{s_{w,i}}.extend(super-sense(s_{w,i}).synsets());
    for s in R_{s_{w,i}} do
        sim.append(C_s · P_w);
    sim(w, s_{w,i}) = C_{s_{w,i}} · P_w + max(sim);
return sim(w, s_{w,i}) (i = 1, 2)
```

4.3 Try-again Mechanism

In this section, we introduce the try-again mechanism against the first and second most

similar potential senses for every ambiguous word. This is based on the observation from the experimental result of the 1st WSD. It shows that after ranking potential senses according to the calculated similarity, 71.8% of the correct senses are ranked 1st, which represents the F1 score of the 1st WSD. Furthermore, 16% of the correct senses are ranked 2nd, which means our system's top 2 performance is 87.8%. This becomes a trigger to our experiment on whether synsets from different relations or the super-sense connection can benefit a 2nd WSD merely against the top 2 potential senses.

Algorithm 1 illustrates the detailed try-again mechanism, where both the 1st and 2nd WSD similarities are employed to select the final predicted sense. Precisely, for ambiguous word w ($\vec{P}_w$ as its contextual embedding), $\vec{C}_{s_{w,i}}$ is the enhanced sense embedding for one of its potential sense $s_{w,i}$. $R_{s_{w,i}}$ is all the directly connected senses from different WordNet relations except *verb_group*. In particular, if the top 2 potential senses belong to different super-sense categories, $R_{s_{w,i}}$ also contains all the senses that belong to the same super-sense as the potential sense. For instance, *medicine.n.01* belongs to *noun.cognition* while *medicine.n.02* is in *noun.artifact* category. In other words, the final WSD approach utilizes both the sense embedding of the potential sense itself and those of its related senses from WordNet relations and the super-sense category.

5 Experiment Setup

In this section, we evaluate our system using the evaluation framework provided by Raganato et al. (2017b). This framework includes five standard all-words WSD datasets: SensEval-2 (SE2, Palmer et al., 2001), SensEval-3 (SE3, Snyder and Palmer, 2004), SemEval-2007 (SE07, Pradhan et al., 2007), SemEval-2013 (SE13, Navigli et al., 2013) and SemEval-2015 (SE15, Moro and Navigli, 2015). We also show how our system performs on lexical sample datasets including SensEval-2 (SE2-LS, Kilgarriff, 2001) and SensEval-3 (SE3-LS, Mihalcea et al., 2004). We use the preprocessed datasets from UFSAC (Vial et al., 2018).

5.1 SREF

We have implemented both knowledge-based and supervised version of our system.

	ALL	N	V	A	R
SREF$_{kb}$	**73.5**	**78.5**	**56.6**	**79**	76.9
-w/o second_wsd	71.8 (-1.7)	77	54.4	77.2	76.3
-w/o gloss_augment	72.5 (-1.0)	77.7	55.4	76.8	77.7
-w/o other_relations	72.5 (-1.0)	77.5	**56.6**	75.2	**78.3**
-w/o hyper_hypo	70.7 (-2.8)	74.8	54.2	78.8	77.2

Table 3: Ablation Study on ALL (F1-%)

SREF$_{kb}$: the augmented gloss is utilized to learn a basic sense embedding from BERT by summing its last 4 layers at all output positions. Then synset relations are used to enhance each basic sense embedding. Finally, a nearest neighbor method is implemented against every ambiguous word's context embedding to its potential senses' enhanced embeddings before the try-again mechanism.

SREF$_{sup}$: Semcor is exploited to learn a supervised sense embedding for each labeled sense. The exact approach is proposed in LMMS (Loureiro and Jorge, 2019) but the learned sense embeddings are not extended with WordNet relations because we already have a knowledge-enhanced sense embedding learned from WordNet, detailed in section 4.2. Then we concatenate the **SREF$_{kb}$** sense embedding with the corresponding one learned from SemCor if the sense is labeled in SemCor, otherwise itself. Each context embedding $\vec{P}_w$ is concatenated with itself for vector dimension matching because the vector dimension of each sense embedding has doubled.

5.2 Systems for Comparison

We compare our experimental results with the state-of-the-art in both knowledge-based and supervised categories.

Knowledge-based systems: Lesk$_{enhanced}$ (Basile et al., 2014), UKB (Agirre et al., 2018), Babelfy (Moro et al., 2014), WSD-TM (Chaplot and Salakhutdinov, 2018) and KEF (Wang et al., 2020).

Supervised systems: EWISE (Kumar et al., 2019), GLU (Hadiwinoto et al., 2019), LMMS (Loureiro and Jorge, 2019), GlossBERT (Huang et al., 2019) and SENSEMBERT (Scarlini et al., 2020). We also include two systems that are available after the submission of this paper, namely BEM (Blevins and Zettlemoyer, 2020)

	Models	Test Datasets					Concatenation of all Test Datasets				
		SE2 (n=2282)	SE3 (1850)	SE07 (455)	SE13 (1644)	SE15 (1022)	ALL (7253)	N	V	A	R
Knowledge-based	Lesk$_{enhanced}$ (2014) †	63	63.7	56.7	66.2	64.6	63.7	69.8	51.2	51.7	80.6
	Babelfy (2014)	67	63.5	51.6	66.4	70.3	65.5	68.6	49.9	73.2	79.8
	UKB (2018) †	68.8	66.1	53	68.8	70.3	67.3	71.2	50.7	75.0	77.7
	WSD-TM (2018)	69	66.9	55.6	65.3	69.6	66.9	69.7	51.2	76.0	**80.9**
	KEF (2020) †	69.6	66.1	56.9	68.4	72.3	68	71.9	51.6	74	80.6
	SREF$_{kb}$	**72.7**	**71.5**	**61.5**	**76.4**	**79.5**	**73.5**	**78.5**	**56.6**	**79.0**	76.9
Supervised	EWISE (2019)	73.8	71.1	67.3*	69.4	74.5	71.8*	74	60.2	78	82.1
	GLU (2019)	75.5	73.6	68.1*	71.1	76.2	73.7*	-	-	-	-
	LMMS (2019)	76.3	75.6	68.1	75.1	77	75.4	78.0	64.0	80.7	83.5
	GlossBERT (2019)	77.7	75.2	**72.5***	76.1	80.4	76.8*	-	-	-	-
	SENSEMBERT (2020)	-	-	-	-	-	-	80.4	-	-	-
	SREF$_{sup}$	**78.6**	**76.6**	72.1	**78**	**80.5**	**77.8**	**80.6**	**66.5**	**82.6**	**84.4**
	EWISER (2020) ‡	78.9	78.4	71	78.9	79.3*	78.3*	81.7	66.3	81.2	85.8
	BEM (2020) ‡	79.4	77.4	74.5*	79.7	81.7	79*	81.4	68.5	83	87.9

Table 4: F1-% Performance on all-words WSD datasets, * represents those performance obtained (partially) as a development set. † denotes the systems that make use of the prior knowledge of MFS for unseen lemmas during testing. ‡ are systems proposed after this paper was submitted. **Bold** and underlined figures indicate the current (submission time) and previous state-of-the-art performance on the evaluation framework, respectively.

and EWISER (Bevilacqua and Navigli, 2020).

For lexical sample tasks, we compare our system with IMS+embeddings (Iacobacci et al., 2016), context2vec (Melamud et al., 2016), NN-CWEs (Wiedemann et al., 2019) and GLU (Hadiwinoto et al., 2019).

6 Results

In this section, an ablation study is first implemented to illustrate how the proposed factors contribute to the final WSD performance and a test set example is given regarding the try-again mechanism. Then, we compare our systems' performance on all-words and lexical sample datasets with state-of-the-art systems. Also, we demonstrate how the number of labeled sentences in SemCor affects the performance of SREF$_{sup}$ and LMMS. Finally, we experiment on how the knowledge-enhanced sense embeddings can benefit several similarity-calculating and ranking tasks with simple attempts.

6.1 Ablation Study

Table 3 shows the ablation analysis of SREF$_{kb}$ on the combined dataset and its POS portions, demonstrating the contribution of each proposed factor. In detail, gloss augmentation manages to boost the system's performance by 1 F1, equal to the contribution of *other_relations* which is manually defined in WordNet. This has revealed the potential of such a fine-grained WordNet gloss utilization, and the employment of more valuable resources such as Wikipedia rather than web mentions for further investigation. Another noteworthy observation is that the sense embedding enhancement damages adverb disambiguation performance.

Figure 4 provides an example about how the try-again mechanism in SREF$_{kb}$ selects the correct sense of *bell*. Here, the word is first falsely predicted to be *bell.n.03* which means *the sound of a bell* rather than *bell.n.01* that means *a hollow device made of metal*. The try-again mechanism manages to detect a more similar sense to the word's context, *fire_bell.n.01*, which is a hyponym of *bell.n.01*. In this case, the hyponymy relation helps the system to correctly disambiguate *bell*. There are also other cases where the super-sense relation contributes.

6.2 All-words WSD

Table 4 illustrates how different systems perform

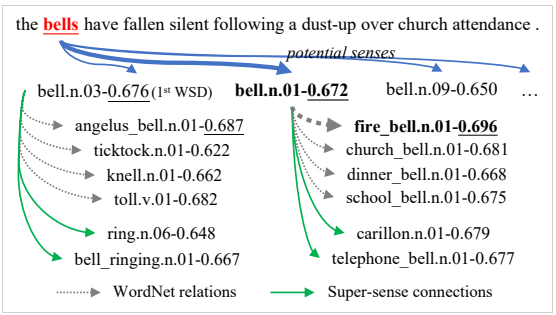

Figure 4: A Test Set Example of the Second WSD

Models	SE2-LS	SE3-LS	SE07
IMS+emb (2016)	69.9	75.2	62.6
context2vec (2016)	-	72.8	61.3
NN-CWEs (2019)	76.5	79.6	59.8
GLU (2019)	76.9	80.0	68.1
SREF$_{sup}$	**77.5**	**80.3**	**72.1**

Table 5: Performance on Lexical Sample Datasets

on standard WSD datasets separately (SE2, SE3, SE07, SE13, and SE15) and on their combined dataset (ALL). It has also shown those systems' performance on 'ALL' from POS perspectives.

For dataset-level performance, our relation-enhanced system, SREF$_{kb}$, achieves new state-of-the-art performance among the systems in the same category, surpassing the previous best system by 5.5 F1.

When our relation-enhanced sense embeddings are combined with the supervised sense embeddings learned from SemCor, our system (SREF$_{sup}$) also obtains new state-of-the-art performance among supervised systems, beating GlossBERT by 1 F1. GlossBERT utilizes SE07 as a developing set and tunes parameters on it. It is the first supervised system that performs over 70 F1 on SE07. SREF$_{sup}$, in contrast, requires no parameter tuning and reaches 72.1 F1 on SE07. It is also worth noting that SREF$_{sup}$ outperforms LMMS, a similar system, by almost 2.5 F1, revealing the tremendous benefits of explicit exploitation of WordNet sense relations.

Our systems also obtain state-of-the-art results in terms of POS disambiguation in both categories, achieving advantageous performance on more ambiguous word types (cf. Table 1) including verb, adjective and noun.

6.3 Lexical Sample WSD

We also conduct experiments on the English lexical sample tasks. For a fair comparison, we use the associated training dataset instead of SemCor to learn the supervised sense embeddings.

As is shown in Table 5, SREF$_{sup}$ obtains new state-of-the-art performance on lexical sample tasks, although the margin between previous best performance is relatively small. NN-CWEs and GLU are systems that employ BERT as a feature-extraction tool for their supervised learning framework but neglect WordNet sense knowledge. Therefore, although the systems can perform well on senses that are given sufficient labeled data for training, they do not have a good generalization ability to disambiguate rare or unseen senses. This is typically illustrated in their SE07 performance.

Models	MFS	LFS
WordNet S1	100	0
Lesk$_{enhanced}$	92.7	9.4
Babelfy	**93.9**	12.2
BiLSTM	93.4	22.9
EWISE	93.5	31.2
LMMS	87.6	52.6
SREF$_{kb}$	83.2	**55.2**
SREF$_{sup}$	91	53.2

Table 6: Performance on MFS and LFS against the 'ALL' dataset, where senses are partitioned into MFS and LFS according to their rank in WordNet.

6.4 Performance on Rare Senses

Except for the above regular experiments, we also set up an experiment regarding how our system performs on those synsets that are ranked first (MFS) in WordNet and the others (LFS, least frequent sense) in the 'ALL' dataset. We compare our results with those provided by EWISE, which is a zero-shot WSD system that makes use of sense gloss and relations in WordNet. EWISE has an overwhelming advantage of disambiguating unseen or rare senses and thus achieve much better results on LFS disambiguation. However, our systems (SREF$_{kb}$, SREF$_{sup}$) have better performance on LFS, although the margin between LMMS is not significant.

Table 6 demonstrates the performance on MFS and LFS for different systems. Although EWISE surpasses BiLSTM (Raganato et al., 2017a) on LFS disambiguation by a large margin, our supervised system still beats EWISE's performance by over 20 F1 while maintains a competitive performance on MFS disambiguation. This has shown our system's generalization ability of disambiguating rare sense.

6.5 Semcor Instance Utilization

Figure 5 demonstrates how the number of utilized Semcor sentences influences the performance of SREF$_{sup}$, LMMS and the sense embeddings learned from BERT and SemCor. For stable performance, we fix the sentence order in SemCor and incrementally extract a proportion of sentences to perform the experiments with a 10% step size. It is shown that even with 10% labeled data, SREF$_{sup}$ can outperform LMMS with full labeled data by 0.5 F1. Furthermore, SREF$_{sup}$

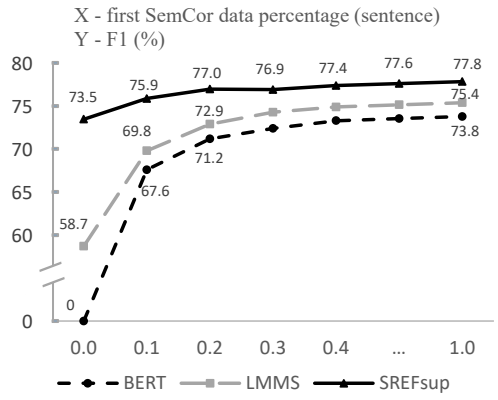

Figure 5: System Performance with Different Proportion of Semcor Sentences

obtains a new state-of-the-art result with only 20% labeled data.

6.6 Sense Embedding Application

In order to reveal the potential of $SREF_{sup}$ sense embeddings to other tasks, we experiment with three similarity-based tasks including SemEval-2017-Semantic Textual Similarity (Cer et al., 2017, SE17-STS-en-en), SemEval-2017 Task 3-SubtaskA and SubtaskB (Nakovet al., 2017, SE17-Task3-SubtaskA and SubtaskB). The similarity calculation is achieved by using merely BERT embeddings or concatenating them with the sum of $SREF_{sup}$ sense embeddings after disambiguating the text. The whole process is conducted in an unsupervised approach.

Table 7 shows that the utilization of sense embeddings is beneficial to these tasks. Nonetheless, a more plausible approach might be to utilize sense embeddings in a supervised framework, requiring further explorations.

6.7 Error Analysis

To implement the error analysis from a general perspective, we calculate the average ambiguity level (total number of potential senses divided by total number of ambiguous words) of those correctly and falsely disambiguated words by our system, 5.1 and 8.4 respectively. In a detail perspective, among the falsely disambiguated words, many competing senses are highly ambiguous and similar, and even their super-senses are hard to distinguish. For example, in '*The medicine can only be obtained with a prescription*' from SE15, the correct and predicted sense for *prescription* are so similar that algorithms that cannot spot the gloss focus

Tasks	BERT	$SREF_{sup}$
SE17-STS (Pearson correlation)	0.66	0.69
SE17-Task3-SubtaskA (MAP-%)	68.0	69.0
SE17-Task3-SubtaskB (MAP-%)	40.2	41.1

Table 7: Performance on Similarity-based Tasks

(*instruction* or *drug*) would fail, requiring the sense embedding to carry separate information regarding what the object is and what features it has.

Correct - *written instructions from a physician or dentist to a druggist concerning the form and dosage of a drug to be issued to a given patient.*

Predicted - *a drug that is available only with written instructions from a dentist to a pharmacist.*

7 Conclusion

We have introduced SREF, a synset relation-enhanced framework with a try-again mechanism that takes into account WordNet relations and augments WordNet glosses with mentions from the web under simple hypotheses and rules. Empirical experiments have proved the effectiveness of SREF from both knowledge-based and supervised perspectives, obtaining major and minor improvements over previous state-of-the-art performance, respectively.

For future work, we intend to scale $SREF_{kb}$ to a multilingual version and explore the possibilities of using the multilingual WordNet so that abundant knowledge regarding English can be transferred to other languages. It is also worth investigating regarding how to better incorporate sense embedding into other downstream tasks.

Acknowledgments

We thank the anonymous reviewers and Jianzhang Zhang for their insightful comments. This work was supported by the National Natural Science Foundation of China (under Project No. 61375053) and the graduate innovation fund of Shanghai University of Finance and Economics (under Project No. CXJJ-2019-395).

References

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

Eneko Agirre, Oier López de Lacalle, and Aitor Soroa. 2018. The risk of sub-optimal use of Open Source NLP Software: UKB is inadvert-ently state-of-the-

art in knowledge-based WSD. *In Proceedings of Workshop for NLP Open Source Software*, pages 29-33, Melbourne, Australia: Association for Computational Linguistics.

Satanjeev Banerjee and Ted Pedersen. 2002. An adapted Lesk algorithm for word sense disambiguation using wordnet. *In International conference on intelligent text processing and computational linguistics*, pages 136-145. Springer.

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, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Michele Bevilacqua and Roberto Navigli. 2020. Breaking Through the 80% Glass Ceiling: Raising the State of the Art in Word Sense Disambiguation by Incorporating Knowledge Graph Information. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2854-2864. Association for Computational Linguistics.

Sudha Bhingardive, Dhirendra Singh, Rudra Murthy V, Hanumant Redkar and Pushpak Bhattacharyya. 2015. Unsupervised most frequent sense detection using word embeddings. *In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1238-1243, Denver, Colorado, Association for Computational Linguistics.

Terra Blevins and Luke Zettlemoyer. 2020. Moving Down the Long Tail of Word Sense Disambiguation with Gloss Informed Bi-encoders. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1006-1017. Association for Computational Linguistics.

Jose 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.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, Lucia Specia. 2017. SemEval-2017 Task 1: Semantic Textual Similarity - Multilingual and Cross-lingual Focused Evaluation. *In Proceedings of the SemEval workshop at ACL 2017*, pages 1-14.

Devendra Singh Chaplot and Ruslan Salakhutdinov. 2018. Knowledge-based word sense disambiguation using topic models. *In Thirty-Second AAAI Conference on Artificial Intelligence*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171-4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Christian Hadiwinoto, Hwee Tou Ng and Wee Chung Gan. Improved Word Sense Disambiguation Using Pre-Trained Contextualized Word Representations. *In EMNLP-IJCNLP 2019*, pages 3507-3512, Hong Kong, China.

Luyao Huang, Chi Sun, Xipeng Qiu and Xuanjing Huang. GlossBERT: BERT for Word Sense Disambiguation with Gloss Knowledge. *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3507-3512, Hong Kong, China, Association for Computational Linguistics.

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, Berlin, Germany. Association for Computational Linguistics.

Adam Kilgarriff. 2001. English lexical sample task description. *In Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 17–20, Toulouse,France.

Sawan Kumar, Sharmistha Jat, Karan Saxena and Partha Talukdar. 2019. Zero-shot Word Sense Disambiguation using Sense Definition Embeddings. *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5670-5681, Florence, Italy. 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, SIGDOC '86*, pages 24-26, New York, NY, USA. ACM.

Daniel Loureiro and Alípio Mário Jorge. 2019. Language modelling makes sense: Propagating representations through WordNet for full coverage word sense disambiguation. *In Proceedings of the 57th Annual Meeting of the Association for*

Computational Linguistics, pages 5682-5691, Florence, Italy. Association for Computational Linguistics.

Fuli Luo, Tianyu Liu, Qiaolin Xia, Baobao Chang, and Zhifang Sui. 2018. Incorporating glosses into neural word sense disambiguation. *In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2473-2482, Melbourne, Australia. Association for Computational Linguistics.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2007. Unsupervised acquisition of predominant word senses. *Computational Linguistics*, 33(4):553-590.

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning generic context embedding with bidirectional LSTM. *In Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 51-61.

Rada Mihalcea and Dan I. Moldovan. 2001. Extended wordNet: progress report. *In Proceedings of NAACL Workshop on WordNet and Other Lexical Resources*, pages 95-100.

Rada Mihalcea, Timothy Chklovski, and Adam Kilgarriff. 2004. The senseval-3 English lexical sample task. *In Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 25–28, Barcelona, Spain.

George A. Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G. Thomas. 1994. Using a semantic concordance for sense identification. *In HUMAN LANGUAGE TECHNOLOGY: Proceedings of a Workshop held at Plainsboro*, New Jersey, March 8-11, 1994.

George A. Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 41(2): 39-41.

Tristan Miller and Iryna Gurevych. 2015. Automatic disambiguation of English puns. *In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL–IJCNLP 2015)*, pages 719-729, Beijing, China. Association for Computational Linguistics.

Tristan Miller, Chris Biemann, Torsten Zesch, Iryna Gurevych. 2012. Using Distributional Similarity for Lexical Expansion in Knowledge-based Word Sense Disambiguation. *In Proceedings of COLING 2012*, pages 1781-1796, Mumbai, India. The COLING 2012 Organizing Committee.

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, Denver, Colorado. Association for Computational Linguistics.

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.

Preslav Nakov, Doris Hoogeveen, Lluís Màrquez, Alessandro Moschitti, Hamdy Mubarak, Timothy Baldwin, Karin Verspoor. 2017. SemEval-2017 Task 3: Community Question Answering. *In Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017)*, pages 27-48.

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, Atlanta, Georgia, USA. Association for Computational Linguistics.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2):10:1-10:69.

Martha Palmer, Christiane Fellbaum, Scott Cotton, Lauren Delfs, and Hoa Trang Dang. 2001. English tasks: All-words and verb lexical sample. *In Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 21-24, Toulouse, France. Association for Computational Linguistics.

Simone Papandrea, Alessandro Raganato and Claudio Delli Bovi. 2017. SupWSD: A Flexible Toolkit for Supervised Word Sense Disambiguation. *In Proceedings of the EMNLP 2017 System Demonstrations*. pages 103-108, Copenhagen, Denmark, Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227-2237, New Orleans, Louisiana. Association for Computational Linguistics.

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, Uppsala, Sweden. Association for Computational Linguistics.

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, Prague, Czech Republic. Association for Computational Linguistics.

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, Copenhagen, Denmark. Association for Computational Linguistics.

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, Valencia, Spain. Association for Computational Linguistics.

Bianca Scarlini, Tommaso Pasini, Roberto Navigli. 2020. SENSEMBERT: Context-Enhanced Sense Embeddings for Multilingual Word Sense Disambiguation. *In Thirty-fourth AAAI Conference on Artificial Intelligence*.

Benjamin Snyder and Martha Palmer. 2004. The English all-words task. *In Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 41-43, Barcelona, Spain. Association for Computational Linguistics.

George Tsatsaronis, Michalis Vazirgiannis and Ion Androutsopoulos. 2007. Word Sense Disambiguation with Spreading Activation Networks Generated from Thesauri. *In Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1725-1730, Hyderabad, India.

Loïc Vial, Benjamin Lecouteux and Didier Schwab. Sense Vocabulary Compression through the Semantic Knowledge of WordNet for Neural Word Sense Disambiguation. *In proceedings of the 10th Global WordNet Conference*.

Loïc Vial, Benjamin Lecouteux, and Didier Schwab. 2018. UFSAC: Unification of Sense Annotated Corpora and Tools. *In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.

Yinglin Wang, Ming Wang and Hamido Fujita. 2020. Word Sense Disambiguation: A Comprehensive Knowledge Exploitation Framework. *Knowledge-Based Systems*, 10530.

Gregor Wiedemann, Steffen Remus, Avi Chawla and Chris Biemann. Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings. *In KONVENS 2019*.

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, Uppsala, Sweden. Association for Computational Linguistics.