Enhancing the Context Representation in Similarity-based Word Sense Disambiguation

Ming Wang[1, ‡], **Jianzhang Zhang**[2, ‡] and **Yinglin Wang**[3, ‡, *]
[‡]School of Information Management and Engineering,
Shanghai University of Finance and Economics
[‡]Alibaba Business School, Hangzhou Normal University
[1]wangming@163.sufe.edu.cn, [2]jianzhang.zhang@foxmail.com,
[3]wang.yinglin@shufe.edu.cn

Abstract

In previous similarity-based WSD systems, studies have allocated much effort on learning comprehensive sense embeddings using contextual representations and knowledge sources. However, the context embedding of an ambiguous word is learned using only the sentence where the word appears, neglecting its global context. In this paper, we investigate the contribution of both word-level and sense-level global context of an ambiguous word for disambiguation. Experiments have shown that the Context-Oriented Embedding (COE) can enhance a similarity-based system's performance on WSD by relatively large margins, achieving state-of-the-art on all-words WSD benchmarks in knowledge-based category.

1 Introduction

Word sense disambiguation (WSD) is aimed at selecting the correct sense for a word given its context. Potential senses of a word are from a sense inventory such as WordNet (Miller, 1995). WSD can be classified into lexical sample WSD and all-words WSD. The former focuses on disambiguating some particular words in many sentences, while the latter conducts WSD on every ambiguous word in the provided text.

The nature of all-words WSD allows the task to be more compatible to downstream applications. Nevertheless, the task becomes more difficult (Pradhan et al., 2007) while it also provides more context information (rather than a single sentence). Utilizing such global context can assist the systems to tackle WSD from an overall perspective.

Recent development of contextual representation models, has accelerated the progress of WSD. Many systems are proposed to tackle WSD by employing BERT either by extracting features (Vial et al., 2019; Loureiro and Jorge, 2019) or fine-tuning (Peters et al., 2019; Levine et al., 2020). However, these systems are mostly implemented with a single sentence context, especially for the systems (Huang et al., 2019; Blevins and Zettlemoyer, 2020) that fine-tune BERT (Devlin et al., 2019). As for the others (Scarlini et al., 2020a; Wang and Wang, 2020, Scarlini et al., 2020b), efforts are allocated to construct sense embeddings using WordNet or SemCor (Miller et al., 1994), while context embeddings for ambiguous words are learned merely from a single sentence. This has led to an issue that the information volume of context embeddings and sense embeddings is not balanced.

In this paper, we introduce COE, a context-oriented embedding technique to learn comprehensive context representations for ambiguous words. This is aimed at enhancing the context embeddings by considering both the global and local sentences in the provided document. In summary, our approach has the following contributions:

- We propose a novel technique to capture both local and global context information for context representation learning. The obtained

* corresponding author

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 8965–8973
November 7–11, 2021.

context embeddings are further enhanced with the embeddings of senses appeared in the context.

- We show that the proposed technique can elevate previous systems' performance on all-words WSD to new state-of-the-art in the knowledge-based category.

2 Method

2.1 Similarity-based WSD

Given a document d that contains several sentences, a system is required to determine the correct sense $s_{k,w_{i,j}}$ of each word $w_{i,j} \in \{w_{i,1}, w_{i,2}, \dots, w_{i,m}\}$ in sentence $S_i \in \{S_1, S_2, \dots, S_n\}$. $s_{k,w_{i,j}}$ is one of the potential senses in $S_{w_{i,j}}$ retrieved from WordNet by the lemma and part-of-speech (POS) of word $w_{i,j}$. In previous similarity-based WSD models (Loureiro and Jorge, 2019; Scarlini et al., 2020a; Wang and Wang, 2020; Scarlini et al., 2020b), sense embeddings of all WordNet senses are first learned using their definitions and other available resources. Then, in order to disambiguate $w_{i,j}$, the sense embedding $V_{s_{k,w_{i,j}}}$ of its potential sense $s_{k,w_{i,j}}$ is retrieved from the learned sense embedding pool. Then, the dot product of each potential sense embedding $V_{s_{k,w_{i,j}}}$ and the context embedding $P_{w_{i,j}}$ is used to select the optimal sense $\hat{s}_{w_{i,j}}$, shown in formula (1). $P_{w_{i,j}}$ is learned using only the sentence where $w_{i,j}$ appears.

$$\hat{s}_{w_{i,j}} = \underset{s_{k,w_{i,j}} \in S_{w_{i,j}}}{\operatorname{argmax}} V_{s_{k,w_{i,j}}} \cdot P_{w_{i,j}} \quad (1)$$

Typically, $P_{w_{i,j}}$ is the sum of BERT's last four layers at the position of $w_{i,j}$, taking s_i as its input. When $w_{i,j}$ is tokenized into several pieces, the sum of all its pieces' embeddings is taken as $P_{w_{i,j}}$. However, this naïve context representation learning process has limited the system's ability to capture global context information. In order to relieve this issue, we devise several methods to learn more comprehensive context embeddings by combining S_i and the other sentences in the same document. Note that, this work does not involve any attempt on sense embedding learning.

2.2 Context Embedding Learning

Local Context Embedding Following the approaches in prior works (Agirre et al. 2018, Wang et al., 2020), we utilize the directly surrounding sentences $\{S_1, \dots, S_{i-1}, S_{i+1}, \dots, S_n\}$ of the ambiguous sentence S_i for a more effective local context embedding. Here, we use a development set to select the optimal number of surrounding sentences on both sides of S_i and use the expanded sentence set as BERT's input to get the local context embedding $P^l_{w_{i,j}}$.

Global Context Embedding Except for the sentences that are in the same small window as the ambiguous sentence S_i, distant sentences are also beneficial for understanding the words in S_i in many cases. Here, we transform the problem into a sentence selection problem, i.e., to determine which sentences can better incorporate global context information for the disambiguation of the words in S_i.

We hence formally define the problem as follows: for each sentence $S_i \in \{S_1, S_2, \dots, S_n\}$ under disambiguation, we aim at ranking the other sentences in the same document according to their contributions from different perspectives. Then, we use S_i and its top ranked sentences to learn the global context embedding $P^g_{w_{i,j}}$. We design three methods to rank the sentences: word overlap (WO), TF-IDF score (TF-IDF WO), gloss-expanded word overlap (GeWO).

- Word overlap: the overlap count between S_i and S_j, i.e., the sum of the number of times that S_i's words appear in S_j.
- TF-IDF weighted word overlap: we regard each sentence $S_i \in \{S_1, S_2, \dots, S_n\}$ as a document and calculate the TF-IDF score of each word in the sentences; the TF-IDF score is then used to weight the overlap count between S_i and S_j for each word. The score of S_j with respect to S_i is calculated as follows:

$$score^{S_i}_{S_j} = \sum_{w \in S_i} tfidf_w * count(w, S_j) \quad (2)$$

- Gloss-expanded word overlap: we first expand each sentence $S_i \in \{S_1, S_2, \dots, S_n\}$ with all the synsets' definition words of each monosemous word $w_{i,j}$ and then calculate the overlap between expended S_i and S_j.

After we obtain the score of sentence $S_j \in \{S_1, \dots, S_{i-1}, S_{i+1}, \dots, S_n\}$ with respect to S_i, we rank them based on the scores and combine S_i and its top related sentences to learn a global context embedding. We note that, the sentence order is maintained when using them to learn the context

embedding. For instance, if S_{i-4} and S_{i+9} are the top 2 related sentences of S_i, we take $\{S_{i-4}, S_i, S_{i+9}\}$ as BERT's input for learning the global context embedding of each word in S_i. We also employ a development set to acquire the optimal number of related sentences for the global context embedding learning.

Sense-aware Context Embedding In most cases, the words in a given document are not always polysemous. This is verified by the statistics that 16.4% of words are monosemous in SemCor. These monosemous words can provide some general background information about the whole document. Here, we utilize the sense embeddings of the monosemous words to compose a sense-aware context embedding $P^s_{w_{i,j}}$.

In detail, all the sense embeddings of the monosemous words in the same document as $w_{i,j}$ are added together to obtain $P^s_{w_{i,j}}$ only when $w_{i,j}$ is a noun or verb. This is because the disambiguation of adjectives and adverbs tend to rely more on the local context information, indicating that it is a modifier (adjective or adverb) of which word (noun or verb) in the same sentence. We note that, for the knowledge-based approach, we also use the sense embedding of WordNet 1st sense for polysemous words in the document.

We combine the above local and global context embeddings after normalization to get the final enhanced context embedding $\hat{P}_{w_{i,j}}$, detailed in formula (3).

$$\hat{P}_{w_{i,j}} = P^l_{w_{i,j}} + P^g_{w_{i,j}} + P^s_{w_{i,j}} \quad (3)$$

2.3 Try-again Mechanism (TaM)

Wang and Wang (2020) proposed a try-again mechanism that exploits WordNet synset relations and super-sense connections to conduct a second WSD. Precisely, when disambiguating $w_{i,j}$, the method takes into account two similarity scores. One is from Formula (1). The other is calculated from a broader perspective, e.g., the maximal similarity between $P_{w_{i,j}}$ and one potential sense' ($s_{k,w_{i,j}}$) related synsets ($R_{s_{k,w_{i,j}}}$). These related synsets are connected to $s_{k,w_{i,j}}$ by WordNet synset relations and the super-sense connection. Here, synsets that are in the same super-sense category are regarded as connected by the super-sense connection. For example, *toy.n.03* (toy) {a device regarded as providing amusement} and *bell.n.01* (bell) {a hollow device made of metal that makes a ringing sound when struck} are both in the super-sense category of *noun.artifact*.

Formula (4) illustrates the final WSD calculation. The method manages to boost the knowledge-based system's performance by a relatively large margin, while slightly damages the performance of the supervised system.

$$\hat{s}_{w_{i,j}} = \underset{s_{k,w_{i,j}} \in S_{w_{i,j}}}{\operatorname{argmax}} \left(V_{s_{k,w_{i,j}}} \cdot P_{w_{i,j}} + \max_{s_l \in R_{s_{k,w_{i,j}}}} V_{s_l} \cdot P_{w_{i,j}} \right) \quad (4)$$

We improve the original mechanism by utilizing a higher quality of synset category named coarse sense inventory (CSI, Lacerra et al., 2020). CSI defines 45 labels in its inventory and covers 83,000 WordNet synsets. We replace the super-sense connection with CSI in the modified try-again mechanism. The revised mechanism leads our model to a better performance.

	ALL	A	N	R	V
COE$_{kb}$	**76.3**	80.5	**80.6**	**81.8**	**61.4**
-w/o local	75.1	**80.7**	79.4	81.2	59.3
-w/o global	74.8	80.6	78.9	**81.8**	59.4
-w/o sense	74.3	80.5	78.7	**81.8**	57.7
-w/o TaM	75.1	79.2	79.8	80.9	59.3
-w/o local+global	73.9	80.0	78.0	79.2	58.7
-w/o all (SREF$_{kb}$)	73.9	79.0	78.4	77.7	58.6

Table 1: Ablation Study on ALL (F1 in %)

3 Experiment

3.1 Datasets and Systems

We use the evaluation framework in (Raganato et al., 2017b) to evaluate our method's effectiveness.

In the following section, we report the performance of systems in the knowledge-based category for all-words WSD task, in comparison with ours. They consist of UKB (Agirre et al.,

COE$_{kb}$	ALL	A	N	R	V
-w/ WO	**76.3**	80.5	**80.6**	**81.8**	**61.4**
-w/ TF-IDF WO	76.0	80.8	80.2	81.5	61.2
-w/ GeWO	76.1	**81.0**	80.3	81.5	61.1

Table 2: COE$_{kb}$ performance on ALL with different scoring strategies for the global context embedding learning

Pretrained Model	Systems	Test Datasets					Concatenation of all Test Datasets				
		SE2	SE3	SE07	SE13	SE15	ALL	N	V	A	R
/	UKB (2018)	68.8	66.1	53.0	68.8	70.3	67.3	71.2	50.7	75.0	77.7
	WSD-TM (2018)	69.0	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	68.0	71.9	51.6	74.0	80.6
	SyntagNet (2019)	71.2	71.6	59.6	72.4	75.6	71.5	-	-	-	-
BERT	SENSEMBERT (2020)	70.8	65.4	58.0	74.8	75.0	70.1	75.9	50.3	74.3	80.9
	SREF$_{kb}$ (2020)	72.7	71.5	61.8	76.4	79.5	73.5	78.5	56.6	79.0	76.9
	COE$_{kb}$	**76.0**	**74.2**	**69.2***	**78.2**	**80.9**	**76.3***	**80.6**	**61.4**	**80.5**	**81.8**

Table 3: All-words WSD performance on different partitions of ALL, including dataset and POS (noun-N, verb-V, adjective-A and adverb-R) partitions. * indicates the performance that are obtained (partially) as a development set. Bold and underlined figures represent the current and previous state-of-the-art performance.

2018), Babelfy (Moro et al., 2014), WSD-TM (Chaplot and Salakhutdinov, 2018), KEF (Wang et al., 2020), SyntagNet (Maru et al., 2019) and SREF (Wang and Wang, 2020).

Throughout the whole paper, we utilize the knowledge-based version of SREF (Wang and Wang, 2020) sense embeddings to validate the effectiveness of our method.

Except for the knowledge-based version, we also implement the proposed method in some supervised similarity-based systems, achieving better performance than their original versions. However, the margin is not significant. Details are shown in Appendix.

4 Evaluation

4.1 Ablation Analysis

Table 1 demonstrates the ablation study on the combined dataset (ALL). An overall conclusion can be drawn that each of the proposed factors manages to raise the system's performance. F1 measure is reported in percentage in all the tables.

As one can see, although the sense-aware context embedding is simple and easy to implement, the strategy alone enhances the system's performance by 2 F1. This astonishing contribution owes to a fine quality sense embedding and the employment of WordNet 1st senses, an essential prior knowledge in WordNet. As for the other two factors regarding context sentence usage, the contribution of each factor is not as significant.

Viewing from another perspective, when both the local and global context embeddings are removed, the performance drop exceeds that of the system that ignores the sense-aware embeddings. This has illustrated a fact that both word-level and sense-level context embeddings are crucial for WSD. It is interesting to note that merely adding the sense-aware context embedding can ruin the contribution of TaM, which makes the last two systems (use only $P_{w_{i,j}}$ as the context embedding) perform identically on ALL.

		SREF$_{kb}$	COE$_{kb}$
overlap		5052	
non-overlap		310	482
	Ambiguity	7.17	8.27
	Noun	54%	55%
	Verb	33%	31%
	Adjective	10%	10%
	Adverb	3%	5%

Table 4: Correctly predicted instances by two models in ALL

In Table 2, the performance of COE$_{kb}$ on ALL has shown that the simplest strategy (WO) has led to the best performance, although the margin is not significant.

4.2 Overall Results

Table 3 shows how different systems perform on several partitions of ALL. Our system in both categories produces a new state-of-the-art.

The knowledge-based version of our system, COE$_{kb}$, outperforms the previous state-of-the-art system (SREF) on ALL by a relatively large margin, 2.8 F1. From the perspective of POS performance, COE$_{kb}$ is the first system that reaches 80 F1 on noun disambiguation, surpassing the previous SOTA by 3.1 F1.

In fact, the performance of COE$_{kb}$ has exceeded that of many supervised systems including GLU. GLU utilizes BERT as a feature extraction tool in a supervised manner. The fact that it merely relies on SemCor hampers the system's generalization ability since SemCor only covers a small proportion of WordNet senses. It is shown that those systems (EWISE and GLU) that fail to

lemma	contact (semeval2015.d003.s022.t005-noun)		
sentence (in lemma)	what be the precaution for the person who give the medicine or come_into **contact** with the animal?		
senses	contact.n.01	2.202	close interaction
	contact.n.02	2.182	the state or condition of touching or of being in immediate proximity
	contact.n.03	2.174	the physical coming together of two or more things
	contact.n.04	2.168	the act of touching physically
	contact.n.05	2.113	(electronics) a junction where things (as two electrical conductors) touch or are in physical contact

Table 5: A falsely predicted instance by COE_{kb} from SE15. Gold senses are in **bold**.

incorporate WordNet knowledge (especially definitions) perform poorly on SE13 and cannot outperform many lately proposed knowledge-based systems such as SyntagNet, SREF and COE. The performance of the systems in supervised category is shown in Appendix.

4.3 Case Study

In this subsection, we compare the experimental result of $SREF_{kb}$ and COE_{kb} in a detailed manner so as to find out on what aspects COE_{kb} performs well and poorly respectively. Table 4 shows the number of instances in ALL that are correctly disambiguated by $SREF_{kb}$ or COE_{kb} only (non-overlap). It also details the ambiguity (average number of potential senses per instance) and POS proportions of the above instances.

A key factor is revealed that COE_{kb} does not outperform $SREF_{kb}$ incrementally, which means COE_{kb} has falsely predicted many, 310, instances that are correctly predicted by $SREF_{kb}$. In this case, although COE_{kb} can disambiguate more ambiguous instances, it has somehow compromised the ability of disambiguating easier instances. This has triggered a question regarding how to customize the context exploitation for different instances. Nevertheless, the POS proportions of the instances that are only correctly predicted by each model is almost identical.

4.4 Error analysis

In Table 5, a falsely predicted example, among others, from SE15 is given to demonstrate what kind of instance COE_{kb} are typically weak at disambiguating. It is shown that the similarity of the top ranked senses to the context of *contact* is very close to each other. This is logical since the definition of these senses are semantically similar, which are hard to distinguish even for human beings.

The above dilemma has raised concerns about whether the systems have reached the upper bound of their capability, 80%. This is an estimated inter-annotator agreement in Navigli (2009), which means the percentage of words tagged with the same sense by two or more human annotators. Further, if a system's performance outperforms this upper bound, is it because of overfitting? To tackle the above issue, a plausible choice might be to construct a coarse-grained sense inventory, similar to Navigli et al. (2007). This might also lead to an easier application of WSD to downstream tasks.

5 Conclusion

In this paper, we have presented COE, a context-oriented embedding technique for similarity-based WSD systems. It takes better advantage of both word-level and sense-level information from the document where an ambiguous word appears. Experiments have shown that the proposed method can enhance a system's performance on all-words WSD by relatively large margins. The ablation study has shown the contribution of each proposed factor. The source code will be made available at GitHub for further development.

6 Ethics Impact Statement

This paper does not involve the presentation of a new dataset, an NLP application and the utilization of demographic or identity characteristics in formation.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (under Project No. 61375053), the graduate innovation fund of Shanghai University of Finance and Economics (under Project No. CXJJ-2019-395) and Hangzhou Normal University Scientific Research Staring Foundation (4135C50220204073).

References

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.

Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. 2014. An enhanced Lesk word sense disambiguation algorithm through a distributional semantic model. *In COLING 2014*, pages 1591-1600, Dublin, Ireland.

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.

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.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *In NAACL 2019*, pages 4171-4186, Minneapolis, Minnesota.

William A. Gale, Kenneth W. Church, and David Yarowsky. 1992. One sense per discourse. *In Proc. of the workshop on Speech and Natural Language. ACL*, pages 233-237.

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

Luyao Huang, Chi Sun, Xipeng Qiu and Xuanjing Huang. GlossBERT: BERT forWord Sense Disambiguation with Gloss Knowledge. *In EMNLP-IJCNLP 2019*, pages 3507-3512, Hong Kong, China.

Caterina Lacerra, Michele Bevilacqua, Tommaso Pasini, Roberto Navigli. 2020. CSI: A Coarse Sense Inventory for 85%Word Sense Disambiguation. *In Proceedings of The Thirty-Fourth AAAI Conference on Artificial Intelligence*, Pages 8123-8130. Association for the Advancement of Artificial Intelligence.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. *In SIGDOC '86*, pages 24-26, New York, NY, USA. ACM.

Yoav Levine, Barak Lenz, Or Dagan, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua and Yoav Shoham. 2020. SenseBERT: Driving Some Sense into BERT. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Pages 4656-4667. Association for Computational Linguistics.

Daniel Loureiro and Alípio Mário Jorge. 2019. Language modelling makes sense: Propagating representations through WordNet for full coverage word sense disambiguation. *In ACL 2019*, pages 5682-5691, Florence, Italy.

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.

Marco Maru, Federico Scozzafava, Federico Martelli, and Roberto Navigli. 2019. SyntagNet: Challenging Supervised Word Sense Disambiguation with Lexical-Semantic Combinations. *In Proc. Of EMNLP*, pages 3525-3531. Association for the Advancement of Artificial Intelligence.

Tristan Miller, Chris Biemann, Torsten Zesch and Iryna Gurevych. 2012. Using Distributional Similarity for Lexical Expansion in Knowledge-based Word Sense Disambiguation. *In COLING*, pages 1781-1796.

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.

Andrea Moro and Roberto Navigli. 2015. SemEval-

2015 task 13: Multilingual all-words sense disambiguation and entity linking. *In SemEval 2015*, pages 288-297, Denver, Colorado.

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.

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. SemEval-2013 task 12: Multilingual word sense disambiguation. *In SemEval 2013 *SEM*, pages 222-231, Atlanta, Georgia, USA.

Roberto Navigli, Kenneth C. Litkowski, and Orin Hargraves. 2007. Semeval-2007 task 07: Coarse grained english all-words task. *In Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 30–35, Prague, Czech Republic.

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*, pages 21-24, Toulouse, France.

Matthew E. Peters, Mark Neumann, Robert L. Logan IV, Roy Schwartz, Vidur Joshi, Sameer Singh and Noah A. Smith. Knowledge Enhanced Contextual Word Representations. *In EMNLP-IJCNLP 2019*, pages 3507-3512, Hong Kong, China.

Sameer Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. SemEval-2007 task-17: English lexical sample, SRL and all words. *In SemEval-2007*, pages 87-92, Prague, Czech Republic.

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 EACL 2017*, pages 99-110, Valencia, Spain.

Bianca Scarlini, Tommaso Pasini, Roberto Navigli. 2020a. SENSEMBERT: Context-Enhanced Sense Embeddings for Multilingual Word Sense Disambiguation. *In AAAI 2020*.

Bianca Scarlini, Tommaso Pasini and Roberto Navigli. 2020b. With More Contexts Comes Better Performance: Contextualized Sense Embeddings for All-Round Word Sense Disambiguation. *In the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Benjamin Snyder and Martha Palmer. 2004. The English all-words task. *In Senseval-3*, pages 41-43, Barcelona, Spain.

George Tsatsaronis, Michalis Vazirgiannis and Ion Androutsopoulos. 2007. Word Sense Disambiguation with Spreading Activation Networks Generated from Thesauri. *In IJCAI 2007*, 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*.

Ming Wang and Yinglin Wang. 2020. A Synset Relation-enhanced Framework with a Try-again Mechanism for Word Sense Disambiguation. *In the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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

Zhi Zhong and Hwee Tou Ng. 2010. It makes sense: A wide-coverage word sense disambiguation system for free text. *In ACL 2010 System Demonstrations*, pages 78-83, Uppsala, Sweden.

Appendix

1 COE$_{sup}$

To implement the supervised version of our system, we utilize the supervised sense embedding from SREF$_{sup}$. For COE$_{sup}$, the context embedding is a concatenation of two embeddings, with one from COE$_{kb}$ ($\hat{P}_{w_{i,j}}$) and the other ($P_{w_{i,j}}$) from the output of BERT using only the original sentence S_i as input. This is to guarantee an information symmetry of the embeddings since LMMS supervised sense embeddings in SREF$_{sup}$ are learned from SemCor with one sentence as input at each time. The calculation before TaM is shown in formula (5). To be consistent with the sense embeddings, we use BERT$_{\text{LARGE_CASED}}$ to learn the context embeddings.

$$\hat{s}_{w_{i,j}} = \underset{s_{k,w_{i,j}} \in S_{w_{i,j}}}{\operatorname{argmax}} [V^{kb}_{s_{k,w_{i,j}}}; V^{lmms}_{s_{k,w_{i,j}}}] \cdot [\hat{P}_{w_{i,j}}; P_{w_{i,j}}] \quad (5)$$

2 Systems

Supervised systems include EWISE (Kumar et al., 2019), LMMS (Loureiro and Jorge, 2019), GlossBERT (Huang et al., 2019), GLU (Hadiwinoto et al., 2019), Sense Vocabulary Compression (SVC, Vial et al., 2019), SENSEMBERT (Scarlini et al., 2020a), SREF (Wang and Wang, 2020), ARES (Scarlini et al., 2020b), BEM (Blevins and Zettlemoyer, 2020) and EWISER (Bevilacqua and Navigli, 2020). In this category, we only report the performance obtained by using SemCor as the training set for a fair comparison.

3 Results

3.1 Overall Performance

In Table 6, COE_{sup}, outperforms its direct competitor, SREF, by 1.8 F1, although the margin between the newly proposed systems that fine-tunes BERT (BEM) is smaller. BEM is a system that fine-tunes two separate BERT for encoding context and gloss respectively. The whole training process takes 2 to 3 days with 2 GPUs, which is comparatively expensive in terms of time and device. In comparison, COE_{kb} and COE_{sup} take less than half an hour to learn all the necessary sense embeddings.

3.2 Rare Lemma or Sense disambiguation

In this subsection, we implement two experiments on rare sense or lemma disambiguation. Following the setting in SREF and other previous works, we partition ALL into two subsets according to the gold label of each lemma, with one containing those lemmas whose sense is ranked 1st in WordNet (ALL_{WN_1st}) and the others (ALL_{WN_other}). The 1st sense of each lemma in WordNet can be regarded as the most frequent sense (MFS). This was manually sorted with the statistics from some sense-annotated corpora.

Table 7 compares different systems' performance on the two subsets of ALL. It shows that COE_{kb} has obtained an advantageous position at disambiguating rare senses, with a 2.5 F1 higher than its direct competitor, $SREF_{kb}$, while maintained a better performance on lemmas of MFS. COE_{sup} has also outperformed it direct opponent, $SREF_{sup}$, on rare sense disambiguation with a larger margin, 3 F1. In comparison to BEM, our system can scale much better to unseen or rare senses while still have a competitive capability of disambiguating MFS.

Following Scarlini et al. (2020b), we also conduct an experiment on those lemmas or senses that are in ALL but not in the training data, SemCor. For zero-shot lemmas/words, 1139 instances are extracted from ALL (ALL_{LFW}). In terms of senses that do not appear in SemCor, we extract 222 polysemous instances from ALL (ALL_{LFS}).

Table 8 shows that COE_{kb} has attained the best performance on both subsets, outperforming $SREF_{kb}$ 1.6 and 4.9 F1 on ALL_{LFS} and ALL_{LFW}, respectively. The margin between COE_{kb} and other newly proposed systems is even larger, revealing the tremendous potential of our system regarding zero-shot learning in WSD. It is also worth mentioning that COE_{sup} performs 8.4 F1 lower than the knowledge-based version on ALL_{LFS}. This has raised a question regarding how to balance the exploitation of the sense embeddings learned from SemCor and WordNet knowledge. In addition, an essential conclusion can be drawn that knowledge-based systems ($SREF_{kb}$ and COE_{kb}) have an overwhelming advantage on zero-shot sense disambiguation.

3.3 Sense Embeddings

Table 9 shows the performance of our systems using different sense embeddings, compared with the original system. Precisely, the proposed method is proven valid and robust when utilizing three different sense embedding sets. The largest margin is obtained in the knowledge-based

Models	ALL_{WN_1st} (n=4728)	ALL_{WN_other} (n=2525)
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
BEM	**94.1**	52.6
$SREF_{kb}$	83.2	55.2
$SREF_{sup}$	91.0	53.2
COE_{kb}	86.3	**57.7**
COE_{sup}	92.0	56.2

Table 7: Performance on Lemmas Whose Sense Label is Ranked 1st in Wordnet and the Others

Sup. (SemCor)	/	EWISE (2019)	73.8	71.1	67.3*	69.4	74.5	71.8*	74.0	60.2	78.0	82.1
	BERT (fine-tune)	GlossBERT (2019)	77.7	75.2	72.5*	76.1	80.4	76.8*	-	-	-	-
		GLU (2019)	75.5	73.6	68.1*	71.1	76.2	73.7*	-	-	-	-
		BEM (2020)	79.4	77.4	**74.5***	79.7	81.7	79.0*	81.4	68.5	**83.0**	**87.9**
	BERT (feature-extract)	SVC (2019)	77.5	77.4	69.5	76.0	78.3	76.7	79.6	65.9	79.5	85.5
		LMMS (2019)	76.3	75.6	68.1	75.1	77.0	75.4	78.0	64.0	80.5	83.5
		ARES (2020)	78.0	77.1	71.0	77.3	**83.2**	77.9	80.6	68.3	80.5	83.5
		SREF$_{sup}$ (2020)	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**	71.0	78.9	79.3*	78.3*	81.7	66.3	81.2	85.8
		COE$_{sup}$	**80.3**	77.6	73.6*	**80.7**	82.3	**79.6***	**82.3**	**68.9**	82.7	87.0

Table 6: All-words WSD performance for both supervised (Sup.) and knowledge-based (Know.) categories on different partitions of ALL, including dataset and POS (noun-N, verb-V, adjective-A and adverb-R) partitions. * indicates the performance that are obtained (partially) as a development set. Bold and underlined figures represent the current and previous state-of-the-art performance, respectively.

category, 2.4 F1. On the contrary, the proposed approach has only elevated ARES's performance by 0.5 F1.

Models	ALL$_{LFS}$ (n=1139)	ALL$_{LFW}$ (n=222)
LMMS	61.6	74.8
GlossBERT	62.0	75.6
ARES	65.2	81.1
SREF$_{kb}$	75.9	82.9
SREF$_{sup}$	67.3	82.4
COE$_{kb}$	**77.5**	**87.8**
COE$_{sup}$	69.1	**87.8**

Table 8: Performance on Lemmas or Senses in ALL with no annotation in SemCor

Sense Embedding	Model	F1	Δ
LMMS	† LMMS	75.4	-
	COE$_{sup}$	**76.5**	1.1
SREF	† SREF$_{kb}$	73.9	-
	COE$_{kb}$	**76.3**	2.4
	† SREF$_{sup}$	77.8	-
	COE$_{sup}$	**79.6**	1.8
ARES	† ARES	77.7	-
	COE$_{sup}$	**78.2**	0.5

Table 9: Systems' Performance on ALL with different sense embedding set. † indicates our