A Survey of Pre-trained Language Model-based Word Sense Disambiguation Models

Donghun Yang
University of Science and Technology,
Korea Institute of Science and Technology Information,
Daejeon, South Korea
yangdonghun3@kisti.re.kr

Myunggwon Hwang
University of Science and Technology,
Korea Institute of Science and Technology Information,
Daejeon, South Korea
mgh@kisti.re.kr

ABSTRACT

Transformer-based pre-trained language models have recently demonstrated excellent results in various Natural Language Processing (NLP) applications, including Word Sense Disambiguation (WSD) tasks. The WSD is a fundamental and core task in the NLP field, which aims to assign the appropriate meaning (sense) to ambiguous words in context from a list of sense candidates, but it is a long-standing challenge and open problem. To solve this problem, various WSD studies have been conducted, such as knowledge-based methods, supervised methods, and neural network-based methods. In recent years, studies on approaches utilizing a transformer-based pre-trained language model and its contextualized representation have been actively conducted and outperformed previous approaches. In this paper, we present a broad overview of recent trends in pre-trained language model-based WSD models, including the state-of-the-art WSD approaches, along with the characteristics, strong points and weak points of each model. Additionally, we point out the limitations of current WSD approaches, and suggest future research directions.

KEYWORDS

Word sense disambiguation, WSD, Transformer, Pre-trained language model, Contextualized representation

1 INTRODUCTION

Word Sense Disambiguation (WSD) is a fundamental and core task, but it is a long-stand challenge and open problem, in various Natural Language Processing (NLP) tasks, such as machine translation [1], named entity recognition [2], and information retrieval [3]. The goal of WSD is to assign the appropriate meaning (sense) to ambiguous words in context from a list of sense candidates [4]. In detail, given a variable-length sentence consisting of a sequence of words $X = \{w_0, w_1, \dots, w_n\}$ including target words to be disambiguated $T = \{t_0, t_1, \dots, t_k\}$, WSD system needs to predict appropriate sense $Y = \{s_0, s_1, \dots, s_k\}$ for each target word t_k. Each predicted sense s_k is selected from a pre-defined sense inventory S_t. Unfortunately, although it is easy for humans to infer the correct meaning of a target word from a context, it is not an easy challenge for WSD systems.

In order to solve WSD problem, various WSD studies have been conducted. Classical WSD approaches can be classified into knowledge-based and supervised models [4]. Knowledge-based WSD models rely on lexical resources such as semantic information networks (e.g., WordNet [5]) and sense definition (e.g., Gloss) [6-8]. Supervised WSD models exploit annotated datasets (e.g., SemCor [9]) to train statistical WSD models (e.g., Word Expert) [10,11]. While the supervised approaches typically perform better, it is less flexible than the knowledge-based approaches due to the lack of high coverage annotated datasets. Recent WSD studies have focused primarily on neural network-based supervised methods [12,13], as well as methods of incorporating various lexical knowledge into neural networks to alleviate the supervision bottleneck, and have outperformed classical approaches [14-16].

More recently, studies on the WSD approaches utilizing a transformer-based pre-trained language model (PLM) has been actively conducted, outperforming previous approaches. The transformer-based PLM, i.e., BERT [17], RoBERTa [18], DeBERTa [19], and etc., and their contextualized representations have recently demonstrated excellent performance not only in WSD tasks but also in most NLP applications. The transformer-based PLM is typically composed of multi-head self-attention layers, and is trained on a large corpus (e.g., Wikipedia) in a self-supervised manner (e.g., Masked Language Modeling (MLM)) [17]. During training, the model learns a contextualized representation of each word through the attention mechanism with respect to its context words [20]. Unlike conventional non-contextualized representations (e.g., Word2Vec [21]), this contextualized representation is able to assign different representations for a single word in different contexts. This representation can be used directly in the WSD task with various lexical knowledge, outperforming the previous WSD approaches. When the transformer-based PLM is fine-tuned on annotated WSD datasets, it can contribute to improved WSD performance, demonstrating state-of-the-arts results.

In this paper, we present a broad overview of recent trend on PLM-based WSD models, including the state-of-the-art approaches,

along with the characteristics, strong points and weak points of each model, divided into two branches: 1) WSD approaches that use the contextualized representations of the transformer-based PLM directly without fine-tuning (Frozen approaches); 2) WSD approaches that fine-tune the transformer-based PLM on the annotated WSD datasets with additional information (Fine-tuning approaches). Finally, based on these survey results, we point out the limitations of current WSD approaches, and suggest future research directions, at the end of this paper.

2 CONTEXTUALIZED REPRESENTATION-BASED WSD APPROACHES (Frozen)

As a first trial to use contextualized representation of the transformer-based PLM, a softmax-based WSD study (**BERTlw** and **BERTglu)** was conducted that feed the output representation of the last layer of a pre-trained BERT, as well as the output representations of all layers to a feedforward network [22]. In this study, a simple linear projection (LW) method and gated linear unit (GLU) method were proposed for weighed summing representations of all layers in pre-trained BERT. These approaches demonstrated the potential of capturing lexical and semantic information for WSD even if the contextualized representation is directly used.

Following that, several studies were conducted to compute high coverage sense representation, which is capable of covering senses that do not appear in annotated WSD datasets (SemCor [9]) by integrating lexical knowledge into contextualized representations. In these studies, after obtaining the sense representations, the closest sense is assigned to the target word via a similarity test (1-nn) between the target word representation and the sense representations. **LMMS** provided a sense representation computed by concatenating the contextualized representations of the sense definition (Gloss) and the WordNet relations [23]. As an extended version of LMMS, **SparsLMMS** was proposed, which makes the sparse sense representation by applying sparse coding to LMMS representation, demonstrating the efficiency of sparse vector for WSD task [24]. Furthermore, **ARES** presented the sense representation based on Personalized Page Rank algorithm (PPR) and K-means clustering with WordNet relationships, as well as co-occurrence words provided by SynagNet [25]. The most effort in this approach is SensEmBERT and SREF. **SensEmBERT** utilized BabelNet-provided Wikipedia pages as enhanced gloss information for each sense to generate higher quality contextualized representations, being limited, however, to noun sense only [26]. **SREF** presented high-quality sense embedding based on the WordNet relations and additional gloss information retrieved directly from the web, as well as a Try-again mechanism (TAM) that considered the second similar sense once more, resulting in significantly improved WSD performance [27].

In the other direction, **EWISER** was introduced as an extension of EWISE, which is based on BI-LSTM and ConvE with wordnet relations [16], providing a hybrid version of the knowledge-based and supervised WSD approaches. EWISER performs WSD tasks by integrating a contextualized representation of BERT and a sparse adjacency matrix of WordNet into a softmax-based classifier, attaining better WSD performance than previous approaches [28].

3 FINE-TUNING-BASED WSD APPROACHES (Fine-tuning)

After the Transformer architecture was introduced [20], **SenseBERT** was proposed to further perform WSD tasks in conjunction with masked language modeling (MLM) of BERT [17] during training [29]. In addition to this, studies on the fine-tuning approach for WSD have been actively conducted. In the **GlossBERT**, a context sentences with a target word and a sense definition (Gloss) sentence are fed together into a pre-trained BERT, and WSD is performed through binary classification for the Gloss sentence [30]. As an extended version of GlossBERT, **ESR** enhanced Gloss sentences by concatenating Gloss sentences of each sense with synonyms, hypernyms, and examples from WordNet, and performs WSD tasks similarly to GlossBERT through fine-tuning based on pre-trained RoBERTa [18], resulting in significantly improved performance [31].

On the other hand, in the **BEM** model, a context sentence and each Gloss sentence for target word are input to different BERT encoder, and WSD is performed via the inner product between the context output representation and each Gloss output representations [32]. As an improved version of BEM, **SACE** was proposed, which devised an interactive sense embedding learning mechanism (called a selective attention layer) into the BEM model that takes into account previously assigned senses in context during training [33]. As another extension of BEM, **SemEq** enhanced Gloss sentences by leveraging six additional sense inventories: Oxford Advanced Learner's Dictionary (Turnbull, 2010), Merriam Webster's Advanced Learner's Dictionary (Perrault,2008), Collins COBUILD Advanced Dictionary (Sinclair, 2008), Cambridge Advanced Learner's Dictionary (Walter, 2008), and Longman Dictionary of Contemporary English (Summers, 2003) [34]. In this approach, each gloss of the six different sense inventories, corresponding to the same sense is aligned using pre-trained SentenceBERT [35], and WSD tasks are performed via contrastive learning based on RoBERTa, in a slightly different manner than the BEM, demonstrating close to state-of-the-art performance.

Meanwhile, **ESCHER** presented a span extraction framework for WSD tasks [36]. Given a context sentence including target word concatenated with all its possible Gloss sentences, the ESCHER model predicts span of the most appropriate Gloss sentence for target word. As an extension, **ConSeC** was proposed, which enhanced input by concatenating ESCHER input sentence with a Gloss list of already disambiguated words [37]. In this model, a modified relative positional embedding is also introduced for a more efficient attention mechanism between the words in the context sentence and Gloss sentences, based on DeBERTa [19], resulting in the state-of-the-art WSD performance.

In the other direction, **WMLC** was proposed, which performs WSD tasks by fine-tuning on a multi-label classification problem instead of a softmax-based classification, as a simple modification

of BERTglu [38]. Additionally, **SVC** presented sense vocabulary compression approaches based on WordNet [39]. By leveraging WordNet relations such as hypernymy, hyponymy, and synonymy to keep only the minimum senses required to classify all senses, the SVC approach provides the benefits of a high coverage sense and the reduced WSD model size.

4 CONCLUSIONS

In this paper, we surveyed a broad overview of recent trends in pre-trained language model-based WSD approaches, divided into two main branches. The first branch is WSD approaches that use the contextualized representations of the transformer-based PLM directly without fine-tuning (Frozen approaches). The other branch is WSD approaches that fine-tune the transformer-based PLM on the annotated WSD datasets with additional information (Fine-tuning approaches). In both directions, various studies have been conducted to integrate lexical knowledge such as Gloss, WordNet relations, and external knowledge into contextualized representations or pre-trained language models. However, most of the studies were conducted independently, and it was considered that the performance of WSD can be improved if multiple approaches are used together. Therefore, we suggest that ablation studies be performed on the existing WSD approaches to investigate ways to improve WSD performance, as a further study and future research directions.

ACKNOWLEDGMENTS

This research was supported by Korea Institute of Science and Technology Information (KISTI).

REFERENCES

[1] Dengji Guo, Zhengrui Ma, Min Zhang, and Yang Feng. 2022. Prediction Difference Regularization against Perturbation for Neural Machine Translation. *In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers). 7665–7675.

[2] Chirawan Ronran, Seungwoo Lee, and Hong Jun Jang. 2020. Delayed combination of feature embedding in bidirectional LSTM CRF for NER. *Applied Sciences* 10, 21 (2020), 7557.

[3] Jimmy Lin. 2022. A proposed conceptual framework for a representational approach to information retrieval. *In ACM SIGIR Forum*, Vol. 55. ACM New York, NY, USA, 1–29.

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

[5] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[6] 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*. 24–26.

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

[8] 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 (2014), 231–244.

[9] George A Miller, Claudia Leacock, Randee Tengi, and Ross T Bunker. 1993. A semantic concordance. In Human Language Technology: *Proceedings of a Workshop Held at Plainsboro*, New Jersey, March 21-24, 1993.

[10] 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*. 78–83.

[11] 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). 897–907.

[12] Mikael Kågebäck and Hans Salomonsson. 2016. Word Sense Disambiguation using a Bidirectional LSTM. COLING 2016 (2016), 51.

[13] Alessandro Raganato, Claudio Delli Bovi, and Roberto Navigli. 2017. Neural sequence learning models for word sense disambiguation. *In Proceedings of the 2017 conference on empirical methods in natural language processing*. 1156–1167.

[14] 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). 2473–2482.

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

[16] 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*. 5670–5681.

[17] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *In Proceedings of NAACL-HLT*. 4171–4186.

[18] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint* arXiv:1907.11692 (2019).

[19] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DEBERTA: DECODING-ENHANCED BERT WITH DISENTANGLED ATTENTION. *In International Conference on Learning Representations*.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[21] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* 26 (2013).

[22] Christian Hadiwinoto, Hwee Tou Ng, and Wee Chung Gan. 2019. Improved word sense disambiguation using pre-trained contextualized word representations (to appear*). In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing*.

[23] Daniel Loureiro and Alipio 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*. 5682–5691.

[24] Gábor Berend. 2020. Sparsity Makes Sense: Word Sense Disambiguation Using Sparse Contextualized Word Representations. *In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 8498–8508.

[25] Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020. With more contexts comes better performance: Contextualized sense embeddings for all-round word sense disambiguation. *In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 3528–3539.

[26] Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020. Sensembert: Context-enhanced sense embeddings for multilingual word sense disambiguation. *In Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 8758–8765.

[27] Ming Wang and Yinglin Wang. 2020. A synset relation-enhanced framework with a try-again mechanism for word sense disambiguation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 6229–6240.

[28] 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*. 2854–2864.

[29] Yoav Levine, Barak Lenz, Or Dagan, Ori Ram, 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*. 4656–4667.

[30] Luyao Huang, Chi Sun, Xipeng Qiu, and Xuan-Jing Huang. 2019. 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). 3509–3514.

[31] Yang Song, Xin Cai Ong, Hwee Tou Ng, and Qian Lin. 2021. Improved Word Sense Disambiguation with Enhanced Sense Representations. *In Findings of the Association for Computational Linguistics: EMNLP 2021*. 4311–4320.

[32] 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*. 1006–1017.

[33] Ming Wang and Yinglin Wang. 2021. Word sense disambiguation: Towards interactive context exploitation from both word and sense perspectives. *In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (Volume 1: Long Papers). 5218–5229.

[34] Wenlin Yao, Xiaoman Pan, Lifeng Jin, Jianshu Chen, Dian Yu, and Dong Yu. 2021. Connect-the-Dots: Bridging Semantics between Words and Definitions via Aligning Word Sense Inventories. *In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics*, Online and Punta Cana, Dominican Republic, 7741–7751.

[35] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *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). Association for Computational Linguistics*, Hong Kong, China, 3982–3992.

[36] Edoardo Barba, Tommaso Pasini, and Roberto Navigli. 2021. ESC: Redesigning WSD with Extractive Sense Comprehension. *In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics*, Online, 4661–4672

[37] Edoardo Barba, Luigi Procopio, and Roberto Navigli. 2021. ConSeC: Word Sense Disambiguation as Continuous Sense Comprehension. *In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics*, Online and Punta Cana, Dominican Republic, 1492–1503.

[38] Simone Conia and Roberto Navigli. 2021. Framing Word Sense Disambiguation as a Multi-Label Problem for Model-Agnostic Knowledge Integration. *In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Association for Computational Linguistics*, Online, 3269–3275.

[39] Loïc Vial, Benjamin Lecouteux, and Didier Schwab. 2019. Sense Vocabulary Compression through the Semantic Knowledge of WordNet for Neural Word Sense Disambiguation. *In Proceedings of the 10th Global Wordnet Conference*. Wroclaw, Poland.