Word Sense Disambiguation Based on Memory Enhancement Mechanism

Baoshuo Kan[1], Wenpeng Lu[1](✉), Xueping Peng[2], Shoujin Wang[3],
Guobiao Zhang[1], Weiyu Zhang[1], and Xinxiao Qiao[1]

[1] School of Computer Science and Technology, Qilu University of Technology
(Shandong Academy of Sciences), Jinan, China
{lwp,zwy,qxxyn}@qlu.edu.cn
[2] Australian Artificial Intelligence Institute, University of Technology Sydney,
Sydney, Australia
xueping.peng@uts.edu.au
[3] School of Computer Science and IT, RMIT University, Melbourne, Australia
shoujin.wang@rmit.edu.au

Abstract. Word sense disambiguation (WSD) is a very critical yet challenging task in natural language processing (NLP), which aims at identifying the most suitable meaning of ambiguous words in the given contexts according to a predefined sense inventory. Existing WSD methods usually focus on learning the semantic interactions between a special ambiguous word and the glosses of its candidate senses and thus ignore complicated relations between the neighboring ambiguous words and their glosses, leading to insufficient learning of the interactions between words in context. As a result, they are difficult to leverage the knowledge from the other ambiguous words which might provide some explicit clues to identify the meaning of current ambiguous word. To mitigate this challenge, this paper proposes a novel neural model based on memory enhancement mechanism for WSD task, which stores the gloss knowledge of previously identified words into a memory, and further utilizes it to assist the disambiguation of the next target word. Extensive experiments, which are conducted on a unified evaluation framework of the WSD task, demonstrate that our model achieves better disambiguation performance than the state-of-the-art approaches (Code: https://github.com/baoshuo/WSD).

Keywords: Word sense disambiguation · Gloss information · Memory mechanism · Memory enhancement

1 Introduction

Word sense disambiguation (WSD) aims at identifying the most suitable sense of an ambiguous word in its given context, which is a classical and challenging task in the natural language processing (NLP) area. Specifically, WSD is an essential and critical component in broad NLP applications, such as text classification [19], machine translation [7] and dialogue tasks [12]. Numerous WSD solutions

G. Memmi et al. (Eds.): KSEM 2022, LNAI 13369, pp. 249–260, 2022.
https://doi.org/10.1007/978-3-031-10986-7_20

have been introduced, which can be generally categorized into knowledge-based and supervised methods.

Knowledge-Based Methods try to fully utilize the knowledge in lexical knowledge bases such as WordNet and BabelNet [17,24]. They either consider the overlap and similarity between the context and glosses of each ambiguous word's senses, or construct a graph for all candidate senses in the context and employ graph-based algorithms. Although knowledge-based methods are flexible and have achieved successes in WSD task, they usually show inferior performance than their supervised counterpart.

Supervised Methods treat WSD as a classification task, and these methods rely on the semantically-annotated corpus for training the classification models. Recently, the effectiveness of supervised methods have been demonstrated in WSD task [9,21]. Particularly, the methods based on neural models have achieved exceptional successes, and show great potentials for addressing the WSD task [5,11]. To be specific, early neural WSD methods leverage neural sequence models, such as LSTM and Seq2Seq, to disambiguate target words. They focus on learning interactive relations between senses of ambiguous words and their context only [10,22]. Although those neural WSD methods could model the dependencies between the candidate senses and the context, they fail to consider the valuable lexical knowledge employed by the knowledge-based counterpart. Aiming at addressing this deficiency, some works attempt to leverage lexical knowledge to optimize supervised neural methods [5], which incorporate the gloss knowledge together with the context into neural WSD models. Although they break the barriers of supervised methods and knowledge-based ones, they mostly model glosses and context with two independent encoders. These approaches unable to capture the glosses-context interactions to strengthen the representations of each other. Therefore, some works propose to learn sense and context representation interactively by generating sense-level context for WSD task [16,26]. However, these methods only show marginal improvements. The possible reason is that they merely focus on the learning of the glosses of the target ambiguous word and the context, while neglecting the glosses of the other neighboring ambiguous words. In real scenarios, when human identify an ambiguous word, it is natural to utilize the gloss information of the previously identified senses of its neighboring words. However, such practice has not been modeled by existing methods.

As well known, when a person reads a sentence containing multiple ambiguous words, he will memorize the identified senses of ambiguous words, and utilize the sense knowledge to assist the disambiguation of the following words. As shown in Table 1, the context contains three ambiguous words, i.e., *monitor*, *table*, and *mouse*. According to their order in the context, once the senses of *monitor* and *table* are identified, their corresponding glosses, i.e., G^{s_1} for *monitor* and G^{s_2} for *table*, will be memorized and utilized to identify the sense of the following ambiguous word, i.e., *mouse*. With the context and the glosses of identified neighboring ambiguous words, a person can identify the right sense

Table 1. Ambiguous words in context and their sense glosses. The ellipsis "..." indicates the remainder of the gloss.

Context		He looks at the **monitor** on **table** and operate it with a **mouse**
Gloss	Monitor	g1: electronic equipment that is used to check the quality or
		g2: someone who supervises (an examination)
	Table	g1: a set of data arranged in rows and columns
		g2: a piece of furniture having a smooth flat top that is usually ...
	Mouse	g1: any of numerous small rodents
		g2: a hand-operated electronic device

easily, i.e., G^{s_2} for *mouse*. However, existing methods neglect to consider the knowledge from the identified ambiguous words and fail to introduce a suitable mechanism to store them to help the disambiguation of the following words. As a result, the interactions between the glosses of identified ambiguous word's senses and the current ambiguous word are missing, which inevitably hurt the performance on WSD task. To this end, how to enhance the learning on the interactions between identified glosses and current ambiguous word is critical for the further performance improvement.

To overcome these limitations, we propose a novel WSD model based on memory enhancement mechanism. Intuitively, memory mechanism can simulate the human reading behaviors to store and memorize the known information, and infer the unknown information [18]. It provides us with the flexibility and capability to capture interaction enhancement between previously identified glosses and the current ambiguous word in our model. Specifically, we first encode the context of the target word and each candidate gloss of the target word by the context-encoder unit and the gloss-encoder unit, respectively. Next, we propose a memory-enhancement unit to enhance the learning of the current target word by making interactions with the glosses of the identified neighboring words stored in the memory previously. Then, we introduce a prediction unit to score each candidate sense of the target word to select the right sense, which is stored into the memory and is employed to enhance the learning of the following ambiguous words.

We summarize the contributions of this paper as follows:

- We propose a novel model for WSD task, i.e., word sense disambiguation based on memory enhancement mechanism (MEM). As far as we know, this is the first work to leverage memory mechanism to model and enhance the interactions between target ambiguous words and the previously identified ones.
- We propose a memory enhancement mechanism, which stores the gloss knowledge of previously identified words in a memory, and utilizes the memory to enhance the representation of the next ambiguous word.
- Experiments on the real-world dataset demonstrate that the proposed model achieves better performance than the compared state-of-the-art benchmarks.

2 Related Word

2.1 Knowledge-Based WSD

Knowledge-based approaches rely on the lexical knowledge to justify the right sense of ambiguous word, which can be categorized into similarity-based methods and graph-based ones. The similarity-based methods usually consider the similarity between the context and the sense information, such as the gloss of the candidate sense, and adopt the sense with the highest similarity as the right one [2,3]. The graph-based methods usually build a graph based on the context and semantic relations retrieved from a lexical knowledge base, then evaluate the importance of each candidate sense to identify the right one [1,20,24]. Although knowledge-based approaches are flexible and show better performance on the coverage rate, they are hard to achieve satisfied performance on precision as the supervised approaches.

2.2 Supervised WSD

Supervised approaches treat WSD task as a classification problem, which are trained on the sense-annotated corpus. In recent years, the methods based on neural models have shown great potentials to address the classification problem. Unlike knowledge-based approaches, some supervised methods succeed to achieve excellent performance by utilizing sense embedding and contextual embedding, instead of lexical information in knowledge bases [10,14,22]. To explore the ability of lexical knowledge, GAS injects the gloss information into supervised models [16]. Following the work of GAS, more methods attempt to integrate lexical knowledge into supervised models, such as BEM [5] and EWISER [4]. BEM learns the representations of the target words and context in the same embedding space with a context encoder, and models the sense representations with a gloss encoder [5]. In addition to the gloss of the current sense, EWISER further utilizes the external explicit relational information from WordNet to enhance its ability [4]. However, these methods neglect the interactions between context and glosses, which can not enhance each other. To address this limitation, very recent works [15,26] attempt to model the interactions between context and glosses. CAN proposes a mechanism to generate co-dependent representations of glosses and the context [15]. SACE strengthens the learning of sense-level context, which takes into account senses in context with a selective attention layer [26]. However, the methods only achieve limited improvements. The possible reason is that they neglect to utilize the gloss information of previously identified senses to assist the disambiguation of the following target words according to the behavior pattern as human reads ambiguous sentences. In this paper, we strive to design a memory enhancement mechanism to solve the problem for WSD task.

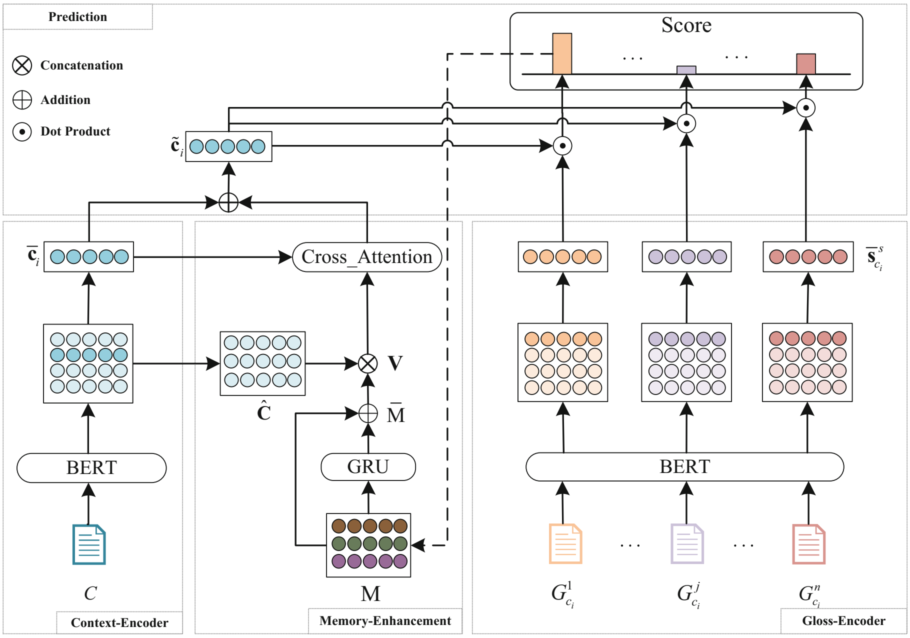

Fig. 1. Overview of our proposed MEM model

3 Methodology

3.1 Task Definition

In the all-words WSD task, the right sense of each ambiguous word in the given context should be identified. Formally, the input of a WSD model is the context $C = \{c_1, c_2, \ldots, c_n\}$ and its output is a sequence of sense predictions $S = \{s^i_{c_1}, s^j_{c_2}, \ldots, s^k_{c_n}\}$, where c_1, c_2, c_n are ambiguous words in the context, and $s^i_{c_1}$, $s^j_{c_2}$ and $s^k_{c_n}$ represent the right senses, i.e., i-th, j-th and n-th sense from the candidate sense sets for c_1, c_2 and c_n. $s \in S_{c_i}$, where S_{c_i} is all candidate senses of the ambiguous word c_i. For each sense s of c_i, we represent its gloss with $G^s_{c_i} = \{g_1, g_2, \ldots, g_n\}$.

3.2 Model Architecture

The overall architecture of our proposed model, called MEM, is shown in Fig. 1. MEM consists of four units, i.e., *the context-encoder unit*, *the gloss-encoder unit, the memory-enhancement unit* and *the prediction unit*. First, the context-encoder unit and the gloss-encoder unit encode the context and the gloss of each candidate sense of the target ambiguous word, respectively. Then, the memory-enhancement unit enhances the representation of the current target word by learning the its interactions with the glosses of the previously identified neighboring words stored in the memory. Finally, the prediction unit score each candidate sense of the target word to select the right sense for it. Such selected sense is stored into the memory and will be employed to enhance the sense disambiguation of the following ambiguous words.

3.3 Context-Encoder and Gloss-Encoder Units

Inspired by the work of BEM [5], we introduce context-encoder unit and gloss-encoder unit respectively, whose structures are shown in Fig. 1. Both encoders are initialized with BERT to benefit from its powerful ability in capturing interactions between words. The inputs of encoders are padded with BERT-specific start and end symbols, i.e., [CLS] and [SEP]. The context and sense gloss of the target ambiguous word are represented as $C = [\text{CLS}], c_1, c_2, \ldots, c_n, [\text{SEP}]$ and $G^s_{c_i} = [\text{CLS}], g_1, g_2, \ldots, g_n, [\text{SEP}]$, respectively.

The context-encoder unit takes the context C as its input, and encodes it to generate the context representation $\mathbf{C}$ with BERT, described with Eq. (1).

$$\mathbf{C} = \text{BERT}\left(C\right), \quad \bar{\mathbf{c}}_i = \mathbf{C}[i], \tag{1}$$

where $\bar{\mathbf{c}}_i$ refers to the representation of the i-th word in the context. When the target word is tokenized into multiple subword pieces by BERT tokenizer, its representation is calculated as the average representation of all its subword pieces.

The gloss-encoder unit takes the gloss $G^s_{c_i}$ as its input, and encodes it with BERT. It then selects the output of BERT at the position of [CLS] as the gloss representation.

$$\bar{\mathbf{s}}^s_{c_i} = \text{BERT}^{CLS}(G^s_{c_i}), \tag{2}$$

where $s \in S_{c_i}$ is one of the candidate senses for the target word c_i, $G^s_{c_i}$ is the gloss of s, and $\bar{\mathbf{s}}^s_{c_i}$ is the gloss representation for s.

3.4 Memory-Enhancement Unit

To enhance the representation of the target word, we propose a memory enhancement mechanism. Specifically, we first build a memory to store the representations of the glosses of the previously-identified neighboring words. Then we encourage the interactions between the representation of the target word and gloss representations of those neighboring words stored in the memory. This practice can model human reading behavior, namely, to utilize the previously-identified sense to assist the understanding of the following ambiguous words.

As shown in Fig. 1, one memory component is utilized to store the glosses representations of the previously-identified neighboring words, i.e., $\mathbf{M} = \{\bar{\mathbf{s}}^y_{c_x}\}$. For each representation $\bar{\mathbf{s}}^y_{c_x}$ in the memory, its subscript c_x indicates the c_x-th ambiguous word before the current target word, and its superscript y indicates the identified y-th sense from the candidate sense sets for the c_x-th ambiguous word. For modeling the sequential relations in different glosses, we employ Gated Recurrent Unit (GRU)to reconstruct their representations. Then, the original representation and the reconstructed one are added together to update the memory representation $\bar{\mathbf{M}}$. The above operations are described using Eq. (3).

$$\bar{\mathbf{M}} = [\mathbf{M} \oplus \mathrm{GRU}(\mathbf{M})], \tag{3}$$

where $\oplus$ refers to the addition operation.

Then, in order to utilize the features from neighboring words, we concatenate the context representation $\hat{\mathbf{C}}$ of the neighboring words of the target word and the memory representation $\bar{\mathbf{M}}$ together to generate the auxiliary information representation $\mathbf{V}$ for the current target word:

$$\mathbf{V} = [\hat{\mathbf{C}}; \bar{\mathbf{M}}], \tag{4}$$

where $\hat{\mathbf{C}}$ is obtained by removing the current target word representation from $\mathbf{C}$.

After obtaining the auxiliary information representation $\mathbf{V}$ of the current target word, we employ a cross-attention mechanism [6] to capture the interactions between the representation of the current target word $\bar{\mathbf{c}}_i$ and $\mathbf{V}$ to generate the enhanced representation $\tilde{\mathbf{c}}_i$ of the current target word. The operations are described with Eq. (5).

$$\tilde{\mathbf{c}}_i = f\left(\bar{\mathbf{c}}_i\right) + \mathrm{CA}(\mathrm{LN}\left[f\left(\bar{\mathbf{c}}_i\right); \mathbf{V}\right]), \tag{5}$$

where $f\left(\cdot\right)$ is the fully connected function, CA indicates the cross-attention mechanism [6], LN refers to the layer normalization. The detailed operations of cross-attention mechanism is described as Eq. (6).

$$\begin{gathered} \mathbf{q} = \bar{\mathbf{c}}_i \mathbf{W}_q, \qquad \mathbf{k} = \mathbf{V}\mathbf{W}_k, \qquad \mathbf{v} = \mathbf{V}\mathbf{W}_v, \\ \mathbf{A} = \mathrm{softmax}(\mathbf{q}\mathbf{k}^T / \sqrt{d/h}), \quad \mathrm{CA}(f\left(\bar{\mathbf{c}}_i\right); \mathbf{V}) = \mathbf{A}\mathbf{v}, \end{gathered} \tag{6}$$

where $\mathbf{W}_q$, $\mathbf{W}_k$ and $\mathbf{W}_v$ are learnable parameters, d and h are the dimension and number of attention heads.

3.5 Prediction Unit

In the prediction unit, we score each candidate sense $s \in S_{c_i}$ for the target word c_i with dot product of $\tilde{\mathbf{c}}_i$ against gloss representation $\bar{\mathbf{s}}^s_{c_i}$ of each sense $s \in S_{c_i}$, described as:

$$score(c_i, s^j_{c_i}) = \tilde{\mathbf{c}}_i \cdot \bar{\mathbf{s}}^j_{c_i}, \tag{7}$$

where $j = 1, \ldots, |S_{c_i}|$ indicates the j-th candidate sense of the target word c_i. According to the scores of all candidate senses, we select the sense $s^h_{c_i}$ with the highest score as the right sense of the target word.

The training object $\mathcal{L}$ is to minimize the focal loss [13]:

$$\begin{aligned} pt &= -\mathrm{CE}(s^r_{c_i}, s^h_{c_i}), \\ \mathcal{L}(c_i, s^j_{c_i}) &= -\alpha(1-\exp(pt))^\gamma * pt, \end{aligned} \tag{8}$$

where $s^r_{c_i}$ represents the true sense of the target word c_i, CE denotes the cross-entropy function, α and γ are the balance parameter and focusing parameter, respectively.

4 Experiments

4.1 Dataset

Table 2. The details of all datasets

Dataset	Noun	Verb	Adj	Adv	Total
SemCor	87002	88334	31753	18947	226036
SE2	1066	517	445	254	2282
SE3	900	588	350	12	1850
SE07	159	296	0	0	455
SE13	1644	0	0	0	1644
SE15	531	251	160	80	1022

To verify the effectiveness of our proposed model, we employ the publicly available dataset SemCor and a representative open evaluation framework [23] to train and evaluate our model. The framework consists of five evaluation datasets including SensEval-2, SensEval-3, SemEval-07, SemEval-13 and SemEval-15, which are marked with SE2, SE3, SE07, SE13 and SE15, respectively. The details of all datasets are shown in Table 2. SemEval-07 dataset are chosen as our development set, the others are selected as evaluation sets. All sense glosses in our approach are retrieved from the widely-used WordNet 3.0.

4.2 Implementation Details

We utilize the pre-trained BERT to initialize our model, whose number of hidden layer is 768, the number of self-attention heads is 12 and the number of the Transformer blocks is 12. When fine-tuning our model, we use the SE07 as the development set to select the optimal hyperparameters. In the fine-tuned model, the dropout probability of CA is 0.1, the number of CA blocks is 5, the number of CA heads is 8, the balance parameter α and the focusing parameter γ are 0.2

and 0.5, respectively. As the limitation of the hardware condition, the batch size of the context encoder and gloss encoder are 1 and 128, respectively. The initial learning rate is 1e−5.

Table 3. Comparison with state-of-the-art models on F_1-score.

Models	Dev	Test datasets				Concatenation of test datasets				
	SE07	SE2	SE3	SE13	SE15	Noun	Verb	Adj	Adv	ALL
SVC [25]	69.5	77.5	77.4	76.0	78.3	79.6	65.9	79.5	85.5	76.7
EWISE [11]	67.3	73.8	71.1	69.4	74.5	74.0	60.2	78.0	82.1	71.8
LMMS [14]	68.1	76.3	75.6	75.1	77.0	78.0	64.0	80.5	83.5	75.4
GlossBERT [9]	72.5	77.7	75.2	76.1	80.4	79.8	67.1	79.6	**87.4**	77.0
GLU [8]	68.1	75.5	73.6	71.1	76.2	–	–	–	–	74.1
EWISER [4]	71.0	**78.9**	78.4	78.9	79.3	**81.7**	66.3	81.2	85.8	78.3
MEM	**74.9**	78.8	**78.6**	**79.3**	**82.3**	81.4	**69.2**	**83.2**	86.9	**79.1**

4.3 Comparison with the State-of-the-Art Baselines

We compare the performance of our model MEM against seven representative and/or state-of-the-art supervised models. These models include SVC [25], EWISE [11], LMMS [14], GlossBERT [9], GLU [8] and EWISER [4]. SVC exploits the semantic relations between senses to compress the sense vocabulary to reduce the number of sense tags to improve WSD performance. EWISE learns sense representation from a combination of sense-annotated data, gloss definition and lexical knowledge base to perform WSD. LMMS adopts the nearest neighbors algorithm on word representation produced by BERT to select the most suitable sense. GLossBERT utilizes BERT to jointly encode the context and glosses of the target word. GLU employs the pretrained contextualized word representation by BERT to improve WSD accuracy. Based on EWISE, EWISER further incorporates prior knowledge with synset embeddings, i.e., the explicit relational information from WordNet.

Table 3 shows F_1-score of our model and all baselines on dataset SemCor obtained by the public evaluation framework [23]. We observe that although our model is inferior to some baselines on SE2, MEM is still able to achieve the best performance on SE3, SE13, SE15, and ALL datasets. Here, ALL means the concatenation of all datasets. This shows that our model is superior to the baselines. The experimental results demonstrate that our model is effective on WSD task. Such satisfied performance of our model is attributed to its memory enhancement mechanism.

4.4 Ablation Study

Table 4. Comparison of ablation variants.

Ablation variants	Dev F_1-score	Δ
MEM	74.9	–
Del-Memory	73.8	−1.1
Only-Memory	74.1	−0.8
Del-CA	74.4	−0.5
Del-Update	74.5	−0.4

We perform an ablation study by comparing the standard MEM model with its four ablation variants: (a) Del-Memory: removes the memory component described in Eq. (3), the contextual representation will interact with the target word via cross-attention directly; (b) Only-Memory: removes the contextual representation described in Eq. (4), the memory representation $\bar{\mathbf{M}}$ will interact with the target word via cross-attention directly; (c) Del-CA: removes the CA mechanism described in Eq. (5); (d) Del-Update: removes the GRU component described in Eq. (3) to stop the update of the gloss representation stored in the memory.

The comparison result of the ablation study is shown in Table 4. According to the table, we have the following observations. First, the performance decrease of Del-Memory and Only-Memory demonstrates that the memory enhancement mechanism is critical for our model. Second, both Del-CA and Del-Update show inferior performance than the standard MEM model, which demonstrates that cross-attention component and memory updating mechanism are effective.

5 Conclusion

This paper proposes a novel model for word sense disambiguation based on memory enhancement mechanism (MEM). To the best of our knowledge, this is the first work to leverage memory mechanism to store the gloss knowledge of previously-identified words to assist the disambiguation of the next target word. Accordingly, we design an effective memory enhancement mechanism to enhance the representation of the target word with the identified glosses. Experimental results on real-world datasets demonstrate that our model outperforms the state-of-the-art models on word sense disambiguation task. This may provide a new perspective for utilizing memory mechanism and gloss knowledge to improve WSD methods.

Acknowledgment. The research work is partly supported by National Nature Science Foundation of China under Grant No. 61502259, and Key Program of Science and Technology of Shandong under Grant No. 2020CXGC 010901 and No. 2019JZZY020124.

References

1. Agirre, E., de Lacalle, O.L., Soroa, A.: Random walks for knowledge-based word sense disambiguation. Comput. Linguist. **40**(1), 57–84 (2014)
2. Banerjee, S., Pedersen, T.: An adapted Lesk algorithm for word sense disambiguation using WordNet. In: International Conference on Intelligent Text Processing and Computational Linguistics, pp. 136–145 (2002)
3. Basile, P., Caputo, A., Semeraro, G.: An enhanced Lesk word sense disambiguation algorithm through a distributional semantic model. In: Proceedings of the 25th International Conference on Computational Linguistics, pp. 1591–1600 (2014)
4. Bevilacqua, M., Navigli, R.: 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, pp. 2854–2864 (2020)
5. Blevins, T., Zettlemoyer, L.: 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, pp. 1006–1017 (2020)
6. Chen, C.F., Fan, Q., Panda, R.: CrossVIT: cross-attention multi-scale vision transformer for image classification. arXiv preprint arXiv:2103.14899 (2021)
7. Emelin, D., Titov, I., Sennrich, R.: Detecting word sense disambiguation biases in machine translation for model-agnostic adversarial attacks. In: The 2020 Conference on Empirical Methods in Natural Language Processing, pp. 7635–7653. Association for Computational Linguistics (2020)
8. Hadiwinoto, C., Ng, H.T., Gan, W.C.: Improved word sense disambiguation using pre-trained contextualized word representations. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 5297–5306 (2019)
9. Huang, L., Sun, C., Qiu, X., Huang, X.J.: 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, pp. 3509–3514 (2019)
10. Kågebäck, M., Salomonsson, H.: Word sense disambiguation using a bidirectional LSTM. In: COLING, pp. 51–56 (2016)
11. Kumar, S., Jat, S., Saxena, K., Talukdar, P.: Zero-shot word sense disambiguation using sense definition embeddings. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 5670–5681 (2019)
12. Le, H., et al.: FlauBERT: unsupervised language model pre-training for French. In: Proceedings of the 12th Language Resources and Evaluation Conference, pp. 2479–2490 (2020)
13. Lin, T., Goyal, P., Girshick, R., He, K., Dollar, P.: Focal loss for dense object detection. IEEE PAMI **42**(2), 318–327 (2018)
14. Loureiro, D., Jorge, A.: 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, pp. 5682–5691 (2019)
15. Luo, F., Liu, T., He, Z., Xia, Q., Sui, Z., Chang, B.: 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, pp. 1402–1411 (2018)

16. Luo, F., Liu, T., Xia, Q., Chang, B., Sui, Z.: Incorporating glosses into neural word sense disambiguation. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pp. 2473–2482 (2018)
17. Maru, M., Scozzafava, F., Martelli, F., Navigli, R.: SyntagNET: challenging supervised word sense disambiguation with lexical-semantic combinations. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 3534–3540 (2019)
18. Miller, A., Fisch, A., Dodge, J., Karimi, A.H., Bordes, A., Weston, J.: Key-value memory networks for directly reading documents. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 1400–1409 (2016)
19. Moreo, A., Esuli, A., Sebastiani, F.: Word-class embeddings for multiclass text classification. Data Min. Knowl. Disc. **35**(3), 911–963 (2021). https://doi.org/10.1007/s10618-020-00735-3
20. Moro, A., Raganato, A., Navigli, R.: Entity linking meets word sense disambiguation: a unified approach. Trans. Assoc. Comput. **2**, 231–244 (2014)
21. Pasini, T., Navigli, R.: Train-O-Matic: supervised word sense disambiguation with no (manual) effort. Artif. Intell. **279**, 103215 (2020)
22. Raganato, A., Bovi, C.D., Navigli, R.: Neural sequence learning models for word sense disambiguation. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 1156–1167 (2017)
23. Raganato, A., Camacho-Collados, J., Navigli, R.: 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, pp. 99–110 (2017)
24. Tripodi, R., Navigli, R.: Game theory meets embeddings: a unified framework for word sense disambiguation. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 88–99 (2019)
25. Vial, L., Lecouteux, B., Schwab, D.: Sense vocabulary compression through the semantic knowledge of WordNet for neural word sense disambiguation. In: Proceedings of the 10th Global WordNet Conference, pp. 108–117 (2019)
26. Wang, M., Wang, Y.: 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, pp. 5218–5229 (2021)