Knowledge Injection with Perturbation-based Constrained Attention Network for Word Sense Disambiguation

Fumiyo Fukumoto
Interdisciplinary Graduate School
University of Yamanashi
4-3-11, Takeda, Kofu, 400-8511, Japan
fukumoto@yamanashi.ac.jp

Shou Asakawa
Graduate School of Engineering
University of Yamanashi
4-3-11, Takeda, Kofu, 400-8511, Japan
G20TK002@yamanashi.ac.jp

Abstract

Supervised Word Sense Disambiguation (WSD) has been studied intensively for over three decades. However, disentangling diverse contexts is still a challenging problem. This paper addresses the problem and proposes a Perturbation-based constrained attention network (Pconan) for injecting lexical knowledge derived from the WordNet. The Pconan allows modeling beneficial dependencies between the segments/words within the input sequence with the mask-attention technique. We incorporate a perturbation method into our model to mitigate the overfitting problem resulting from intensive learning. The experimental results by using a benchmark dataset show that our method is comparable to the SOTA WSD methods. Our source codes are available online[1].

1 Introduction

Computational lexicons such as WordNet (George A. Miller and Miller, 1990) and ACQUILEX (Edward, 1991) have been popular knowledge resources for NLP tasks. There is a large body of WSD work based on neural networks that leverage rich information derived from these resources (Luo et al., 2018b; Vial et al., 2019; Kumar and Talukdar, 2019; Huang et al., 2019; Blevins and Zettlemoyer, 2020; Bevilacqua and Navigli, 2020; Conia and Navigli, 2021). They demonstrated that the external knowledge base is beneficial to disambiguate senses. However, it is often the case that the inputs are long sequences. It hampers WSD attempts with external knowledge. Several authors have attempted to alleviate the issue. (Blevins and Zettlemoyer, 2020) independently embedded the target word with its surrounding contexts and the dictionary definition of each sense. (Bevilacqua and Navigli, 2020) extended (Blevins and Zettlemoyer, 2020) method and integrated relational knowledge into the architecture through a simple additional sparse dot product operation. Their results by using a benchmark dataset were beyond 80%.

The attention mechanism is also one of the major techniques to capture long-term dependencies on their sequence (Vaswani et al., 2017). (Luo et al., 2018a) introduced a co-attention mechanism to generate co-dependent representations to capture both word- and sentence-level information. Their assumption is that lexical knowledge such as gloss sentences and context sentences can help each other to highlight the important words within these sentences, while the sense definition candidates do not all at once take into account during the training process. Several authors focused on the issue (Wang and Wang, 2021). (Barba et al., 2021b) proposed a joint-learning that learns the input context and target word definitions jointly. Subsequently, (Barba et al., 2021a) attempted to process the disambiguation of a target word to be conditioned not only on its context but also on the explicit senses assigned to the surrounding words, while their model has not leveraged external lexical knowledge.

Inspired by the previous work mentioned above, we propose a method to inject lexical knowledge, i.e. example sentences from WordNet to effectively learn a context sentence and lexical knowledge simultaneously. Our model called Perturbation-based constrained attention network (Pconan) allows the modeling of dependencies between the segments/words within the input sequence with the mask-attention technique. The technique makes it possible to concentrate on learning beneficial dependencies only, entirely discarding the others. However, this causes an overfitting problem, especially when the available training data is limited. To alleviate this issue, the Pconan utilizes the perturbation technique (Sato et al., 2019). More specifically, we add noises to the training data and the model learns sense distinctions of the same word by using these noisy data to assign the correct label.

[1] https://github.com/fukumoto-lab/Pconan

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023, pages 171–177
November 1–4, 2023. ©2023 Asian Federation of Natural Language Processing

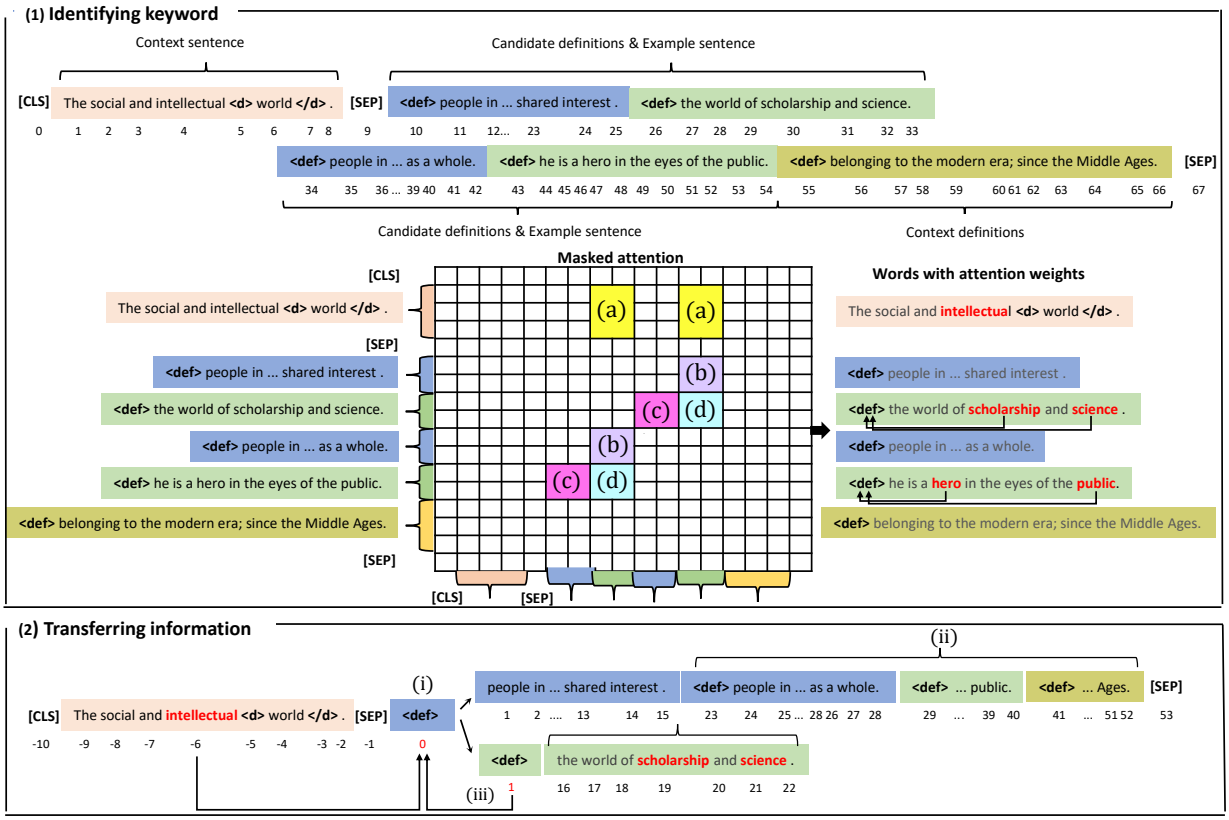

Figure 1: Identifying keyword and transferring information

2 Framework

2.1 The WSD task definition

For each of the candidate definitions (CDs), our model utilizes an example sentences (ESs) provided by the WordNet.

Let $\mathcal{SI}_{tw(i)}$ be a sense inventory, $tw(i)$ be a current, i-th target word appeared in the context sentence c. Let also $\mathcal{D}_{tw(i)}$ = <def>$tw(i)_1^{g_1} \cdots tw(i)_{|g_1|}^{g_1}$<def>$es(i)_1^{s_1} \cdots es(i)_{|s_1|}^{s_1}$ <def>$tw(i)_1^{g_N} \cdots tw(i)_{|g_N|}^{g_N}$ <def> $es(i)_1^{s_N} \cdots es(i)_{|s_N|}^{s_N}$ be the CDs for $tw(i)$, along with ESs, where $tw(i)_j^{g_k}$ and $es(i)_j^{s_k}$ be the j-th word of the k-th CD g_k, and ES s_k ($1 \leq k \leq N$), respectively. N is the total number of CDs and <def> stands for the start of each segment. Let also $\widetilde{\mathcal{D}} = \widetilde{s}_1, \cdots, \widetilde{s}_{i-1}$ be a sequence concatenating the context definitions of the senses previously assigned to $\widetilde{w}_1, \cdots, \widetilde{w}_{i-1}$.

For a given c including <d>$tw(i)$</d>, we create an input sequence [CLS] c [SEP] $\mathcal{D}_{tw(i)}$ $\widetilde{\mathcal{D}}$ [SEP]. Here, the inputs to each encoder are padded with DEBERTa-specific start and end symbols: [CLS] and [SEP] (He et al., 2021). The goal of the WSD task is for the input sequence, to find the correct definition $\widetilde{g} \in \mathcal{SI}_{tw(i)}$.

2.2 Constrained Attention Network

Our model applies (1) identifying keywords, and (2) transferring information to learn relevant contextual features for WSD. The top of Figure 1 illustrates an example of the input sequence X, i.e. [CLS] c [SEP] $\mathcal{D}_{tw(i)}$ $\widetilde{\mathcal{D}}$ [SEP].

For the input sequence X = [CLS] c [SEP] $\mathcal{D}_{tw(i)}$ $\widetilde{\mathcal{D}}$ [SEP], we apply the so-called hard attention technique (Xu et al., 2015; Shen et al., 2018) that a model concentrates solely on learning beneficial dependencies to identify keywords in the sequence, entirely discarding the others. The middle picture of Figure 1 illustrates masked attention for a bi-dimensional matrix. The words aligned on the horizontal axis are heads, and those aligned on the vertical axis are dependents. As illustrated in Figure 1, we discard some segment pairs, each of which consists of a head and dependence on the sequence by masking them as these pairs are not semantically related to each other and do not include keywords that are beneficial to identify the sense of the target word. Table 1 shows pairs that we masked. For example, (b) pairs of candidate definitions and example sentences which do not correspond to each other are masked (purple box

	Head	Dependent
(a)	Context sentence	Example sentence
(b)	Candidate definition	Different example sentence
(c)	Example sentence	Different candidate definition
(d)	Example sentence	Different example sentence

Table 1: Masked attention between head and dependent.

in Figure 1). The output is a sequence of words with attention weights. In Figure 1, keywords having high weight values are marked in red. From the result of keyword identification, for each candidate definition, we created a sequence for the target candidate definition starting from <def> as follows:

- Context sentence segment perceived as immediately before the CD segment. ((i) in Figure 1)
- <def> of the target CD has two branches. One is that it witnesses other CDs, their ES, and context definitions. ((ii) in Figure 1) Another is that its ES. Relative position of the special token <def> is set to 1. ((iii) in Figure 1)

As shown in "(2) Transferring information" of Figure 1, the structure provides a method to leverage the relative positions of <def> including keywords representation to learn a model more accurately.

2.3 Model Architecture

Figure 2 illustrates an overview of our model. Let $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_L] \in \mathbb{R}^{d \times L}$ refers to the concatenation of word/token embeddings for the input sequence X, and $\mathbf{E}_{rpm} \in \mathbb{R}^{d \times L}$ be relative position matrix of X. Here, $\mathbf{e}_k$ and its relative position is obtained by DEBERTa encoding. d refers to the dimension of embedding, and L is the number of words/tokens in X. We further utilize a special symbol in X so that the model can capture the difference between the context sentence and others. Let also $\mathbf{r}_k \in \mathbb{R}^d$ be a perturbation vector for the k-th word x_k in the input X. The perturbed input embedding $\hat{\mathbf{e}}_k$ is computed based on the stochastic gradient descent as follows:

$$\begin{aligned} \hat{\mathbf{e}}_k &= \mathbf{e}_k + \mathbf{r}_k, \\ \mathbf{r}_k &= \epsilon \frac{\mathbf{c}_k}{||\mathbf{c}_k||}, \\ \mathbf{c}_k &= \bigtriangledown_{\hat{\mathbf{e}}_k} \mathcal{L}_1(\theta), \end{aligned} \tag{1}$$

where ϵ refers to a hyperparameter that controls the norm of the perturbation and $\mathcal{L}_1(\theta)$ indicates cross-entropy loss which is given by:

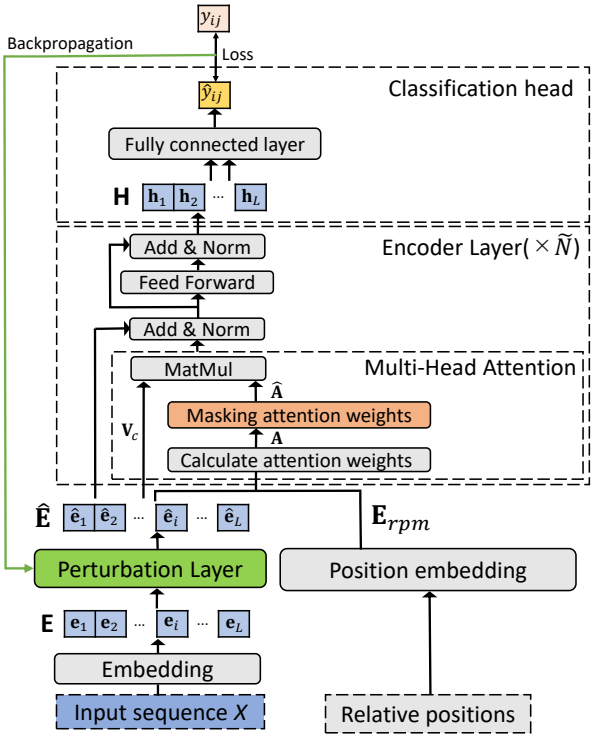

Figure 2: Pconan model architecture.

$$\mathcal{L}_1(\theta) = -\sum_{i=1}^{M} \sum_{j=1}^{N_i} y_{ij} \log \hat{y}_{ij}, \tag{2}$$

where M is the total number of target words in the training data, and N_i is the word sense number of the i-th target word, y_{ij} and $\hat{y}_{ij}$ are true and predicted probability of the i-th target word that belongs to the j-th candidate definition. As shown in Figure 2, for each embedding $\mathbf{e}_k$, we apply Eq.(1) and obtain perturbed sequence. Let $\hat{\mathbf{E}} = [\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \cdots, \hat{\mathbf{e}}_L] \in \mathbb{R}^{d \times L}$ be the concatenation of perturbed sequence. Our constrained attention networks aim to learn relevant contextual keywords for WSD, and finally output attention weights $\mathbf{A}$ for the inputs, $\hat{\mathbf{E}}$ and $\mathbf{E}_{rpm}$. We further obtain $\hat{\mathbf{A}}$ by applying the mask-attention procedure to $\mathbf{A}$. $\hat{\mathbf{E}}$ is linearly projected and we obtain $\mathbf{V}_c$. We multiply $\mathbf{V}_c$ and $\hat{\mathbf{A}}$ by matrix multiplication. Keyword information is transferred by this operation. The result is fed into a feed-forward network, combined with layer normalization and residual connection. Each encoder layer takes the output of the previous layer as input and the number of layers is $\widetilde{N}$. We obtain the matrix $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_L] \in \mathbb{R}^{d \times L}$ as an output of the encoder. Each <def> vector that corresponds to the start of the candidate definition

is extracted from the matrix $\mathbf{H}$, passed to the fully connected layer and finally, we obtain the probability score $\hat{y}_{ij}$ by the softmax function. The final loss $\mathcal{L}(\theta)$ obtained by our model is given by:

$$\begin{aligned}\mathcal{L}(\theta) &= \mathcal{L}_1(\theta) + \alpha\mathcal{L}_2(\theta), \\ \mathcal{L}_2(\theta) &= \frac{1}{|\mathcal{D}|}\sum_{\mathcal{D}} \mathrm{KL}(p(\cdot|X;\theta)||p(\cdot|X,\mathbf{r};\theta)),\end{aligned} \tag{3}$$

where $\mathcal{D}$ refers to the number of training instances. α indicates a hyperparameter. $\mathrm{KL}(p||q)$ denotes KL-divergence between distributions p and q, and $\mathbf{r}$ shows a concatenated vector of $\mathbf{r}_k$ for all x_k ($1 \leq k \leq L$). We train the whole architecture to minimize $\mathcal{L}(\theta)$. Similar to (Barba et al., 2021a) approach, at training time, we use teacher forcing on the context definitions, and at prediction time, we use a greedy decoding strategy and the model deems $\widetilde{g}$ as the most likely definition for a current target word.

3 Experiments

3.1 Dataset

We performed the experiments on English all-words fine-grained WSD datasets (Alessandro Raganato, 2017b), using SemCor (George A. Miller and Bunker, 1993) as the training corpus. The datasets are Senseval/SemEval data consisting of Senseval-2 (SE2) (Edmonds and Cotton, 2001), Senseval-3 (SE3) (Snyder and Palmer, 2004), SemEval-07 (SE07) (Sameer Pradhan and Palmer, 2007), SemEval-13 (SE13) (Roberto Navigli and Vannella, 2013), and SemEval-15 (SE15) (Moro and Navigli, 2015). Similar to other related work, we chose SE07 as the development set.

3.2 Model settings and evaluation metrics

We utilized the hyperparameters with the best performance on SE07 as follows: The dimension of word embedding d was 1,024. The number of maximum words per batch was 1,536. The gradient accumulation and the maximum number of steps were 8.0 and 25,000, respectively. The number of layer $\widetilde{N}$ of DEBERTa was 24 and the learning rate was 3e-6. The initial perturbation was set to 1e-2 and the ϵ value in Eq.(1) was 3e-6. α in Eq. (3) was set to 1.0. We used Rectified Adam as optimizer (Weijie Liu, 2020). The experiments were conducted by using Pytorch on Nvidia GeForce RTX A6000 (48GB memory). We used the F1-score following (Alessandro Raganato, 2017b).

3.3 Comparison Models

We compared our model with the SOTA methods; MSF-SemCor as a frequency based approach, **SVC** (Vial et al., 2019), **GlossBERT** (Huang et al., 2019), **ARES** (Bianca Scarlini, 2020), **EWISER** (Bevilacqua and Navigli, 2020), **BEM** (Blevins and Zettlemoyer, 2020), **WMLC** (Conia and Navigli, 2021), **HCAN** (Luo et al., 2018a), and **KELESC** (Zhang et al., 2022) as a knowledge source integration approach, **SACE** (Wang and Wang, 2021), **ESCHER** (Barba et al., 2021b), and **ConSec** (Barba et al., 2021a) as a joint learning approach.

3.4 Results

The results are summarized in Table 2. The performance of joint-learning approaches was better than those of frequency-based and knowledge source integration approaches in all test sets and part-of-speech (POS) patterns. This indicates that the model learned the input context and target word definitions jointly are effective for disambiguation. Our model was statistically significant compared with the second-best method for test sets and POS patterns except for SE3, 15, Adj, and Adv.

3.5 Ablation study

We conducted ablation studies to empirically examine our mask-attention technique and perturbation (Prtb). Table 3 shows the results. When we did not utilize the mask-attention and perturbation procedures, the F1-score was 82.1% which is no significant difference compared with ConSec (82.0%) even though ESs are injected. When we applied the mask-attention technique, the improvement was 0.5% at maximum, 82.6%. Among masked attentions, there is also no statistically significant difference between the masked attention (a) context sentence and ES pairs (82.5%) and the combination of (b)$\sim$ (d) (82.3%). However, we gained 0.6% improvement by using (a) $\sim$ (d) and further gained 0.4% improvement by perturbation. From these observations, we can conclude that the perturbation approach helps the mask-attention procedure to boost the WSD task performance.

3.6 Qualitative analysis of errors

We performed an error analysis to provide feedback for further improvement of our method. The number of errors for each POS was 590 noun words, 420 for a verb, 120 for an adjective, and 38 for an adverb, 1,168 words in all. The average senses

	Model	Dev Set	Test Sets				Concatenation of all Datasets				
		SE7	SE2	SE3	SE13	SE15	Noun	Verb	Adj	Adv	ALL
SemCor	MFS-SemCor	54.5	65.6	66.0	63.8	67.1	67.7	49.8	73.1	80.5	65.5
	SVC (hypernyms)	69.5	77.5	77.4	76.0	78.3	79.6	65.9	79.5	85.5	76.7
	GlossBERT	72.5	77.7	75.2	76.1	80.4	79.8	67.1	79.6	87.4	77.0
	ARES	71.0	78.0	77.1	78.7	75.0	80.6	68.3	80.5	83.5	77.9
	EWISER	71.0	78.9	78.4	78.9	79.3	81.7	66.3	81.2	85.8	78.3
	BEM	74.5	79.4	77.4	79.7	81.7	81.4	68.5	83.0	87.9	79.0
	WMLC	72.2	78.4	77.8	76.7	78.2	80.1	67.0	80.5	86.2	77.6
	HCAN	-	72.8	70.3	68.5	72.8	72.7	58.2	77.4	84.1	71.1
	KELESC	76.7	82.2	78.1	82.2	83.0	84.3	69.4	84.0	86.7	81.2
	SACE	76.3	82.4	**81.1**	82.5	83.7	84.1	72.2	**86.4**	**89.0**	81.9
	ESCHER	76.3	81.7	77.8	82.2	83.2	83.9	69.3	83.8	86.7	80.7
	ConSec	77.4	82.3	79.9	83.2	**85.2**	85.4	70.8	84.0	87.3	82.0
	Pconan	**79.8**	**83.8**	**81.1***	**83.9**	84.7*	**85.6**	**73.8**	84.8	**89.0***	**83.0**

Table 2: Performance comparison: The best score is in boldface and the second best is underlined. $*$ denotes the method (if any) whose score is not statistically significant compared to the best one. We used a t-test, p-value < 0.01.

Model	ALL
ConSec	82.0
w/o Prtb & w/o (a), (b), (c), and (d)	82.1*
w/o Prtb & w/o (b), (c), and (d)	82.5*
w/o Prtb & w/o (a)	82.3*
w/o Prtb	82.6
Pconan	83.0

Table 3: Ablation test over the Pconan components: "w/o Prtb" refers to the result without pertubation. $*$ denotes the method whose score is not statistically significant compared to the baseline, ConSec.

[Context sentence]
Skilled ringers use their wrists to <d> advance </d> or retard the next swing so that one bell can swap places with another in the following change.
[Candidate sense & definition & example sentence]
advance#1 <def> **move forward**, also in the metaphorical sense. <def> Times marches on.
advance#5 <def> cause to **move forward** <def> Can you **move** the car seat **forward**?

Figure 3: An example with similar context definitions: The correct sense in <d> advance <d> is advance#5.

for these POS words were 6.7 for nouns, 12.2 for verbs, 5.9 for adjectives, and 5.2 for adverbs. We randomly picked up 100 words from 1,168 and found that there are mainly two types of errors:

1. Sense distribution: When the sense distribution of the target word in the training data is unbalanced, most of the target words tend to be assigned to the sense of having much training instances. This was the most frequent error type and 51 words were classified into this type.

2. The similarity between candidate definitions: When words that appear one candidate definition are semantically similar or the same as those of other candidate definitions, it is difficult to model beneficial dependencies to identify keywords. For example, in Figure 3, as "move forward" appears in both candidate definitions and example sentence, only a few words such as "car" and "seat" are clues to predict beneficial dependencies which causes an error. 26 words were classified into this type.

We focused on example sentences extracted from WordNet as lexical knowledge. (Vial et al., 2019; Conia and Navigli, 2021) utilized the semantic relationships between senses such as synonymy, hypernymy, and hyponymy derived from the WordNet and reported that the external knowledge contributes to improving WED performance. This is definitely worth trying with Pconan.

4 Conclusion

We presented WSD approach for injecting lexical knowledg from the WordNet with perturbation-based constrained attention network. The comparative results with the SOTA WSD methods showed the effectiveness of our method. Future work will include: (i) evaluating our model by using other lexical knowledge such as the semantic relationship between senses, (ii) investigating other perturbation techniques (Gal and Ghahramani, 2016; Wang and Wang, 2021) to improve the performance, and (iii) applying methods (Liu et al., 2020; Xiong et al., 2021) to reduce the overall self-attention complexity for further advantages in efficacy.

Acknowledgments

We would like to thank anonymous reviewers for their comments and suggestions. This work is supported by SCAT, JKA, Kajima Foundation's Support Program, and JSPS KAKENHI No.23H03402.

References

Roberto Navigli Alessandro Raganato, Claudio Delli Bovi. 2017b. Neural sequence learning models for word sense disambiguation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, page 1156–1167.

Edoardo Barba, Luigi Procopio, and Roberto Navigli. 2021a. ConSeC: Word sense disambiguation as continuous sense comprehension. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1492–1503.

Edonardo Barba, Tommaso Pasini, and Navigli Roberto. 2021b. Esc: Redesigning wsd with extractive sense comprehension. In *Proceedings of the 2021 Conference on the North American Chappter of the Association for Computational Linguistics: Human Language Technologies*, pages 4661–4672.

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.

Roberto Navigli Bianca Scarlini, Tommaso Pasini. 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)*, page 3528–3539.

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 Association for Computational Linguistics*, pages 1006–1017.

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*, pages 3269–3275.

Philip Edmonds and Scott Cotton. 2001. Senseval2: Overview. In *Proceedings of SENSEVAL2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, page 1–5.

Briscoe Edward. 1991. Lexical issues in natural language processing. In *Proceedings of the Symposium on Natural Language and Speech*, pages 36–68.

Yarin Gal and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 1027–1035.

Christiane D. Fellbaum D. Gross George A. Miller, R.T. Beckwith and K. Miller. 1990. Introduction to wordnet: An online lexical database. In *International Journal of Lexicography, 3(4)*, pages 235–244.

Randee Tengi George A. Miller, Claudia Leacokc and Ross T Bunker. 1993. A semantic concordance. In *Proceedings of the workshop on Human Language Technology*, pages 303–308.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations*.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing 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)*, pages 3507–3512.

Sharmistha Jat Karan Saxena Kumar, Sawan 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.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2020. On the variance of the adaptive learning rate and beyond. In *Proceedings of the Eighth International Conference on Learning Representations (ICLR 2020)*.

Fuli Luo, Tianyu Liu, Zexue He, Qiaolin Xia, Zhifang Sui, and Baobao Chang. 2018a. 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*, pages 1402–1411.

Fuli Luo, Tianyu Liu, Qiaolin Xia, Baobao Chang, and Zhifang Sui. 2018b. 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.

Andrea Moro and Roberto Navigli. 2015. Semeval2015 task 13: Multilingual all-words sense disambiguation and entity linking. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, page 288–297.

David Jurgens Roberto Navigli 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)*, page 222–231.

Dmitriy Dligach Sameer Pradhan, Edward Loper 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)*, page 87–92.

Motoki Sato, Jun Suzuki, and Shun Kiyono. 2019. Effective adversarial regularization for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 204–210.

Tao Shen, Tianyi Zhous, Guodong Long, Jing Jiang, Sen Wang, and chengqi Zhang. 2018. Reinforced self-attention network: a hybrid of hard and soft attention for sequence modeling. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*, pages 4345–4352.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, page 6000–6010.

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*, pages 108–117.

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)*, pages 5218–5229.

Zhe Zhao Zhiruo Wang Qi Ju Haotang Deng Ping Wang Weijie Liu, Peng Zhou. 2020. K-BERT: Enabling language representation with knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2901–2908.

Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. 2021. Nyströmformer: A nystöm-based algorithm for approximating self-attention. In *Proceedings of the Association for the Advancement of Artificial Intelligence*, pages 14138–14148.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Couville, Ruslan Salakhudinov, Rich Zemei, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *32nd International Conference on Machine Learning*, pages 2048–2057.

Guobiao Zhang, Wenpeng Lu, Zueping Peng, Shoujin Wang, Baoshuo Kan, and Rui Yu. 2022. Word sense disambiguation with knowledge-enhanced and local self-attention-based extractive sense comprehension. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4061–4070.