GX at SemEval-2021 Task 2: BERT with Lemma Information for MCL-WiC Task

Wanying Xie
Beijing Language and Culture University, China
xiewanying07@gmail.com

Abstract

This paper presents the GX system for the Multilingual and Cross-lingual Word-in-Context Disambiguation (MCL-WiC) task. The purpose of the MCL-WiC task is to tackle the challenge of capturing the polysemous nature of words without relying on a fixed sense inventory in a multilingual and cross-lingual setting. To solve the problems, we use context-specific word embeddings from BERT to eliminate the ambiguity between words in different contexts. For languages without an available training corpus, such as Chinese, we use neural machine translation model to translate the English data released by the organizers to obtain available pseudo-data. In this paper, we apply our system to the English and Chinese multilingual setting and the experimental results show that our method has certain advantages.[1]

1 Introduction

In recent years, contextual embeddings have drawn much attention. The approaches of calculating contextual embeddings include multi-prototype embeddings, sense-based and contextualized embeddings (Camacho-Collados and Pilehvar, 2018). However, it is not easy to evaluate such multiple embedding methods in one framework. Pilehvar and Camacho-Collados (2019) present a large-scale word in context dataset to focus on the dynamic semantics of words. Following and expanding them, the MCL-WiC task (Martelli et al., 2021) performs a binary classification task that indicates whether the target word is used with the same or different meanings in the same language (multilingual data set) or across different languages (cross-lingual data set). Besides, it is the first SemEval task for Word-in-Context disambiguation (Martelli et al., 2021).

A typical solution to the problems is obtaining context-specific word embeddings, such as Context2vec (Melamud et al., 2016) and Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019). BERT is designed to pre-train deep bidirectional representation in unlabeled texts by jointly conditioning in the left and right contexts of all layers. Due to its powerful capabilities and easy deployment, we use BERT as our major system and fine-tune on the training data released by the organizer to get the context-specific word embeddings.

In this paper, we participate in the sub-task of multilingual settings in English and Chinese. The organizer only provides English training data, and we fine-tune the pre-trained English BERT model based on this data. For Chinese tasks where no training set is available, we train a satisfactory neural machine translation (NMT) model to translate the English training set into Chinese and then fine-tune the Chinese BERT model based on the pseudo-data. The experimental results show that our method achieves 82.7% in English multilingual setting and 76.7% in Chinese multilingual setting.

2 Background

In this section, we will briefly introduce the word-in-context task and the structure of BERT for the sentence pair classification task.

2.1 Word-in-Context

The MCL-WiC task (Martelli et al., 2021) expands the Word-in-Context (WiC) (Pilehvar and Camacho-Collados, 2019) task to be multilingual and cross-lingual settings. For WiC, each instance has a target word *lemma*, which provides it with two contexts. Each context triggers the specific meaning of the word *lemma*. The task is to identify whether *lemma* in two contexts corresponds to

[1] Reproducible code: https://github.com/yingwaner/bert4wic

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 706–712
Bangkok, Thailand (online), August 5–6, 2021. ©2021 Association for Computational Linguistics

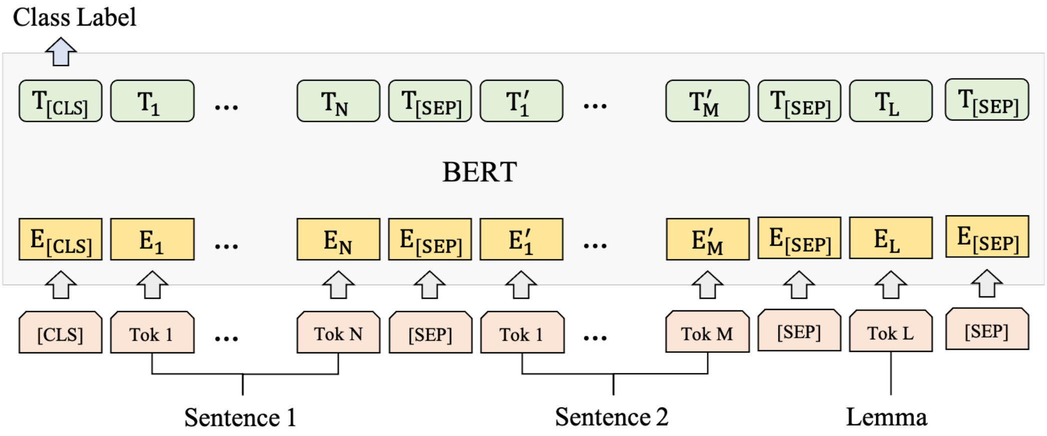

Figure 1: Overall fine-tuning procedures for our system. The token Lemma is the target word that require the system to judge whether it has the same meaning in two sentences.

the same meaning, which has been widely investigated in recent years. Wiedemann et al. (2019) perform word sense disambiguation models with contextualised representations. Hu et al. (2019) prove that the supervised derivation of time-specific sense representation is useful. Giulianelli et al. (2020) present an unsupervised approach to lexical-semantic change that makes use of contextualized word representations. Loureiro and Jorge (2019) compute sense embeddings and the relations in a lexical knowledge base. Scarlini et al. (2020) drop the need for sense-annotated corpora so as to collect contextual information for the senses in WordNet.

2.2 BERT

Neural contextualized lexical representation has been widely used in natural language processing, which benefits from deep learning model in optimizing tasks while learning usage dependent representations, such as ULMFiT (Howard and Ruder, 2018), ELMo (Peters et al., 2018), GPT (Radford et al., 2018, 2019), and BERT (Devlin et al., 2019). BERT is pre-trained by two unsupervised tasks: masked LM task, which is simply masking some percentage of the input tokens at random, and then predicting those masked tokens; and next sentence prediction task, which is whether the next sentence in the sentence pair is the true next sentence. In the fine-tuning phase, task-specific inputs and outputs are plugged into the BERT and all parameters are fine-tuned end-to-end.

The architecture of BERT is a multi-layer bidirectional Transformer encoder based on the original implementation described in Vaswani et al. (2017). There are several encoder layers in the BERT model. For a single layer in Transformer encoder, it consists of a multi-head self-attention and a position-wise feed-forward network. Specifically, there are specialized input and output formats for different downstream tasks. For language pair classification task, the input format is [CLS] + Sentence 1 + [SEP] + Sentence 2 + [SEP]. At output layer, the [CLS] representation is fed into an output layer for the classification task, such as entailment, sentiment analysis, and the word-in-context disambiguation task.

3 System Overview

Systems proposed for both English and Chinese multilingual settings were based on BERT model (Devlin et al., 2019) with task-specific input modifications. We participate in the multilingual setting and divide the system into two parts according to the language: English setting and Chinese setting.

3.1 English Setting

Following Devlin et al. (2019), we initialize our model with the well pre-trained model, which has been trained on the large-scale data set and obtained the general knowledge. Then we fine-tune the model on the English parallel sentences released by the organizers.

Model Architecture The model architecture in the fine-tuning stage is shown in Figure 1. On the basis of the original BERT input, *lemma* token is added, which is the target word that needs the system to judge whether it has the same meaning

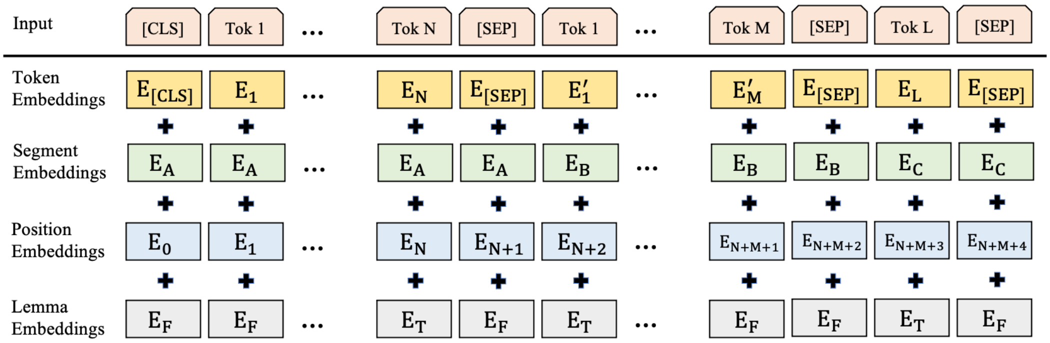
Input
[CLS]
Tok 1
...
Tok N
[SEP]
Tok 1
...
Tok M
[SEP]
Tok L
[SEP]
Token Embeddings
$E_{[CLS]}$ E_1 ... E_N $E_{[SEP]}$ E'_1 ... E'_M $E_{[SEP]}$ E_L $E_{[SEP]}$
Segment Embeddings
E_A E_A ... E_A E_A E_B ... E_B E_B E_C E_C
Position Embeddings
E_0 E_1 ... E_N E_{N+1} E_{N+2} ... E_{N+M+1} E_{N+M+2} E_{N+M+3} E_{N+M+4}
Lemma Embeddings
E_F E_F ... E_T E_F E_T ... E_F E_F E_T E_F

Figure 2: Our system input representation. The input embeddings are the sum of the token embeddings, the segment embeddings, the position embeddings, and the lemma embeddings. It is only at the position of the lemma tokens that the lemma embedding is E_T. In this example, we assume that Tok N in sentence 1 and Tok 1 in sentence 2 are also lemma tokens.

in the sentence pair. For instance, given the sentence pairs ’They opposite to the policies designed to *tackle* inflation.’ and ’I *tackled* him about his heresies.’, the standard input format is:

[CLS] They opposite to the policies designed to tackle inflation . [SEP] I tackled him about his heresies . [SEP] tackle [SEP]
where *tackle* is the lemma token, which is also the word that needs to be judged by the system whether it has the same meaning in two sentences. In this way, we emphasize the target word so that the output $\mathrm{T}_{[\mathrm{CLS}]}$ of the output layer can express whether the lemma token is synonymous in the two sentences.

Input Representation In addition, we also made some modifications for the input representation, which is made up of the sum of the corresponding token, segment, and position embeddings according to Devlin et al. (2019). The input representation of our system is shown in Figure 2. We adjust the segment embeddings of the lemma token to further emphasize the importance of the target word in the whole sentence pair, which is represented as E_C. Moreover, we introduce lemma embeddings in the input representation. Lemma embeddings are similar to segment embeddings, but segment embeddings are to distinguish between sentence 1 and sentence 2, while lemma embeddings are to distinguish between the position of lemma tokens and the position of other tokens. Only the lemmas in sentence 1 and sentence 2 and the final lemma Tok L will be marked E_T, and the other positions will be marked E_F, that is, for a training example, there will be three lemma markers E_T in lemma embeddings. In this way, we enhance the relationship of lemma tokens to make them more closely connected, and at the same time highlight and emphasize the position and importance of lemma tokens, so that the final output can obtain enough lemma token information.

3.2 Chinese Setting

The multilingual setting in Chinese is more difficult because there is no available training data in Chinese, so it is not possible to fine-tune the pre-trained BERT model. In order to solve this problem, we introduce neural machine translation method.

Neural Machine Translation Due to the superior performance of Transformer, we use it as our neural machine translation model. We first train an English-to-Chinese translation model on an open-source dataset and evaluate its performance to ensure that it has sufficient translation quality. Then, we use this translation model to translate sentence 1 and sentence 2 from the English training set released by the organizer into Chinese, respectively, and regard the generated sentences as the training data of Chinese MCL-WiC task. Finally, we use the pre-trained Chinese BERT model to fine-tune this generated data set to get our final model.

Model Architecture The system in Chinese setting is somewhat different from the system in English setting in that there is no lemma token. We use machine translation to translate English training data into Chinese, because every token in a sentence has a context, so sentence to sentence translation does not change the meaning of the whole sentence much. However, a lemma token has no context, so it is difficult for translation model to choose which token to translate into the target

	English				Chinese			
	Valid	Δ	Test	Δ	Valid	Δ	Test	Δ
Fine-tuning	81.2	-	81.6	-	67.9	-	76.7	-
+Lemma Token	81.8	0.6	82.1	0.5	67.1	-0.8	76.0	-0.7
+Segment $\mathrm{E_C}$	82.2	0.4	82.4	0.3	67.3	0.2	76.3	0.3
+Lemma Embeddings	82.5	0.3	82.7	0.3	-	-	-	-

Table 1: Main results on English and Chinese tasks. The measure is accuracy (%). The '+' in systems represent an increase in modules of the system in the previous row. Δ represents the difference between the result of the current system and that of the previous row.

language, because it may correspond to multiple meanings. Therefore, the final submitted system to the task has no lemma token, no segment embedding $\mathrm{E_C}$ and no lemma embeddings. However, in order to analyze the role of lemma tokens in this multilingual setting, we will report the results with lemma token and segment embedding in Table 1 and Section 5.1. In this case, the lemma token will be translated to the most common Chinese word.

4 Experimental Setup

In this section, we will describe the experimental settings for English and Chinese in detail.

4.1 English Setting

Take one pre-trained English cased BERT *base* model[2] with 12 layer, 768 hidden, 12 heads, and 110M parameters, and fine-tune on the English training data in 5 epochs with batch size is 16 and max sequence length is 128. The dropout rate is 0.2 and other settings are followed Devlin et al. (2019).

4.2 Chinese Setting

The fine-tuning setups are the same as the English ones, except that the pre-training model is a Chinese BERT *base*[3] with a layer of 12, hidden size of 768, heads of 12, and parameters of 110M.

For machine translation model, we implement Transformer *base* model (Vaswani et al., 2017) using the open-source toolkit *Fairseq-py* (Ott et al., 2019). The training data of English-Chinese is from UNPC v1.0 and MultiUN v1 in WMT17[4], which are total 30.4M sentences. We trained the model with dropout $= 0.1$ and using Adam optimizer (Kingma and Ba, 2015) with $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 10^{-9}$. The translation is detokenized and then the quality is 35.0, which is evaluated using the 4-gram case-sensitive BLEU (Papineni et al., 2002) with the *SacreBLEU* tool (Post, 2018).[5]. This translation model achieves satisfactory results, which shows that the method of translating English training set into Chinese has theoretical basis and feasibility.

5 Results

In this section, we will first report the main results of English and Chinese multilingual setting and analyze the importance of all the factors in the system. Then, we explore the probability of error for each part of speech.

5.1 Main Results

We conduct our experiments based on BERT frame[6]. The specific meanings of each system in the main experiment are as follows:

Fine-tuning Following the standard fine-tuning format of Devlin et al. (2019), sentences 1 and 2 are connected by [SEP].

+Lemma Token Lemma token is added on the basis of the previous system, and the training input at this time is '[CLS] + Sentence 1 + [SEP] + Sentence 2 + [SEP] + Lemma + [SEP]'.

+Segment $\mathrm{E_C}$ Based on the previous system, the segment embedding of the final lemma token is set to $\mathrm{E_C}$.

+Lemma Embeddings Lemma embedding will be added on the basis of the previous system, and the input representation at this time consists of four parts: token, segment, position, and lemma embeddings.

The main results are shown in Table 1, and the specific analysis is as follows:

[2] https://storage.googleapis.com/bert_models/2018_10_18/cased_L-12_H-768_A-12.zip

[3] https://storage.googleapis.com/bert_models/2018_11_03/chinese_L-12_H-768_A-12.zip

[4] http://www.statmt.org/wmt17/translation-task.html

[5] BBLEU+case.mixed+lang.en-zh+numrefs.1+smooth.exp+test.wmt17+tok.zh+version.1.4.4

[6] https://github.com/google-research/bert

	English		Chinese	
	Num	Acc	Num	Acc
NOUN	528	83.14	554	78.70
VERB	298	83.22	364	75.82
ADJ	144	81.94	62	69.35
ADV	30	73.33	20	60.00
Overall	1000	82.70	1000	76.70

Table 2: Error analysis on the test set. Num represents the number of examples of this part of speech in the test set, and Acc represents the accuracy (%) of the current part of speech.

English Fine-tuning alone can obtain relatively good performance, and the performance has been further improved with the introduction of the three new modules. Here we conduct experiments to check their influence on our method by adding them one by one. With the addition of +Lemma Token, our system has a significant improvement, while the improvement brought by the other two modules is slightly lower, indicating that the presence or absence of lemma token has a greater impact.

Chinese Fine-tuning achieves the best performance in this task. After the introduction of +Lemma Token, the result shows that the effect is greatly reduced, which may be because the translated lemma token is not necessarily the appropriate translation result. Because fine-grained translations of individual tokens have no context, they often fail to translate properly, as mentioned in Section 3.2. However, after the introduction of +Segment E_C, the effect has been slightly improved, which proves that our idea is effective. Because it is difficult to get the exact position of the lemma word after translation, there is no result of +Lemma Embeddings. Based on the above results, we use Fine-tuning as the final system for the Chinese task. In other words, our final model has no lemma token, no segment embedding E_C and no lemma embeddings.

5.2 Error Analysis

Lemma tokens have different parts of speech, so we think about the relationship between the accuracy of system prediction and the parts of speech of lemma tokens. Based on this, we reported the accuracy of each part of speech in the test set, as shown in Figure 2.

There are similar findings of error analysis for English and Chinese tasks. The order of the data volume on the test set is NOUN, VERB, ADJ, and ADV. The distribution of these data types is consistent with the training set, and each part of speech in the training set is 4123, 2269, 1429, and 175, which is also in descending order and occupies roughly the same proportion of each part of speech. Therefore, the parts of speech with more data in the training set often get better performance on the test set, which indicates the importance of data size. More data can enable the model to learn more classification knowledge and behavior, which affects the prediction results.

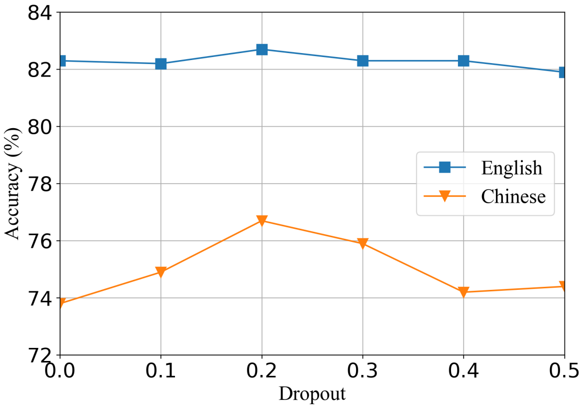

Figure 3: The performance of different dropouts on test set in English and Chinese tasks.

5.3 Impact of Dropout

To analyze the importance of dropout, we conducted experiments by using different dropouts on both English and Chinese test sets, and the results are shown in Figure 3. As we can see, the performance of the two tasks increases with the increment of dropout rate and reach the best performance when dropout rate equals 0.2. As dropout rate continues to increase, the performance deteriorates, which indicates that too many lost parameters may make the model difficult to converge.

Besides, Dropout performs differently in terms of data quality. In general, real corpus (English) should be of better quality than pseudo corpus (Chinese). On this basis, the performance of different dropout models on high-quality real corpus is relatively stable, with a gap of less than 1%, while the performance of pseudo corpus fluctuates greatly, with a gap of 3%.

6 Conclusion

In this paper, we describe the GX system participating in the MCL-WiC task. In order to obtain the general basic knowledge, we use the pre-trained BERT model and then fine-tune it on the data released by the organizer. In order to further emphasize the relationship between sentence pairs and the importance of lemma, we introduce three new factors: lemma token, lemma segment embedding, and lemma embedding, and finally get better results. Our system reaches 82.7% in English multilingual setting and 76.7% in Chinese multilingual setting.

Acknowledgments

We would like to thank the anonymous reviewers for their insightful comments. We thank Shuhao Gu for his helpful suggestions and inspiration.

References

José Camacho-Collados and Mohammad Taher Pilehvar. 2018. From word to sense embeddings: A survey on vector representations of meaning. *J. Artif. Intell. Res.*, 63:743–788.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3960–3973. Association for Computational Linguistics.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 328–339. Association for Computational Linguistics.

Renfen Hu, Shen Li, and Shichen Liang. 2019. Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3899–3908. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Daniel Loureiro and Alípio Jorge. 2019. Language modelling makes sense: Propagating representations through wordnet for full-coverage word sense disambiguation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5682–5691. Association for Computational Linguistics.

Federico Martelli, Najla Kalach, Gabriele Tola, and Roberto Navigli. 2021. SemEval-2021 Task 2: Multilingual and Cross-lingual Word-in-Context Disambiguation (MCL-WiC). In *Proceedings of the Fifteenth Workshop on Semantic Evaluation (SemEval-2021)*.

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning generic context embedding with bidirectional LSTM. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*, pages 51–61. ACL.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Mohammad Taher Pilehvar and José Camacho-Collados. 2019. Wic: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1267–1273. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020. Sensembert: Context-enhanced sense embeddings for multilingual word sense disambiguation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8758–8765. AAAI Press.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. Does BERT make any sense? interpretable word sense disambiguation with contextualized embeddings. In *Proceedings of the 15th Conference on Natural Language Processing, KONVENS 2019, Erlangen, Germany, October 9-11, 2019*.