# Adapting BERT for Word Sense Disambiguation with Gloss Selection Objective and Example Sentences

**Boon Peng Yap, Andrew Koh, Eng Siong Chng**
Nanyang Technological University, Singapore
{boonpeng001, andr0081, aseschng}@ntu.edu.sg

## Abstract

Domain adaptation or transfer learning using pre-trained language models such as BERT has proven to be an effective approach for many natural language processing tasks. In this work, we propose to formulate word sense disambiguation as a relevance ranking task, and fine-tune BERT on sequence-pair ranking task to select the most probable sense definition given a context sentence and a list of candidate sense definitions. We also introduce a data augmentation technique for WSD using existing example sentences from WordNet. Using the proposed training objective and data augmentation technique, our models are able to achieve state-of-the-art results on the English all-words benchmark datasets.[1]

## 1 Introduction

In natural language processing, Word Sense Disambiguation (WSD) refers to the task of identifying the exact sense of an ambiguous word given the context (Navigli, 2009). More specifically, WSD associates ambiguous words with predefined senses from an external sense inventory, e.g. WordNet (Miller, 1995) and BabelNet (Navigli and Ponzetto, 2010).

Recent studies in learning contextualized word representations from language models, e.g. ELMo (Peters et al., 2018), BERT (Devlin et al., 2019) and GPT-2 (Radford et al., 2019) attempt to alleviate the issue of insufficient labeled data by first pre-training a language model on a large text corpus through self-supervised learning. The weights from the pre-trained language model can then be fine-tuned on downstream NLP tasks such as question answering and natural language inference. For WSD, pre-trained BERT has been utilized in multiple ways with varying degrees of success. Notably,

Huang et al. (2019) proposed GlossBERT, a model based on fine-tuning BERT on sequence-pair binary classification task, and achieved state-of-the-art results in terms of single model performance on several English all-words WSD benchmark datasets.

In this paper, we extend the sequence-pair WSD model and propose a new task objective that can better exploit the inherent relationships within positive and negative sequence pairs. Briefly, our contribution is two-fold: (1) we formulate WSD as gloss selection task, in which the model learns to select the best context-gloss pair from a group of related pairs; (2) we demonstrate how to make use of additional lexical resources, namely the example sentences from WordNet to further improve WSD performance.

We fine-tune BERT using the gloss selection objective on SemCor (Miller et al., 1994) plus additional training instances constructed from the WordNet example sentences and evaluate its impact on several commonly used benchmark datasets for English all-words WSD. Experimental results show that the gloss selection objective can indeed improve WSD performance; and using WordNet example sentences as additional training data can offer further performance boost.

## 2 Related Work

BERT (Devlin et al., 2019) is a language representation model based on multi-layer bidirectional Transformer encoder (Vaswani et al., 2017). Previous experiment results have showed that significant improvement can be achieved in many downstream NLP tasks through fine-tuning BERT on those tasks. Several methods have been proposed to apply BERT for WSD. In this section, we briefly describe two commonly used approaches: feature-based and fine-tuning approach.

---

[1] Codes and pre-trained models are available at https://github.com/BPYap/BERT-WSD.

## 2.1 Feature-based Approaches

Feature-based WSD systems make use of contextualized word embeddings from BERT as input features for task-specific architectures. Vial et al. (2019) used the contextual embeddings as inputs in a Transformer-based classifier. They proposed two sense vocabulary compression techniques to reduce the number of output classes by exploiting the semantic relationships between different senses. The Transformer-based classifiers were trained from scratch using the reduced output classes on Sem-Cor and WordNet Gloss Corpus (WNGC). Their ensemble model, which consists of 8 independently trained classifiers achieved state-of-the-art results on the English all-words WSD benchmark datasets.

Besides deep learning-based approach, Loureiro and Jorge (2019) and Scarlini et al. (2020) construct sense embeddings using the contextual embeddings from BERT. The former generates sense embeddings by averaging the contextual embeddings of sense-annotated tokens taken from Sem-Cor while the latter constructs sense embeddings by concatenating the contextual embeddings of BabelNet definitions with the contextual embeddings of Wikipedia contexts. For WSD, both approaches make use of the constructed sense embeddings in nearest neighbor classification (kNN), in which the simple 1-nearest neighbor approach from Scarlini et al. (2020) showed substantial improvement over the nominal category of the English all-words WSD benchmark datasets.

## 2.2 Fine-tuning Approaches

Fine-tuning WSD systems directly adjust the pre-trained weights on annotated corpora rather than learning new weights from scratch. Du et al. (2019) fine-tuned two separate and independent BERT models simultaneously: one to encode sense-annotated sentences and another one to encode sense definitions from WordNet. The hidden states from the 2 encoders are then concatenated and used to train a multilayer perceptron classifier for WSD.

Huang et al. (2019) proposed GlossBERT which fine-tunes BERT on sequence-pair binary classification tasks. The training data consists of context-gloss pairs constructed using annotated sentences from SemCor and sense definitions from Word-Net 3.0. Each context-gloss pair contains a sentence from SemCor with a target word to be disambiguated (context) and a candidate sense definition of the target word from WordNet (gloss). During fine-tuning, GlossBERT classifies each context-gloss pair as either positive or negative depending on whether the sense definition corresponds to the correct sense of the target word in the context. Each context-gloss pair is treated as independent training instance and will be shuffled to a random position at the start of each training epoch. At inference stage, the context-gloss pair with the highest output score from the positive neuron among other candidates is chosen as the best answer.

In this paper, we use similar context-gloss pairs as inputs for our proposed WSD model. However, instead of treating individual context-gloss pair as independent training instance, we group related context-gloss pairs as 1 training instance, i.e. context-gloss pairs with the same context but different candidate glosses are considered as 1 group. Using groups of context-gloss pairs as training data, we formulate WSD as a ranking/selection problem where the most probable sense is ranked first. By processing all related candidate senses in one go, the WSD model will be able to learn better discriminating features between positive and negative context-gloss pairs.

## 3 Methodology

We describe the implementation details of our approaches in this section. When customizing BERT for WSD, we use a linear layer consisting of just 1 neuron in the output layer to compute the relevance score for each context-gloss pair, in contrast to the binary classification layer used in GlossBERT.

Additionally, we also extract example sentences from WordNet 3.0 and use them as additional training data on top of the sense-annotated sentences from SemCor.

### 3.1 Gloss Selection Objective

Following Huang et al. (2019), we construct positive and negative context-gloss pairs by combining annotated sentences from SemCor and sense definitions from WordNet 3.0. The positive pair contains a gloss representing the correct sense of the target word while a negative pair contains a negative candidate gloss. Each target word in the contexts is surrounded with two special [TGT] tokens. We group context-gloss pairs with the same context and target word as a single training instance so that they are processed sequentially by the neural network. As illustrated in Figure 1, the output layer takes the hidden states of the [CLS] token from
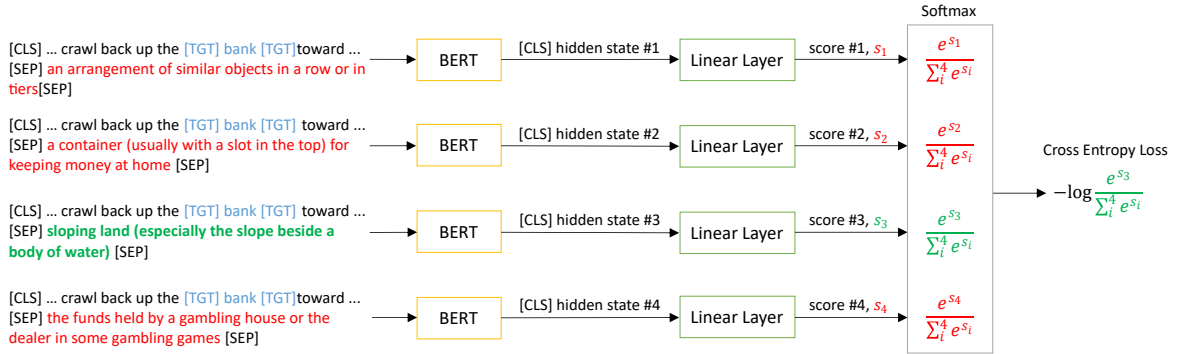
Figure 1: Visualisation of the gloss selection objective when computing the loss value for a training instance. The context "*He turned slowly and began to crawl back up the bank toward the rampart.*" is annotated with the target word "*bank*". A training instance consists of n context-gloss pairs (n=4 in this case), including 1 positive pair (shown in **green**) and n-1 negative pairs (shown in red). The order of the context-gloss pairs within each training instance is randomized during the dataset construction step.

each context-gloss pair as input and calculate the corresponding relevance score. A softmax layer then aggregates the relevance scores from the same group and computes the training loss using cross entropy as loss function. Formally, the gloss selection objective is given as follow:

$$loss = -\frac{1}{m} \sum_{i=1}^{m} [\sum_{j=1}^{n_i} 1(y_i, j) \log(p_{ij})] \quad (1)$$

where $m$ is the batch size, $n_i$ is number of candidate glosses for the $i$-th training instance, $1(y_i, j)$ is the binary indicator if index $j$ is the same as the index of the positive context-gloss pair $y_i$, and $p_{ij}$ is the softmax value for the $j$-th candidate sense of $i$-th training instance, computed using the following equation:

$$p_{ij} = \frac{\exp(Rel(context_i, gloss_{ij}))}{\sum_{k}^{n_i} \exp(Rel(context_i, gloss_{ik}))} \quad (2)$$

where $Rel(context, gloss)$ denotes the relevance score of a context-gloss pair from the output layer. Similar formulation was presented for web document ranking (Huang et al., 2013) and question-answering natural language inference (Liu et al., 2019). In the case of WSD, we are only interested in the top-1 context-gloss pair. Hence, during testing, we select the context-gloss pair with the highest relevance score and its corresponding sense as the most probable sense for the target word.

### 3.2 Data Augmentation using Example Sentences

Most synsets in WordNet 3.0 include one or more short sentences illustrating the usage of the synset

members (i.e. synonyms). We introduce a relatively straightforward data augmentation technique that combines the example sentences with positive/negative glosses into additional context-gloss pairs. First, example sentences (context) are extracted from each synset and target words are identified via keyword matching and annotated with two [TGT] tokens. Then, context-gloss pairs are constructed by combining the annotated contexts with positive and negative glosses. Using this technique, we were able to obtain 37,596 additional training instances (about 17% more training instances).

## 4 Experiments

In this section, we introduce the datasets and experiment settings used to fine-tune BERT. We also present the evaluation results of each model and compare them against existing WSD systems.

### 4.1 Datasets

Both training and testing datasets were obtained from the unified evaluation framework for WSD (Raganato et al., 2017b). Our training dataset for gloss selection consists of 2 parts: a baseline dataset with 226,036 training instances constructed from SemCor and an augmented dataset with 37,596 training instances constructed using the data augmentation method. When constructing the context-gloss pairs for the training datasets, we select a maximum of $n = 6$ context-gloss pairs per training instance; for testing datasets, all possible candidate context-gloss pairs are considered.

The testing dataset contains 5 benchmark datasets from previous Senseval and SemEval competitions, including Senseval-2 (**SE2**), Senseval-3

43

| | Dev | Test | | | | Concatenation of all datasets | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| System | SE07 | SE2 | SE3 | SE13 | SE15 | Noun | Verb | Adj | Adv | ALL |
| **KB** Most frequent sense baseline | 54.5 | 65.6 | 66.0 | 63.8 | 67.1 | 67.7 | 49.8 | 73.1 | 80.5 | 65.5 |
| Lesk$_{ext}$+emb | 56.7 | 63.0 | 63.7 | 66.2 | 64.6 | 70.0 | 51.1 | 51.7 | 80.6 | 64.2 |
| Babelfy | 51.6 | 67.0 | 63.5 | 66.4 | 70.3 | 68.9 | 50.7 | 73.2 | 79.8 | 66.4 |
| **Sup** IMS+emb | 62.6 | 72.2 | 70.4 | 65.9 | 71.5 | 71.9 | 56.6 | 75.9 | 84.7 | 70.1 |
| LSTM-LP | 63.5 | 73.8 | 71.8 | 69.5 | 72.6 | - | - | - | - | - |
| Bi-LSTM | - | 71.1 | 68.4 | 64.8 | 68.3 | 69.5 | 55.9 | 76.2 | 82.4 | 68.4 |
| HCAN | - | 72.8 | 70.3 | 68.5 | 72.8 | 72.7 | 58.2 | 77.4 | 84.1 | 71.1 |
| **Feat** LMMS$_{2348}$ (BERT) | 68.1 | 76.3 | 75.6 | 75.1 | 77.0 | - | - | - | - | 75.4 |
| SemCor+WNGC, hypernyms (single) | - | - | - | - | - | - | - | - | - | 77.1 |
| SemCor+WNGC, hypernyms (ensemble) | 73.4 | 79.7 | **77.8** | 78.7 | 82.6 | 81.4 | 68.7 | **83.7** | 85.5 | 79.0 |
| SENSEMBERT$_{sup}$ | - | - | - | - | - | 80.4 | - | - | - | - |
| *BEM*[2] | *74.5* | *79.4* | *77.4* | *79.7* | *81.7* | *81.4* | *68.5* | *83.0* | *87.9* | *79.0* |
| *EWISER$_{hyper}$*[2] | *75.2* | *80.8* | *79.0* | *80.7* | *81.8* | *82.9* | *69.4* | *83.6* | *87.3* | *80.1* |
| **FT** BERT$_{def}$ | - | 76.4 | 74.9 | 76.3 | 78.3 | 78.3 | 65.2 | 80.5 | 83.8 | 76.3 |
| GlossBERT (Sent-CLS-WS) | 72.5 | 77.7 | 75.2 | 76.1 | 80.4 | 79.3 | 66.9 | 78.2 | 86.4 | 77.0 |
| **Ours** BERT$_{base}$ (baseline) | **73.6** | 79.4 | 76.8 | 77.4 | 81.5 | 80.6 | 67.9 | 82.2 | 87.3 | 78.2 |
| BERT$_{base}$ (augmented) | **73.6** | 79.3 | 76.9 | 79.1 | 82.0 | 81.3 | 67.7 | 82.2 | 87.9 | 78.7 |
| BERT$_{large}$ (baseline) | 73.0 | **79.9** | 77.4 | 78.2 | 81.8 | 81.2 | **68.8** | 81.5 | **88.2** | 78.7 |
| BERT$_{large}$ (augmented) | 72.7 | 79.8 | **77.8** | 79.7 | **84.4** | **82.6** | 68.5 | 82.1 | 86.4 | **79.5** |

Table 1: F1-score (%) on the English all-words WSD benchmark datasets in Raganato et al. (2017b). The systems are grouped into 5 categories: i) knowledge-based system (*KB*), i.e. the most frequent sense baseline, Lesk$_{ext}$+emb (Basile et al., 2014) and Babelfy (Moro et al., 2014), ii) supervised models (*Sup*), i.e. IMS+emb (Iacobacci et al., 2016), LSTM-LP (Yuan et al., 2016), Bi-LSTM (Raganato et al., 2017a) and HCAN (Luo et al., 2018), iii) featured-based approach using contextual embeddings from BERT (*Feat*), i.e. LMMS$_{2348}$ (Loureiro and Jorge, 2019), SemCor+WNGC (Vial et al., 2019), SENSEMBERT$_{sup}$ (Scarlini et al., 2020), BEM (Blevins and Zettlemoyer, 2020) and EWISER$_{hyper}$ (Bevilacqua and Navigli, 2020), iv) fine-tuning approach using BERT (*FT*), i.e. BERT$_{def}$ (Du et al., 2019) and GlossBERT (Huang et al., 2019), v) our models (*Ours*).

(SE3), SemEval-07 (**SE07**), SemEval-13 (**SE13**), and SemEval-15 (**SE15**). Following Huang et al. (2019) and others, we choose SemEval-07 as the development set for tuning hyperparameters.

### 4.2 Experiment Settings

We experiment with both uncased BERT$_{base}$ and BERT$_{large}$ models. BERT$_{base}$ consists of 110M parameters with 12 Transformer layers, 768 hidden units and 12 self-attention heads while BERT$_{large}$ consists of 340M parameters with 24 Transformer layers, 1024 hidden units and 16 self-attention heads. We use the implementation from the transformers package (Wolf et al., 2019). In total, we trained 4 models on 2 setups: (1) **BERT$_{base/large}$ (baseline)**, using only the baseline dataset; (2) **BERT$_{base/large}$ (augmented)**, using the concatenation of baseline and augmented dataset.

At fine-tuning, we set the initial learning rate to 2e-5 with batch size of 128 over 4 training epochs. The remaining hyperparameters are kept at the default values specified in the transformers package.

### 4.3 Evaluation Results

We evaluate the performance of each model and report the F1-scores in Table 1, along with the results from other WSD systems.

All 4 of our models trained on the proposed gloss selection objective show substantial improvement over the non-ensemble systems across all benchmark datasets, which signifies the effectiveness of this task formulation[3]. The addition of augmented training set further improves the performance, particularly in the noun category. It is worth noting that Du et al. (2019) and Huang et al. (2019) reported slightly worse or identical results when fine-tuning on BERT$_{large}$, but both of our models fine-tuned on BERT$_{large}$ obtain considerable better results than the BERT$_{base}$ counterparts. This may be partially attributed to the fact that we were using the recently released whole-word masking variant of BERT$_{large}$, which was shown to have a better performance on the Multi-Genre Natural Language Inference (MultiNLI) benchmark. Although the BERT$_{large}$ (augmented) model has lower F1-

---

[2]For reference, we included the results from ACL2020. Since these results were not available at the time of writing this paper, we did not compare with the results in Section 4.3.

[3]Statistically different from previously reported results (with p=0.05) under one-sided randomization test on the F1-scores in concatenated dataset.

score on the development dataset, it outperforms the ensemble system consisting of eight independent $BERT_{large}$ models on three testing datasets and achieves the best F1-score on the concatenation of all datasets.

To illustrate that the improvement of WSD performance comes from the gloss selection objective instead of hyperparameter settings, we fine-tune a $BERT_{base}$ model on the unagumented training set using the same hyperparameter settings as Gloss-BERT (Huang et al., 2019), i.e. setting learning rate and batch size to 2e-5 and 64 respectively, and using 4 context-gloss pairs for each target word. As shown in Table 2, our model fine-tuned on the proposed gloss selection objective consistently outperforms GlossBERT across all benchmark datasets under the same hyperparameter settings.

|  | SE07 | SE2 | SE3 | SE13 | SE15 |
|---|---|---|---|---|---|
| GlossBERT | 72.5 | 77.7 | 75.2 | 76.1 | 80.4 |
| $BERT_{base}$ | 73.0 | 79.1 | 77.3 | 77.4 | 81.0 |

Table 2: Comparison of F1-score (%) on different benchmark datasets between GlossBERT and a $BERT_{base}$ model fine-tuned with gloss selection objective.

## 5 Conclusion

We proposed the gloss selection objective for supervised WSD, which formulates WSD as a relevance ranking task based on context-gloss pairs. Our models fine-tuned on this objective outperform other non-ensemble systems on five English all-words benchmark datasets. Furthermore, we demonstrate how to generate additional training data without external annotations using existing example sentences from WordNet, which provides extra performance boost and enable our single-model system to surpass the state-of-the-art ensemble system by a considerable margin on a number of benchmark datasets.

## Acknowledgements

## References

Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. 2014. An enhanced lesk word sense disambiguation algorithm through a distributional semantic model. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1591–1600.

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, Online.

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1006–1017, Online.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Jiaju Du, Fanchao Qi, and Maosong Sun. 2019. Using bert for word sense disambiguation. *arXiv preprint arXiv:1909.08358*.

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)*, pages 3500–3505.

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 2333–2338.

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)*, pages 897–907.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496.

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*, pages 5682–5691.

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*, pages 1402–1411.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

George A Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G Thomas. 1994. Using a semantic concordance for sense identification. In *Proceedings of the workshop on Human Language Technology*, pages 240–243. Association for Computational Linguistics.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244.

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

Roberto Navigli and Simone Paolo Ponzetto. 2010. Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 216–225. Association for Computational Linguistics.

Matthew 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, Volume 1 (Long Papers)*, pages 2227–2237.

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.

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

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017b. 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: Volume 1, Long Papers*, pages 99–110.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020. Sensembert: Context-enhanced sense embeddings for multilingual word sense disambiguation. In *Proc. of AAAI*.

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*, pages 5998–6008.

Loïc Vial, Benjamin Lecouteux, and Didier Schwab. 2019. Sense vocabulary compression through the semantic knowledge of wordnet for neural word sense disambiguation. In *Wordnet Conference*, page 108.

Thomas Wolf, L Debut, V Sanh, J Chaumond, C Delangue, A Moi, P Cistac, T Rault, R Louf, M Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv, abs/1910.03771*.

Dayu Yuan, Julian Richardson, Ryan Doherty, Colin Evans, and Eric Altendorf. 2016. Semi-supervised word sense disambiguation with neural models. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1374–1385.

# Appendix

## A  Additional Details on Experiment Settings

All models are trained using a single Nvidia Tesla K40 GPU with 12 GB of memory. Gradient accumulation is used to accommodate large batch size.

For hyperparameters search, we manually tune for the optimal hyperparameter combinations using the following candidate values:

- **BERT variant**: {cased, uncased}

- **Maximum number of glosses per context**: {4, 6}

- **Batch size**: {32, 64, 128}

- **Initial learning rate**: {2e-5, 3e-5, 5e-5}

- **Warm-up steps**: {0, 0.1 * total steps}

At testing stage, model checkpoints with the highest F1 score on the development dataset, i.e. SemEval-07, evaluated at every 1000 steps over 4 training epochs, are selected for evaluation on the testing dataset. We use the scoring script downloaded from `http://lcl.uniroma1.it/wsdeval/home`.