GPT-2 Contextual Data Augmentation for Word Sense Disambiguation

Rakia Saidi
LIMTIC Laboratory,
Faculty of science of Tunisia
UTM University / Tunisia

Fethi Jarray
LIMTIC Laboratory
ISI Medenine
Gabes University

Jeongwoo Jay Kang
LIG Laboratory, Emvista
Univ. Grenoble Alpes
France

Didier Schwab
LIG Laboratory
Univ. Grenoble Alpes
France

{saidi.rakya,fjarray}@gmail.com
{didier.schwab,jeongwoo.kang}@univ-grenoble-alpes.fr

Abstract

Most Word-Sense Disambiguation (WSD) systems rely on machine learning approaches that require large-scale corpora for effective training. So, the quality of a WSD system degrades when trained in a low-resource language such as Arabic. To improve WSD, we design a novel data augmentation technique by properly fine-tuning GPT-2 to generate a gloss or a phrase for a selected word. The generated training data is then combined with the original dataset to train a BERT-based WSD classifier. Experimental results show that integrating this augmentation technique improves WSD quality for both low-resource language (Arabic) and high-resource languages (English).

1 Introduction

Polysemous is a word with several related meanings. Some words, such as run or set have more than thirty different meanings. Polysemous words cause ambiguity in contexts where the meaning is different from the primary meaning of the word. For example, more than 40% of English words have more than one meaning (Nagy, 1995). In natural language processing, word sense disambiguation (WSD) consists of identifying the intended meaning (sense) of a polysemous word in a given context.

Two main approaches to WSD can be distinguished: Knowledge-Based Approach and Machine Learning-Based Approach. The former relies on external lexical resources, such as the well-known WordNet Knowledge Base. Machine learning based approaches use sense-annotated corpora to train the WSD system. The success of ML approaches in general and neural networks in particular highly depends on the availability of two resources inputs: 1) a dictionary or a sense inventory such as WordNet to specify senses for each word from some lexicon, 2) a sense annotated corpus for training such as SemCor. Deep neural networks achieved state-of-the-art performance for WSD. However, they are prone to overfitting on small datasets because they learn millions or billions of parameters while building the model. To avoid overfitting, the best solution is to use more training data by augmenting existing data.

WSD tasks suffer from insufficient training data or unavailability of annotated corpora. To overcome this challenge, we propose to generate new corpora or increase the size of existing ones. More specifically, we propose data augmentation techniques based on generative language models to improve the accuracy of learning methods.

In this paper, we start by discussing data augmentation (DA) in general and data augmentation for WSD in particular, with a focus on GPT(Generative Pre-trained Transformer) such as a generative language models. To validate our proposed method, we carried out experiments on GlossBert for English WSD and on ArabicGlossBert for Arabic. By conducting experiments on various different WSD tasks, we show that the proposed GPT data augmentation performs better than the baselines, existing techniques that are mainly devoted to the machine translation task.

The main contributions are as follows.

- We propose a GPT2 based augmentation method. The method allows GPT2 to augment sentences without altering the target word meaning. Our approach can further be applied to other NLP tasks such as semantic matching and natural language inference.
- Experimental results show that our approach achieves better performance, comparing with existing data augmentation methods.

Our contribution falls into centralized DL techniques, contrary to federated learning (Boughorbel et al., 2019). We suppose that the dataset set is clean, as opposed to the noisy data set (Boughorbel et al., 2018). To our knowledge, this is the first

attempt to utilize GPT-2 to augment data for WSD tasks.

The remainder of this paper is structured as follows. Section 2 presents the related work on WSD-DA. Section 3 explains our DA method for WSD. Section 4 presents our integrated WSD system, where BERT models are used for feature extraction. Section 5 discusses the results obtained. We conclude this paper with a summary of our contribution and we mention some future extensions.

2 Related work

Data augmentation strategy is used in computer Vision and Natural Language Processing(NLP) to overcome the challenge of deal data scarcity and data diversity insufficiency and to prevent overfitting. Data augmentation in NLP is not sufficient to replace a word with its synonym because the context will be different and many downstream tasks depending on the context, so it is not as easy as in computer vision where image cropping showed good improvements,

Data augmentation is basically performed based on human knowledge of invariance, rules, or heuristics. According to Li et al. (2022), data augmentation techniques can be categorized into three main categories: paraphrasing, noising and sampling. Methods based on paraphrasing replace a word by a generated one with its synonyms (Zhang et al., 2015) or with the most similar word after a similarity computation(Wang and Yang, 2015). This approach can't ensure a good diversity due to the limitation of lexical ontology such as WordNet and thus the diversity. Noising-based approaches add discrete or continuous noise to expand the training set, such as randomly swapping two words or replacing a word with another(Xie et al., 2017). Even if noising is a popular technique in computer vision, it may alter the semantics of the sentence and will change the context. For example, in sentiment analysis, changing the word happy by sad will change the class of sentences.

To address the limitations of the aforementioned approaches, sampling methods based on contextual augmentation have been proposed. Fadaee et al. (Fadaee et al., 2017) focused on parallel corpora and replaced high-frequency words with rare words in the target language according to a language model, then changed source words accordingly. Kobayashi (Kobayashi, 2018) and Wu et al. (Wu et al., 2019) applied a language model on a target word and predict a new word that will replace the original one. Gao et al. (Gao et al., 2019) introduced a soft contextual augmentation that occurs at the word embedding level, where the target word embedding is replaced by the expectation of word embeddings of predicted words by a language model. Yand et al. (Yang et al., 2019) fine-tuned the BERT model on different data sets for Open-Domain Question Answering and considered it also as a data augmentation. Papanikolaou and Pierleoni(Papanikolaou and Pierleoni, 2020) proposed a GPT-2 based approach to augment data training in relation extraction task. Kun et al. (Li et al., 2020), proposed a conditional augmentation method based on sequence to sequence generation for aspect term extraction task.

The contextual augmentation preserves the semantic since the new word is generated from the entire context, not only from the current target as it done by the other approach. Nevertheless, none of the above methods is applicable for WSD, as they don't take into account the ambiguous target word. There are few works dedicated to WSD DA. Yap et al. (2020) noted that the majority of the synsets in WordNet contain illustrative short sentences. He included illustrative sentence-gloss on the set of context-gloss pairings. Using this technique, they obtained 37,596 additional training instances (about 17% more training instances).

Kohli (2021) adopted a back-translation strategy for data augmentation. They fine-tuned the GlossBERT model on the augmented corpus and evaluated it on semEval. The disadvantage of this technique is that some information may be lost in back-translating process.

Lin and Giambi (2021) used BERT and WordNet to investigate alternative data augmentation approaches on context-gloss pairs to improve the performance of WSD. They demonstrated that augmentation procedures at the sentence and word levels are effective solutions for WSD. They also discovered that adding hypernym glosses from a lexical knowledge base can increase performance. Their procedure is available only for the French, German and Russian languages (not for Arabic and English).

Yuan et al. (2016) constructed a semi-supervised system for WSD data augmentation. It consists of increasing the tagged sample sentences

with a huge number of unlabeled sentences from the Web to solve these shortcomings by using the LSTM model. It is possible to add many unlabeled sentences by adding many unlabeled sentences. The data sets used in this work are SemEval and Semcor based on WordNet. The major drawback of this approach is that the generated sentences depend on the seed set.

Our work falls into the sampling techniques for text, but we focus on word-sense disambiguation.

3 Proposed method

To the best of our knowledge, this is the first contribution dedicated to data augmentation in WSD by the GPT family. GPT (Generative Pre-Training) is a large transformer-based language model with billions of parameters and trained on gigabytes of text scraped off the Internet. GPT-2 has shown impressive text generation results and low perplexity on several benchmarks. GPT-2 was trained with a standard language modeling objective and is therefore powerful in predicting the next word in a sequence of seen words. There are other variants of GPT that typically differ according to the number of layers, the size of the layers and the corpus used to train them such as GPT3, GPT-Neo and GPT-NeoX but we can't use it due to the huge size of these models and the unavailability of GPT-3.

In this paper, we aim to use GPT-2 to automatically generate WSD training data and combine it with existing datasets to train better WSD models. More concretely, given a training dataset for WSD such as Semcor, we fine-tuned the pre-trained GPT-2 model on this dataset to encourage GPT-2 to produce synthetic sentences that preserve the meaning of a target word.

The main challenge is how to deal with the target word and under which hypotheses, it retains its meaning. Our data augmentation is shown in Algorithm 3. Given an original tagged sentence, our strategy is to freeze the target word and to generate a gloss or phrase for a context word, such as the last word of a sentence. The number of context words replaced by a gloss can be seen as an extra hyperparameter.

[H]
Set a generation strategy
each sentence with an annotated target
- Freeze the target word
- Select context words according to the generation strategy
- Generate a gloss or a phrase for the selected word

We studied three strategies for increasing the number of examples based on a context selection of a chosen sense.

1. Entire context (Selection of the whole context): we replace each context word with a sequence of words. In this strategy, the sentence loses its structure;

2. Context-tail (Selection of the end of the context): we replace the last word of the context with a sequence of words.

3. Context-head (Selection of the beginning of the context): we replace the first word of the context with a sequence of words.

Figure 1 shows an example for each strategy of context selection. We note that we can exhibit other selection strategies, such as the random selection of a set of context words.

We validate our system on low-resource language (Arabic) and high-resource languages (English). So, we used a GPT-2 and a BERT models different for each language.

4 Experimental Evaluation

In this contribution, we are mainly interested in data augmentation for the task of lexical disambiguation for the Arabic and English languages. We use UFSAC (Vial et al., 2018) (*Unification of Sense Annotated Corpora and Tools*) which brings together the annotated corpus from the WordNet that have been automatically ported in Arabic and French via lexical transfer (Salah et al., 2018a).

4.1 Setup

To fine-tune GPT-2, we employed the medium model (355M) with a learning rate of 10^{-5}, a restore_from of fresh and a batch size of 16. The fine-tuning objective was to minimize the binary cross-entropy loss between the predicted senses and the golden senses. We applied the three above-mentioned strategies for Arabic and the English because it is possible to select any number of context words and replace them by a gloss.

We used pre-trained BERT models as a WSD classifier which we fine-tuned on either the gold or the gold+generated datasets. For fine-tuning, we used the pre-trained uncased BERT-BASE model. The total number of parameters in this pre-trained

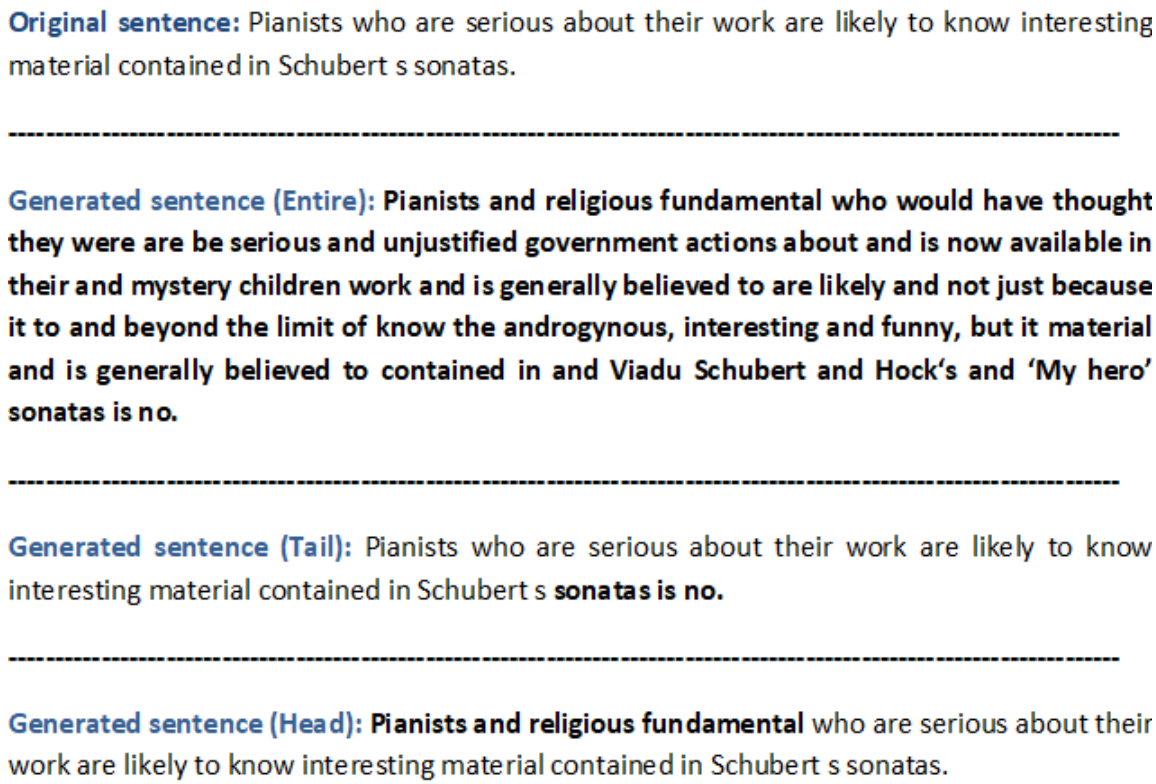

Figure 1: Examples for each of the selection strategies (in order, Entire-context, Context-tail, Context-head)

model is 110M, with 12 Transformer blocks, 768 hidden layer blocks, and 12 self-attention heads. For the optimizer, we used Adam (Kingma and Ba, 2014), a sequence length of 128, a batch size of 64 and a learning rate of 10^{-6}, the dropout probability is set to 0.1. We fine-tuned for 10 epochs, keeping the best model so far. We used the development set semEval2007(SE07) (Raganato et al., 2017) to fix the best parameters for our tests when fine-tuning. Concerning the final word representation, we average the final four layers of the first subword to get the representation of a word. That is, if a word is tokenized into a set of subwords, it's represented by the vector associated to the first subword.

For the English model, we use BERT with selection objective(Yap et al., 2020) and Gloss BERT (Huang et al., 2019) and (Du et al., 2019) that concatenate sentence embedding and gloss embedding for sentence representation. In addition, we tested with Roberta [1].

For the Arabic language, we use most available domain-specific pre-trained BERT models for modern standard Arabic (MSA): AraBERT (Antoun et al., 2020), Arabic-BERT (Safaya et al., 2020), CAMeL-BERT[2], and MARBERT[3] with arabGlossBERT (Al-Hajj and Jarrar, 2022). It's worthy mentioning that the multilingual mBERT (Libovickỳ et al., 2019) can also handle Arabic texts.

4.2 Datasets

For both languages, we carried out experiments on different corpora for data augmentation and for evaluation. Concerning data augmentation, our training is carried out on the SemCor corpus for the English language and its translated Arabic version for Arabic. Semcor is composed of 352 texts of the corpus Brown for a total of 226 040 annotations of meaning. This is the largest hand-annotated corpus available in English.

Regarding system evaluation, there are standard corpora for evaluation that are used or built up for evaluation campaigns.We use SE07 (Raganato

[1]https://huggingface.co/roberta-base

[2]https://huggingface.co/CAMeL-Lab/bert-base-arabic-camelbert-ca

[3]https://huggingface.co/UBC-NLP/MARBERT

et al., 2017) for English and the corpus OntoNotes Release 5.0 (Weischedel et al., 2013) for Arabic language.

SemEval2007 (Navigli et al., 2007) dataset contains 5677 words for 2261 words annotated in WordNet by Raganato et al. (2017). OntoNotes Release 5.0 has three languages (English, Arabic and Chinese). It is a large corpus manually annotated in a legal sense containing several kinds of text (news, telephone conversations, weblogs, usenet newsgroups, broadcast, talk shows). The Arabic portion of OntoNotes Release 5.0 includes 300K words from the Arabic corpus An-Nahar Newswire. This corpus is available [4] and contains 212332 words for 12524 annotated senses from WordNet.

4.3 Results

As baseline, we used GlossBert and context BERT for English and ArabGlossBert for Arabic. Table 1 represents the results of our augmentation approach for the WSD task and a comparison for the English WSD with GlossBERT(Huang et al., 2019) and with ContextBERT(Du et al., 2019). Wei et al. (Wei and Zou, 2019) provided a list of easy and basic data augmentation techniques for text classification.

Concerning the English language, the proposed data augmentation technique outperforms the basic BERT model by 7% and achieves an accuracy score of 92.4% because DA adds diversity to the original dataset. We also note also that RoBERTa outperforms BERT for all the augmentation strategies.

Regarding the Arabic language, we obtain the best accuracy by following the context-tail augmentation strategy and by using the Arabic Bert embedding model. So, Arabic Bert with GPT2 based augmentation achieves a state of the art on Arabic WSD by an accuracy of 88.88% on Ontonote dataset (Salah et al., 2018b).

Words in sentences are known to have different parts and different levels of importance in the sentence. Thus, a particular word replacement or substitution has an effect different from that of replacing another word. A crucial issue arises concerning which word to select and replace it with another word or phrase. Table 1 answers this question and shows that, for any BERT-based model, the context-tail data augmentation outperforms the context-head and context-entire augmentation. This can be explained by the fact that replacing the tail of a sentence by a gloss is more sense preserving than the other two strategies.

[4] https://goo.gl/peHdKQ

	ModelDataAug	GPT-2: c-head	GPT-2:c-entire	GPT-2: c-tail	w/o
3* English Test: Sse-meval	BERT	90.20%	77.96%	91.16%	NA
	RoBERTa	91.63%	79.73%	**92.4%**	NA
	GlossBERT(BERT)B1	NA			80.0%
	ContextBERT(BERT)B2	NA			86.1%
8* Arabic Test: Ontonote 8* 8*	ArabGlossBERT(AraBERTv02)B3	NA			84%
	ArabGlossBERT(CAMeLBERT)B4	NA			82%
	ArabGlossBERT(QARiB)B5	NA			80%
	AraBERTv02	82.25%	65.38%	85.95%	NA
	Arabic BERT	85%	72.01%	**88.88%**	NA
	CAMeLBERT	84%	55%	86.22%	NA
	MARBERT	83.77%	57%	86%	NA
	mBERT	84.69%	59.64%	88.76%	NA

Table 1: Performances of data augmentation techniques for WSD. The final column refers to the basic model without any augmentation. NA stands for not applicable. w/o stands for without. Lines marked with a reference are experiments results from that reference. (Baselines B1 from the work implemented by (Huang et al., 2019), B2 from du2019using and B3, B4 and B5 from the ArabglossBERT (Al-Hajj and Jarrar, 2022)). The English models are tested on Semeval. The Arabic models are tested on Ontonote.

5 Conclusion

In this work, we presented a novel data augmentation framework based on the GPT2 model for word sense disambiguation. The generated training data is then combined with the gold dataset to train a BERT-based WSD classifier. The results of the experiment are very encouraging and show that the proposed contribution outperforms existing augmentation techniques. We also empirically proved that context-tail selection is the better context generation strategy.

As a future extension of this work, we aim to study the effects on other English or Arabic corpora as well as proposing different data augmentation techniques.

References

Moustafa Al-Hajj and Mustafa Jarrar. 2022. Arabglossbert: Fine-tuning bert on context-gloss pairs for wsd. *arXiv preprint arXiv:2205.09685*.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Sabri Boughorbel, Fethi Jarray, Neethu Venugopal, and Haithum Elhadi. 2018. Alternating loss correction for preterm-birth prediction from ehr data with noisy labels. *arXiv preprint arXiv:1811.09782*.

Sabri Boughorbel, Fethi Jarray, Neethu Venugopal, Shabir Moosa, Haithum Elhadi, and Michel Makhlouf. 2019. Federated uncertainty-aware learning for distributed hospital ehr data. *arXiv preprint arXiv:1910.12191*.

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

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. *arXiv preprint arXiv:1705.00440*.

Fei Gao, Jinhua Zhu, Lijun Wu, Yingce Xia, Tao Qin, Xueqi Cheng, Wengang Zhou, and Tie-Yan Liu. 2019. Soft contextual data augmentation for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5539–5544.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. Glossbert: Bert for word sense disambiguation with gloss knowledge. *arXiv preprint arXiv:1908.07245*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. *arXiv preprint arXiv:1805.06201*.

Harsh Kohli. 2021. Transfer learning and augmentation for word sense disambiguation. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43*, pages 303–311. Springer.

Bohan Li, Yutai Hou, and Wanxiang Che. 2022. Data augmentation approaches in natural language processing: A survey. *AI Open*, 3:71–90.

Kun Li, Chengbo Chen, Xiaojun Quan, Qing Ling, and Yan Song. 2020. Conditional augmentation for aspect term extraction via masked sequence-to-sequence generation. *arXiv preprint arXiv:2004.14769*.

Jindřich Libovickỳ, Rudolf Rosa, and Alexander Fraser. 2019. How language-neutral is multilingual bert? *arXiv preprint arXiv:1911.03310*.

Guan-Ting Lin and Manuel Giambi. 2021. Context-gloss augmentation for improving word sense disambiguation. *arXiv preprint arXiv:2110.07174*.

William E Nagy. 1995. On the role of context in first- and second-language vocabulary learning. *Center for the Study of Reading Technical Report; no. 627*.

Roberto Navigli, Kenneth C Litkowski, and Orin Hargraves. 2007. Semeval-2007 task 07: Coarse-grained english all-words task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 30–35.

Yannis Papanikolaou and Andrea Pierleoni. 2020. Dare: Data augmented relation extraction with gpt-2. *arXiv preprint arXiv:2004.13845*.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. 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.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. Kuisail at semeval-2020 task 12: Bert-cnn for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059.

Marwa Hadj Salah, Hervé Blanchon, Mounir Zrigui, and Didier Schwab. 2018a. Un corpus en arabe annoté manuellement avec des sens wordnet. In *25e conférence sur le Traitement Automatique des Langues Naturelles*.

Marwa Hadj Salah, Hervé Blanchon, Mounir Zrigui, and Didier Schwab. 2018b. Un corpus en arabe annoté manuellement avec des sens wordnet. In *25e conférence sur le Traitement Automatique des Langues Naturelles*.

Loïc Vial, Benjamin Lecouteux, and Didier Schwab. 2018. Ufsac: Unification of sense annotated corpora and tools. In *Language Resources and Evaluation Conference (LREC)*.

William Yang Wang and Diyi Yang. 2015. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using# petpeeve tweets. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2557–2563.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

R Weischedel, M Palmer, M Marcus, E Hovy, S Pradhan, L Ramshaw, N Xue, A Taylor, J Kaufman, M Franchini, et al. 2013. Ontonotes release 5.0 ldc2013t19. web download. philadelphia: Linguistic data consortium, 2013.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Conditional bert contextual augmentation. In *Computational Science–ICCS 2019: 19th International Conference, Faro, Portugal, June 12–14, 2019, Proceedings, Part IV 19*, pages 84–95. Springer.

Ziang Xie, Sida I Wang, Jiwei Li, Daniel Lévy, Aiming Nie, Dan Jurafsky, and Andrew Y Ng. 2017. Data noising as smoothing in neural network language models. *arXiv preprint arXiv:1703.02573*.

Wei Yang, Yuqing Xie, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. Data augmentation for bert fine-tuning in open-domain question answering. *arXiv preprint arXiv:1904.06652*.

Boon Peng Yap, Andrew Koh, and Eng Siong Chng. 2020. Adapting bert for word sense disambiguation with gloss selection objective and example sentences. *arXiv preprint arXiv:2009.11795*.

Dayu Yuan, Julian Richardson, Ryan Doherty, Colin Evans, and Eric Altendorf. 2016. Semi-supervised word sense disambiguation with neural models. *arXiv preprint arXiv:1603.07012*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.