Investigating the Effect of Lexical Segmentation in Transformer-based Models on Medical Datasets

Vincent Nguyen
Australian National University
CSIRO Data61
vincent.nguyen@anu.edu.au

Sarvnaz Karimi
CSIRO Data61
Sydney, Australia
sarvnaz.karimi@csiro.au

Zhenchang Xing
Australian National University
Canberra, Australia
zhenchang.xing@anu.edu.au

Abstract

Transformer-based models have been popular recently and have improved performance for many Natural Language Processing (NLP) Tasks, including those in the biomedical field. Previous research suggests that, when using these models, an in-domain vocabulary is more suitable than using an open-domain vocabulary. We investigate the effects of a specialised in-domain vocabulary trained from scratch on a biomedical corpus. Our research suggests that, although the in-domain vocabulary is useful, it is usually constrained by the corpora size because these models needs to be trained from scratch. Instead, it is more useful to have more data, perform additional pretraining steps with a corpus-specific vocabulary.[1]

1 Introduction

In the natural language processing domain, there is a requirement for a fixed-sized vocabulary during training which could lead to *Out-Of-Vocabulary (OOV)* problem (Luong et al., 2015). This problem is when the fixed vocabulary model encounters an unseen word during inference, and the model is unable to handle it appropriately. WordPiece tokenisation, initially used in machine translation systems (Wu et al., 2016), has been widely successful in addressing the OOV problem by segmenting unseen words into *word pieces* as a representation for the unknown word. Previous research has either replaced unseen words with a special token (Luong et al., 2015), used character word embeddings (Labeau and Allauzen, 2017) as a fall-back, or ignored these words completely. These techniques have shortcomings as they do not attempt to represent the unseen word or require additional processing and memory as with character embeddings. WordPiece tokenisation is a trade-off, where there is no need for special handling of out-of-vocabulary, as unseen words are segmented into sub-word units. It allows a limited vocabulary to represent an infinitely sized vocabulary space.

WordPiece: *arthralgias* → *art-hra-al-gia-s*
Ideal: *arthralgias* → *arthr-algias*
arthr- means joints, -algias means pain

Figure 1: Word segmentation in WordPiece and the ideal segmentation using medical morphemes.

Models that successfully use WordPiece tokenisation include the transformer-based architectures: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and XLNet (Yang et al., 2019). BERT uses WordPieces as morphemes to aid the contextual representation of words. BERT performs at a state-of-the-art level on the GLUE tasks (Wang et al., 2019) due to its ability to fine-tune specifically to each task. Given this success, the model has also been applied to the biomedical domain through models such as BioBERT (Lee et al., 2019), which applies additional pretraining on the MEDLINE and PubMed corpora for biomedical text representation. However, these BioBERT models sometimes do not perform well on biomedical tasks, and in some instances are even worse than vanilla BERT (Zhu et al., 2019; Peng et al., 2019; Nguyen et al., 2019).

We hypothesise that a reason for this failure could be due to the vocabulary limitation of BioBERT, where the authors keep the open-domain vocabulary of BERT. This is problematic because the original BERT vocabulary is not suited for the biomedical domain due to the lack of medical suffixes and prefixes in its vocabulary leading to incorrect segmentation of the words (see the example in Figure 1). This is important because the suffixes and prefixes (the morphemes) in medical terminology carry distinct meanings,

[1] Our code is publicly available at Lexical-Segmentation-Transformer.

and almost the entire medical vocabulary can be constructed from prefix and suffix combinations (Stanfield et al., 2008). Thus, we aim to validate the importance of having the additional biomedical vocabulary for downstream tasks.

2 Background

Most of biomedical natural language processing is adapted from open-domain state-of-the-art techniques, from word embeddings (Chiu et al., 2016), to BiLSTM-CRF (Kalyan and Sangeetha, 2019). BERT is a deeply bidirectional encoder that is based on the transformer architecture. It uses self-attention as a mechanism of encoding input contextually by attending to different aspects of the sentence using multi-headed self-attention head, passed through layer normalisation and a Multi-Layer Perceptrons (MLP). The BERT model has been successful in the open-domain as it scored State-of-the-Art (SOTA) performance on the SQUAD (Rajpurkar et al., 2016) and GLUE datasets because it addresses the polysemy problem (Molla and Gonzlez, 2007), through contextual clues, for richer representations.

However, it was realised that directly applying the model to a closed domain can be problematic due to two factors: (1) The BERT model is trained on Wikipedia and BookCorpus meaning that the internal representations for specialised words may not be properly learned for a specialised domain; and, (2) The internal vocabulary that BERT has learned is suitable for tasks in the open-domain and a separate or additional vocabulary is needed (Beltagy et al., 2019).

To address the first problem, BioBERT takes the original BERT model and performs additional pretraining steps on academic biomedical literature, PubMed and MEDLINE, to improve downstream medical tasks. However, BioBERT does not change the open-domain vocabulary to a medically-focused one. Furthermore, the language in academic corpora is different to clinical text and patient language. These issues resulted in lower than expected performance on biomedical datasets, such the MEDIQA (Ben Abacha et al., 2019; Nguyen et al., 2019) and in some cases worse than the original BERT models (Zhu et al., 2019) from which they were trained from.

Alleviating the dataset problem, Clinical BERT (Alsentzer et al., 2019), performs further pretraining steps for the BioBERT and BERT models on domain-specific corpora showing marked improvements on downstream clinical tasks. However, they did not change the internal vocabulary of the models as this would require training the models from scratch which may limit performance.

Addressing both the vocabulary and the dataset problem, SciBERT was trained from scratch on the Semantic Scholar corpus and a specialised SentencePiece vocabulary trained on the corpus.

Our paper is a first look empirical study into the effect of vocabulary and dataset in applying BERT-based models to downstream tasks. Although our study is limited in scope, it still explores an important problem and our research suggests that some of previous studies may have drawn incorrect conclusions.

2.1 Sub-word Models

The original WordPiece algorithm addresses the OOV problem and handles arbitrary sequences of characters found on the web. This algorithm greedily maximises the likelihood of the vocabulary over the training data. The algorithm is similar to the byte-pair encoding algorithm, which uses frequency rather than likelihood to train the model. By using word pieces, the tokenisation procedure can break down OOV words into their word sub-units. For instance, jumped can be broken down into `jump ##ed`.

The SentencePiece algorithm (Kudo and Richardson, 2018) is similar to WordPiece except that it performs direct training from raw sentences with language independence. It treats all sentences as a sequence of Unicode characters without a special reliance on spaces, allowing for reliable multi-lingual de-tokenisation.

3 Methods

We propose vocabulary adaption to investigate the segmentation problems for medical text in BERT models and their variants. We propose two different methods to achieve this: (1) Adding additional vocabulary from a common medical vocabulary of suffixes and prefixes[2] to the existing BERT vocabulary and perform additional pretraining steps; and, (2) Training a separate SentencePiece tokeniser and pretraining a BERT model from scratch on the medical corpus. We compare

[2] GlobalRPh Common Medical Suffixes and Prefixes (Accessed Nov 2019)

Model	NLI		RQE		QA	
	Dev Acc.	Test Acc.	Dev Acc.	Test Acc.	Dev Acc.	Test Acc.
Medical vocab 10k steps	0.781	0.743	0.778	0.487	0.739	0.655
Medical Vocab (intermediate)	0.795	0.721	0.762	**0.500**	0.718	0.657
Medical Vocab (final)	0.791	0.751	0.748	**0.500**	0.731	0.634
Medical Vocab (final) - Medical Vocab	0.741	**0.798**	0.768	0.478	0.731	0.640
SentencePiece Vocab (intermediate)	0.769	0.684	0.745	0.491	**0.782**	0.665
SentencePiece Vocab (final)	0.666	0.684	0.431	**0.500**	0.641	0.513
BioBERT v1.0 PMC	0.809	0.768	0.775	0.487	0.778	**0.721**
BioBERT v1.0 PubMed+PMC	0.828	0.778	0.775	0.474	0.765	0.708
BioBERT v1.0 PubMed	0.815	0.775	0.791	0.465	0.744	0.704
BioBERT v1.1 PubMed	**0.833**	0.790	**0.808**	0.487	0.778	0.677
SciBERT + BaseVocab	0.799	0.773	0.745	0.487	0.735	0.697
SciBERT + SciVocab	0.817	0.783	0.785	0.483	0.761	0.713
BERT base	0.786	0.736	0.742	0.483	0.709	0.655

Table 1: Comparing accuracy of all models in three tasks using the MEDIQA datasets.

these methods against BERT, BioBERT and SciBERT models on downstream medical tasks.

3.1 Datasets

We use PubMed Central (PMC)[3] corpus for pretraining our BERT Model. It consists of two million articles, 300 million sentences and one billion tokens at the time of writing. Note that we use the full text of the articles, not just the abstracts, as this was shown to be effective in SciBERT (Beltagy et al., 2019).

For fine-tuning, we select the MEDIQA datasets (Ben Abacha et al., 2019) which contains three tasks: MEDical Natural Language Inference (MEDNLI) (Johnson et al., 2016), Recognising Question Entailment (RQE) (Abacha and Demner-Fushman, 2016), and Question Answering (QA).

3.2 Preprocessing

In order to comply with BERT's formatting for pretraining, for tokenisation and sentence segmentation, we use ScispaCy (Neumann et al., 2019), with a biomedical model (en_core_sci_sm) for its speed and ability to parse biomedical data. We then train a SentencePiece model with a fixed vocabulary size of 32,000 on a subset of 20 million PubMed text articles to extract a vocabulary that maximises likelihood over the dataset. We then adapt the SentencePiece vocabulary to be compatible with BERT by pruning '_' characters, replacing them with '##' and removing start and end of sequence tokens.

[3] PubMed Central Dump

3.3 Pretraining

Due to the large size of PMC and time and computing resources limitation, we randomly select a subset of 60 million sentences for pretraining. We use the default settings for pretraining the BERT model as described in the original paper. We also use the same pretraining schedule as the original BERT implementation where the model is first trained on a sequence length of 128, which we call the *intermediate model*, until convergence before being trained on a sequence length of 512, the *final model*. We set the learning rate of 1e-4 for the SentencePiece model as this is being trained from scratch and 2e-5 for the Medical Vocabulary model.

3.4 Fine-tuning

After pretraining, we fine-tune our model to each task in the dataset. We use a learning rate of 5e-5 for five epochs. We also use a fixed seed of 42 for all libraries. We train our model on the official training data and report our results on the development and test sets of each task.

We fine-tune 12 models to three separate tasks and evaluate on both the development and the test sets due to distribution mismatch (the test sets were made much later than the original training/development sets). We fine-tuned the BERT base model plus medical vocabulary with the models pretrained for 10k, 90k (intermediate), 100k (final) steps and a final model without the medical vocabulary. Similarly, we pretrain the BERT model with a PubMed SentencePiece vocabulary on models for 90k (intermediate) and 100k steps (final). We fine-tune all the BioBERT models,

where all v1.0 models are trained on abstracts of a specific corpus (e.g., Pubmed or PMC), and the v1.1 model is trained on the full-text corpus. We also fine-tune the SciBERT models with BERT base vocabulary (BaseVocab) and Semantic Scholar SentencePiece vocabulary (SciVocab). Finally, we fine-tune our baseline (BERT base). We report our results in Table 1.

4 Results and Discussions

Overall, we found that fine-tuned models , with the exception of our SentencePiece model and Medical Vocab model for QA, outperformed the *BERT base* baseline.

For the NLI task, the SentencePiece models and the Medical Vocab (final) model performed worse on the development set, however the Medical Vocab (final) - Medical Vocab model performed best on the test set. All other models performed scored higher than the BERT base model. The BioBERT models, on average, performed best here as the task involved inference from a medical sentence (a clinical note) to a normalised sentence (summary).

On the RQE dataset, all models performed reasonably on the development set, with the PubMED models scoring the best, with the exception of the final SentencePiece model as the task required interpretation of patient language in addition to academic. However, all models performed poorly on the test set, with no model scoring higher than random guess due to a marginal distribution mismatch between the training, development sets against the test set.

On the QA dataset, the task involved interpreting a patient's naturally formed question to a medical answer from medical articles. Here, BioBERT performed the best on the test set.

In summary, all models performed similarly with only mild discrepancies which we discuss in the following section.

4.1 SciVocab versus BaseVocab

We find that the SciVocab model performed better than the BaseVocab model (see Table 1, rows 11-12). BaseVocab is trained similarly to our medical vocab model where BERT base was fine-tuned with additional data before further tuned to a downstream task. The reason the SciVocab model performed better is that it had learned better representations during the training phase while the BaseVocab model learned noisier representations due to a vocabulary mismatch between the Semantic Scholar dataset and the BERT vocabulary. However, the SciVocab may not be as beneficial due to the academic nature of the vocabulary as the MEDIQA contains a mix of both academic medical terminology and natural patient questions.

4.2 BioBERT versus SciBERT SciVocab

The BioBERT and SciBERT models are both pre-trained/tuned on academic biomedical literature. However, there are two keys differences to note, SciBERT is trained from scratch as it is not possible to completely alter the BERT vocabulary while maintaining the original weights. We found that, contrary to previous research (Zhu et al., 2019), citing a development accuracy of 43% (RQE) and 68% (NLI), the BioBERT models performed better on the development and test sets of the MEDIQA datasets. We attribute BioBERT's strength to the fact that it was fine-tuned rather than trained from scratch, and thus incorporates both open-domain and biomedical-domain knowledge. Further evaluation with a purely biomedical reasoning task such as clinical term extraction (Si et al., 2019) may be suitable for further comparison.

We found that the BioBERT models performed better than previously reported and that the size of corpus matters in the performance of the model as the full-text corpus model is generally better.

4.3 Medical versus SentencePiece Vocab

We found that, on two of the tasks, the medical vocab model performed better due to the nature of the task. The SentencePiece vocab is adapted only for the PMC corpus, which is academically written without misspellings or colloquialism, in contrast with the datasets. That is, having a corpus specific vocabulary might not be sufficient even within the same domain due to the different nature of writing styles; academic and general audience. Furthermore, we found that the SentencePiece vocab do not contain all the punctuation tokens, which further hurts performance when it comes to understanding questions as '?' is replaced with 'unk'.

Consistent with SciBERT and BaseVocab vocabulary overlap, there was a 40% overlap in vocabulary between BERT base vocabulary and the PMC SentencePiece vocabulary, highlighting the vocabulary mismatch between two corpora. Also, there is a 4% overlap in the added medical suffix/prefix vocabulary and the SentencePiece

vocabulary suggesting that the PMC corpus was likely not training the representations of the added prefix and suffix tokens correctly because they do not appear frequently enough. Finally, due to the relatively smaller size of pretraining dataset compared with all the other models, the SentencePiece model most likely overfit as the performance across all datasets worsened with more training steps (see Table 1, rows 6-7).

However, the training with the PMC corpus allowed for better adaption to the downstream tasks. Our models did not perform as well as BioBERT as they are trained on a smaller subset than the original models. We also find that the intermediate SentencePiece model performs better than the final model, and this is because the downstream task had only short sequences, introducing noise and overfitting. The medical vocab model, rather than the SentencePiece model, is more robust against this noise as it is not trained from scratch.

4.4 BioBERT versus Medical Vocab

Although trained similarly, all the BioBERT models outperformed our pretrained models across all datasets. For a direct comparison, we compare *BioBERT v1.1 PubMED* as this shares the same dataset and pretraining procedures. The only notable differences between the BioBERT model and ours is that: 1) 3% of the Medical vocab model is augmented with medical suffix/prefix and 2) We trained only a subset of PubMED on the Medical vocab model. We do a preliminary test by removing additional vocabulary (Table 1, row 4) in our model for a comparison against dataset size. We saw that the performance of the model increased slightly on average, leading to the conclusion that the extra vocabulary was hurting performance as they were not well trained. Overall, we also find that the accuracy is still lower than the BioBERT model, suggesting that additional dataset size is crucial to achieving a better performance.

4.5 Effect of Corpus and Vocabulary

In all the models, although vocabulary helps (e.g., SciVocab vs. BaseVocab), this effect is limited to the pretraining phase when learning representations, but when applying to a downstream task, it is more important to have additional corpus data that is suited to the downstream task. This effect is shown where SciBERT basevocab (fine-tuned from the BERT base model) performed better than the BERT base model. The additional corpus data is useful in the case of BioBERT vs. SciVocab as BioBERT is fine-tuned with additional data on top of the BookCorpus and Wikipedia datasets of the BERT model.

We hypothesise that the best way to maximise all these effects is instead of fine-tuning from one corpus to the other, to combine both the open-domain and target domain corpora and pretrained the model from scratch with a well-tuned vocabulary. We leave this to future work.

5 Limitations and Future Work

There are several limitations to our study which we leave as directions for future work: (1) we only trained on a subset of the PMC dataset for pretraining the Medical Vocab and SentencePiece models as it was computationally intensive to use the full set; (2) we only trained and evaluated on BERT base models. For a complete comparison we need to pretrain all the BioBERT models, SciBERT models, our models and also, for completeness, clinical BERT using the BERT large model, and then 3) we would need to train on datasets of varying sizes to see the effect of the corpus. Furthermore, investigation of character embeddings as a segmentation strategy over the use of wordpieces, avoiding the need for a vocabulary could be useful. However, this would require factorisation of the embedding space to reduce the computational cost of increased sequence length (Lan et al., 2019).

Furthermore, empirically, we did not conduct a significance test due to the use of a fixed seed for all randomisation to emphasise reproducibility, however, in future, re-running each experiment without a fixed seed multiple times to produce reliable statistics is desirable in future work.

6 Conclusions

Previous research suggests that using open-domain vocabulary in BERT-based models affects downstream tasks compatibility and leads to a loss in effectiveness. However, our research suggests that this is not the case. An open-domain vocabulary is more useful than an in-domain vocabulary trained on less data, if it is additionally trained on an in-domain corpus.

Acknowledgements

This research is supported by the Australian Research Training Program and the CSIRO Research Office Postgraduate Scholarship.

References

Ben Abacha and Demner-Fushman. 2016. Recognizing Question Entailment for Medical Question Answering. *American Medical Informatics Association Annual Symposium Proceedings*, 2016:310–318.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. 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 3606–3611, Hong Kong, China. Association for Computational Linguistics.

Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. 2019. Overview of the MEDIQA 2019 shared task on textual inference, question entailment and question answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 370–379, Florence, Italy.

Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. 2016. How to train good word embeddings for biomedical NLP. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 166–174, Berlin, Germany.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Alistair Johnson, Tom Pollard, Lu Shen, Li-wei Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Anthony Celi, and Roger Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035.

K.S Kalyan and S. Sangeetha. 2019. SECNLP: A survey of embeddings in clinical natural language processing. *Computing Research Repository*, abs/1903.01039.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Matthieu Labeau and Alexandre Allauzen. 2017. Character and subword-based word representation for neural language modeling prediction. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 1–13, Copenhagen, Denmark. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *Computing Research Repository*, arXiv:1909.11942.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *Computing Research Repository*, abs/1907.11692.

Thang Luong, Ilya Sutskever, Quoc Le, Oriol Vinyals, and Wojciech Zaremba. 2015. Addressing the rare word problem in neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 11–19, Beijing, China. Association for Computational Linguistics.

Diego Molla and Jos Gonzlez. 2007. Question answering in restricted domains: An overview. *Computational Linguistics*, 33:41–61.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy.

Vincent Nguyen, Sarvnaz Karimi, and Zhenchang Xing. 2019. ANU-CSIRO at MEDIQA 2019: Question answering using deep contextual knowledge. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 478–487, Florence, Italy.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the Workshop on Biomedical Natural Language Processing*, pages 58–65, Florence, Italy.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. *Computing Research Repository*, abs/1606.05250.

Yuqi Si, Jingqi Wang, Hua Xu, and Kirk Roberts. 2019. Enhancing clinical concept extraction with contextual embedding. *Computing Research Repository*, abs/1902.08691.

P. Stanfield, Y.H. Hui, and N. Cross. 2008. *Essential Medical Terminology*. Jones and Bartlett. Chapter 2.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *In the Proceedings of International Conference on Learning Representations*, pages 353–355, Brussels, Belgium.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *Computing Research Repository*, page arXiv:1609.08144.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Computing Research Repository*, page arXiv:1906.08237.

Wei Zhu, Xiaofeng Zhou, Keqiang Wang, Xun Luo, Xiepeng Li, Yuan Ni, and Guotong Xie. 2019. PANLP at MEDIQA 2019: Pre-trained language models, transfer learning and knowledge distillation. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 380–388, Florence, Italy.