Cross-Domain Language Modeling: An Empirical Investigation

Vincent Nguyen[1,2] **Sarvnaz Karimi**[1] **Maciej Rybinski**[1] **Zhenchang Xing**[2]
[1]CSIRO Data61, Sydney, Australia
[2]The Australian National University, Canberra, Australia
{firstname.lastname}@csiro.au
{zhenchang.xing}@anu.edu.au

Abstract

Transformer encoder models exhibit strong performance in single-domain applications. However, in a cross-domain situation, using a sub-word vocabulary model results in sub-word overlap. This is an issue when there is an overlap between sub-words that share no semantic similarity between domains. We hypothesize that alleviating this overlap allows for a more effective modeling of multi-domain tasks; we consider the biomedical and general domains in this paper. We present a study on reducing sub-word overlap by scaling the vocabulary size in a Transformer encoder model while pretraining with multiple domains. We observe a significant increase in downstream performance in the general-biomedical cross-domain from a reduction in sub-word overlap.

1 Introduction

Contemporary language models are pretrained on massive, linguistically diverse corpora (Lan et al., 2020; Devlin et al., 2019a). It is not uncommon for these models to excel at benchmark downstream tasks (Wang et al., 2019a), given the use of contextual representations (Devlin et al., 2019b) that are trained on a variety of source *domains*—a term used to describe a distribution of language on a given topic or genre (for example BIOMEDICAL, SCIENTIFIC)—or GENERAL domain. However, the benefit of GENERAL domain pretraining for specialized application is questionable, as applying these language models (Gu et al., 2020) to specialized tasks is worse than using specialized counterparts (Beltagy et al., 2019). This degradation still occurs after sequential pretraining on specialized domains (Shin et al., 2020) when *fine-tuned* (updates to pretraining) to downstream tasks.

We hypothesize some of this degradation lies in the use of a *sub-word vocabulary* (Si et al., 2019). Sub-word vocabularies (Sennrich et al., 2016; Wu et al., 2016a,b) allow for efficient modeling of a source language distribution with a limited vocabulary size. However, problematically sub-words can be shared between different words—for example *hypotension* and *hypocritical*—with different meanings. This potentially conflates the vector representation of a sub-word (or *wordpiece*) causing *sub-word overlap*. When this overlap occurs with sub-words appearing in multiple domain contexts we call this *cross-domain sub-word overlap*.

As a pilot empirical study, we investigate reducing *cross-domain sub-word overlap*, by increasing vocabulary size, in language models pretrained in the GENERAL and BIOMEDICAL cross-domain. To evaluate the effect of sub-word overlap, general and biomedical domain benchmarks are used in this study as the *task distribution* includes different linguistic phenomena such as grammar, sentiment, textual similarity, natural language inference (Wang et al., 2019a). Interestingly, we find that disjoint sub-word vocabulary sets are not ideal. Some sub-word overlap is necessary and unavoidable, and a different level of overlap is ideal for each target domain. We also find a positive trend occurs when reducing cross-domain sub-word overlap, suggesting that there is a trade-off depending on the target downstream task and domain.

To better understand the results, we look at the impact of the pretraining data domain on downstream benchmark performance. Surprisingly, we found that inclusion of the general domain with a specialized domain improves downstream performance for that specialized domain's tasks, but *not* the other way around. This suggests that specialized domains should be trained in tandem with a general one.

Our contribution is a pilot study that investigates a pretraining strategy to reduce *cross-domain sub-word overlap* between GENERAL and BIOMEDICAL domains. We train cross-domain language models with varied vocabulary sizes and evaluate them on downstream classification tasks. We show that a

significant improvement can be achieved on two benchmark datasets ((Wang et al., 2019a), (Peng et al., 2019)) when reducing overlap. Further experiments point to the importance of selecting appropriate pretraining data for specialized domains.

2 Related Work

We discuss strategies from the literature to adapt the GENERAL domain language model, in particular a Transformer (Vaswani et al., 2017) encoder (Devlin et al., 2019b), to a specialized domain.

Domain-specific pretraining Many studies have adapted BERT, a popular Transformer encoder, to a specialized domain. However, as BERT was pretrained with a general domain sub-word vocabulary and trained on general domain data (BookCorpus and Wikipedia), domain adaptation is needed. For example, in the BIOMEDICAL domain, BioBERT (Lee et al., 2019) benefited from additional pretraining of the pretrained BERT model on academic biomedical corpora (PubMed Open Access and MEDLINE), showing a marked improvement on downstream biomedical tasks. DAPT (Gururangan et al., 2020) showed similar improvements.

However, BioBERT's approach was less effective in clinical applications; thus, ClinicalBERT (Alsentzer et al., 2019) was trained on domain-specific clinical corpora to improve upon downstream clinical tasks. Similarly, BlueBERT (Peng et al., 2019) was pretrained on a combination of domain-specific data, including PubMed abstracts and clinical notes. However, these approaches were only specialized for narrow task distributions rather than the entire BIOMEDICAL domain (Nguyen et al., 2019) and were trained sequentially (general to biomedical) rather than combined initially, which may suffer from effects such as catastrophic forgetting (McCloskey and Cohen, 1989).

Vocabulary Insertion Other studies considered extending a Transformer-based model's vocabulary without repeating the expensive pretraining step. In particular, one study replaced unused vocabulary elements with medical suffixes and prefixes (Nguyen et al., 2019). Additional pretraining steps were used so that the model learned the new vocabulary. They found that vocabulary insertion did not help as much as an increase in pretraining data. A similar observation is found by Shin et al. (2020) and Beltagy et al. (2019). However, another study using a domain-specific tokenizer for vocabulary insertion (Tai, 2019) found improvements in the German legal domain. However, improvements from vocabulary insertion are minimal, as there is still an interaction between the original vocabulary embeddings and the embeddings added during the fine-tuning step, resulting in sub-word overlap.

Wang et al. (2019b) proposes an enrichment of the BERT vocabulary by using embeddings from other models and learns a projection to the BERT embedding space in a multilingual setting. exBERT (Tai et al., 2020) extends the embedding dimension with domain-specific vocabulary. The model's original weights and embeddings are frozen during extended vocabulary training. Within the same class of approaces, (Poerner et al., 2020) propose a method where general domain embeddings are aligned with target-domain-specific word2vec embeddings. However, vocabulary insertion approaches circumvent the pretraining stage with domain-specific data which may potentially be more important than a vocabulary change (Shin et al., 2020).

Domain-specific vocabulary pretraining An extension to these methods is to pretrain on a target domain corpus with a custom vocabulary. SciBERT (Beltagy et al., 2019) showed that pretraining from scratch with a domain-specific vocabulary is better than a general-purpose vocabulary despite having fewer combined pretraining examples. Similarly, BioMegatron (Shin et al., 2020) showed that a larger custom vocabulary is useful for biomedical named entity recognition tasks and that a domain-specific vocabulary is more valuable than a larger model. They also show that a larger vocabulary size caused a reduction in over-segmentation, a problem that occurs when using a general vocabulary on specialized tasks (Chalkidis et al., 2020) that increases sub-word overlap.

Our work is a pilot study that extends upon domain-specific vocabulary pretraining to investigate cross-domain sub-word modeling. We pretrain models with varying vocabulary sizes to reduce sub-word overlap. In particular, we focus on cross-domain pretraining, which was previously unexplored in vocabulary experiments.

3 Datasets and Tasks

We use the combined English snapshot of Wikipedia (a proxy for the general domain) and

PubMed Open Access Full-Text corpora (biomedical domain) taken on the 1st of April 2020 for pretraining the language models and tokenizers. The PubMed corpus, consisting of 8.3 billion tokens, is preprocessed to remove references, while the Wikipedia corpus, consisting of 2.0 billion tokens, is extracted and cleaned with *wikiextractor* (Attardi, 2015). We use this pretraining data combination as a cross-domain proxy of the GENERAL and BIOMEDICAL domain. We use the training and validation sets of the GLUE benchmark (Wang et al., 2019a) to fine-tune our models for general domain benchmarking. Likewise, we use the publicly available subset of the BLUE tasks collection for the biomedical domain (Peng et al., 2019).

4 Experiments

We perform pretraining with a cross-domain corpus with the ALBERT model, which results in a high degree of cross-domain sub-word overlap. In addition, we experiment with models that have different vocabulary sizes (5000 to 100,000), each with a varying degree of sub-word overlap during pretraining. In Transformer models, the embedding dimension is coupled with the model's hidden dimension, causing the vocabulary size to control the model size—a larger vocabulary size exponentially increases the model's size. To remedy this, we use the ALBERT model (Lan et al., 2020), which projects the embedding dimension to a latent vocabulary dimension before projecting it to the model's hidden dimension. This projection allows scaling of the vocabulary size without significantly impacting the model's size.

Task performance and vocabulary size After pretraining, for each vocabulary size, we then evaluate our language models on downstream BLUE and GLUE benchmark datasets to determine how downstream performance is affected by the amount of sub-word overlap.

Determining Sub-word Overlap To determine the amount of sub-word overlap in relation to vocabulary size, we tokenize each general domain and biomedical task in GLUE and BLUE for each vocabulary size and compute the Jaccard index (Jaccard, 1912). The GLUE and BLUE tasks, are used as a cross-domain proxy between the GENERAL and BIOMEDICAL domains.

Experimental Setup For each model (vocabulary size $|V|$), we train a separate tokenizer using Byte-Pair Encoding (Sennrich et al., 2016). We use masked language modeling to train the largest model, ALBERT$_{|V=100{,}000|}$, on the combined corpora of Wikipedia and PubMed for two weeks using four V100 GPUs with an effective batch size of 256. We use the LAMB (You et al., 2020) optimizer and a maximum model sequence length of 512. All other hyperparameters are left as default, as described by Lan et al. (2020). For each model, we select the checkpoint such that validation performance (perplexity) is equal for all models. We then evaluate each model on both general domain and biomedical benchmark tasks. Specifically, we fine-tune each model for a maximum of 15 epochs for all the biomedical tasks, taking the best model on the validation set for inference over the test set. For the general domain tasks, to reduce overfitting (false convergence), we train each task for five epochs and report the validation performance as the test set labels are not publicly available.

However, scaling vocabulary size itself can lead to performance increases (Shin et al., 2020). Hence, we use the checkpoint where validation performance for masked language modeling is equal across all models; meaning that all models have similar capacity for language modeling with the only difference being vocabulary size during downstream updates via fine-tuning. The increase in parameter count due to vocabulary embeddings is negligible as the embeddings are all projected into the same sized latent dimension before being used by the model.

The classification layer used is created for each individual task and is not shared by any model. We use the default classification layer, with the correct label output layer as provided by the huggingface library (Wolf et al., 2019).

$\|V\|$	CoLA	MNLI	MRPC	QNLI	QQP	RTE	SST-2	STS-B	WNLI
5000	13.7	64.9	79.6	64.3	75.7	**56.7**	78.0	17.7	53.5
10000	12.8	70.1	79.3	67.4	79.9	51.3	80.5	19.9	33.8
20000	9.30	70.8	78.6	78.8	**81.5**	51.6	83.4	61.0	46.5
30000	20.7	70.6	78.4	78.8	81.3	54.9	83.1	**63.7**	56.3
40000	14.8	71.2	**80.9**	78.7	80.0	54.5	82.3	21.3	43.7
50000	15.3	71.1	79.2	79.5	80.9	54.5	83.4	28.9	46.5
60000	16.9	**71.4**	77.3	79.6	80.3	53.1	82.1	25.6	42.3
70000	17.4	71.0	79.3	78.8	79.7	55.6	**85.7**	26.9	36.6
80000	17.3	71.0	80.3	79.0	81.3	53.8	84.8	31.0	**56.3**
90000	21.2	71.1	80.1	**79.6**	81.2	50.2	84.3	25.8	56.3
100000	**21.9**	71.3	79.0	79.0	80.4	52.4	83.7	34.5	46.5

Table 1: Evaluation of the general domain tasks against varied $|V|$ of the ALBERT model.

$\|V\|$	Jaccard Index	Num. Overlaps	Num. Overlaps/$\|V\|$	$\|V\|$ in use
5000	94.6	4710	94.2%	99.5%
10000	87.8	8730	87.3%	99.5%
20000	73.8	14600	73.0%	99.0%
30000	62.8	18480	61.6%	98.2%
40000	54.6	21200	53.0%	97.1%
50000	48.2	23100	46.2%	95.8%
60000	43.0	24360	40.6%	94.4%
70000	38.9	25200	36.0%	92.7%
80000	35.6	25840	32.3%	90.8%
90000	32.8	26280	29.2%	89.1%
100000	30.4	26600	26.6%	87.3%

Table 2: Jaccard Index and overlap proportion for varying vocabulary sizes.

5 Results and Discussion

We trained masked language models of varying vocabulary sizes, each with its own degree of sub-word overlap and evaluate on downstream general and biomedical language understanding benchmarks. We found that cross-domain sub-word overlap reduction benefited the cross-domain between the general (Table 1) and biomedical domain (Figure 1) as sub-word overlap decreased (Table 2).

In terms of sub-word overlap, we find that the Jaccard index decreases sharply with vocabulary size (Table 2), indicating that biomedical and general domain tasks share common elements. This overlap decreases rapidly, especially at larger vocabularies (26.6% overlap at $|V| = 100,000$). A similar overlap percentage is reported by Beltagy et al. (2019) when measuring overlap between scientific and general domain vocabulary.

We also report the sub-word overlap proportional to vocabulary size (Table 2) and observe that it also falls sharply in a similar pattern. Although sub-word overlap proportion decreases, at least 87.3% of the vocabulary is still used, meaning vocabulary elements are not underused. Generally, reducing the overlap from approximately 60% Jaccard Index ($|V| < 30000$) to 40% ($|V| \geq 70000$) increases effectiveness in the biomedical domain while producing small improvements in the general domain

Benchmark	Pretraining Corpora	Effectiveness
BLUE (F1)	Wiki	0.6973
	PubMed	0.6706
	PubMed+Wiki	**0.7186**†
GLUE (Acc)	Wiki	0.7090
	PubMed	0.7060
	PubMed+Wiki	0.6906

Table 3: Pretraining data selection and downstream benchmark performance. BLUE is measured in terms of F1-score, while GLUE is measured in Accuracy. The BLUE benchmarks have a confidence interval higher than 0.95 using a sign test.

Domain	Task	S	L	L-S
General Domain	CoLA	14.3	14.7	+0.40
	MNLI	69.5	71.1†	+1.60
	MRPC	79.4	79.6	+0.20
	QNLI	73.6	79.3	+5.70
	QQP	79.7	80.6	+0.90
	RTE	53.8	53.5	-0.50
	SST-2	81.5	84.0†	+2.50
	STS-B	36.7	28.8	-7.90
	WNLI	46.8	47.5	+0.70
Biomedical	biosses	13.6	19.0	+5.40
	chemprot	59.4	65.2	+5.80
	DDI	66.9	71.2†	+4.30
	HoC	81.4	82.1	+0.70
	MedNLI	67.6	70.2†	+2.60

Table 4: Performance of vocabulary sizes larger (L) than 50,000, and vocabulary sizes smaller (S) than 50,000 on language understanding general (GLUE) and biomedical (BLUE) tasks. An independent t-test is used to calculate statistical significance ($P < 0.05$) denoted by †. Metrics are given in Appendix 8.2.

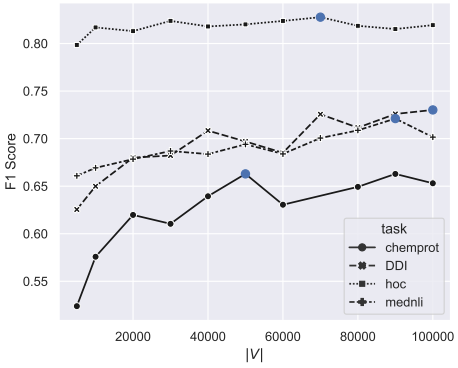

Figure 1: Evaluation of the biomedical tasks against varied $|V|$. A bold blue circle indicates the peak of the curve.

(Table 1). This indicates that reducing sub-word overlap does not reduce vocabulary usage and that downstream fine-tuning with a larger vocabulary size alleviates overlap and improves performance.

However, we find that few tasks perform best with a maximal separation of the biomedical and general domain vocabulary, with the only tasks performing well are CoLA (grammar detection), MedNLI (inference classification) and DDI (relation extraction). This suggests that a degree of overlap in a cross-domain is beneficial and that these domains share similarities. This shared similarity is also observed by Toews and Holland (2019).

BLUE tasks seem to benefit from a larger separation of vocabularies, as suggested by an improved F1-score with increased vocabulary size

($|V|$) in Figure 1. However, this benefit is less significant for GLUE tasks, as validation model selection (used in BLUE) could not be applied.

We find that GLUE results are worse when using combined (PubMed+Wiki) rather than individual pretraining corpora (see Table 3), while interestingly, the opposite appears to be true for BLUE. However, both benchmarks together show that the pretraining data and a larger vocabulary size helps in a cross-domain setting. Though it does not significantly *hurt* performance in the general domain, it significantly *improves* performance in the biomedical domain. Interestingly, pretraining with PubMed alone performed worse than pretraining with the Wikipedia corpus. A detailed table of results can be found in Appendix 5.

We observe that inference tasks fared better with a larger vocabulary (Table 4), indicating that inference tasks are more affected by sub-word overlap. For textual entailment (RTE) and paraphrase detection (QQP), larger $|V|$ had no positive effect. For SST-B (Textual Similarity) the model overfits as data size is small compared to the other tasks. Furthermore, while the default $|V|$ in transformers is 30,000, only a few tasks perform well at this size, suggesting that $|V|$ is an important consideration during pretraining depending on downstream task.

6 Limitations

This study only considers the biomedical and general domains; we hypothesize these principles can be applied to other domains, such as multilingual machine translation. One particular observation relevant to our setup is that the general domain corpus is smaller than that of the target domain, which should also be considered when extrapolating our findings. Another limitation is that training of the language models was not performed to completion. However, language modeling effectiveness was fixed for a fair comparison. These limitations will be explored in future work.

We are also aware that the fixed perplexity does not fully disentangle the impacts of vocabulary overlap and vocabulary size on the downstream effectiveness. We plan to extend our study with further experiments to ensure the robustness of results presented here.

7 Conclusions

When applying general domain Transformer language models to specialized ones, the use of sub-word modeling results causes sub-word overlap leading to decreased performance. We showed that increasing the vocabulary size of the model alleviates this performance penalty and improves downstream task performance on GENERAL and BIOMEDICAL benchmarks. Furthermore, we show that specialized domains improve significantly from a combination of specialized and general domain pretraining data. Our work is a pilot study into improving downstream performance on specialized domains with potential application in cross-domain tasks. In the future, we would extend this study to other applications such as machine translation and cross-lingual language modeling.

Acknowledgements

Vincent is supported by the Australian Research Training Program and the CSIRO Research Office Postgraduate Scholarship. This work is funded by the CSIRO Precision Health Future Science Platform.

References

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.

Giuseppe Attardi. 2015. Wikiextractor. https://github.com/attardi/wikiextractor.

Simon Baker, Ilona Silins, Yufan Guo, Imran Ali, Johan Högberg, Ulla Stenius, and Anna Korhonen. 2015. Automatic semantic classification of scientific literature according to the hallmarks of cancer. *Bioinformatics*, 32(3):432–440.

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.

Ilias Chalkidis, Manos Fergadiotis, Sotiris Kotitsas, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. An empirical study on large-scale multi-label text classification including few and zero-shot labels. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 7503–7515, Online.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*, pages 4171–4186, Minneapolis, MN.

Alvar EllegaRd. 1960. Estimating vocabulary size. 16:219–244.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing. *Computing Research Repository*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online.

María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. 2013. The ddi corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of Biomedical Informatics*, 46(5):914 – 920.

Paul Jaccard. 1912. The distribution of the flora in the alpine zone.1. *New Phytologist*, 11:37–50.

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

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016b. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035.

Martin Krallinger, O. Rabal, S. A. Akhondi, M. Pérez, J. Santamaría, Gael Pérez Rodríguez, G. Tsatsaronis, Ander Intxaurrondo, J. A. López, Umesh Nandal, E. V. Buel, A. Chandrasekhar, Marleen Rodenburg, A.G Lægreid, Marius A. Doornenbal, J. Oyarzábal, A. Lourenço, and A. Valencia. 2017. Overview of the biocreative vi chemical-protein interaction track. In *Proceedings of BioCreative*, pages 141–146.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *Proceedings of the 8th International Conference on Learning Representations*.

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*, 36(4):1234–1240.

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. volume 24 of *Psychology of Learning and Motivation*, pages 109–165. Academic Press.

Vincent Nguyen, Sarvnaz Karimi, and Zhenchang Xing. 2019. Investigating the effect of lexical segmentation in transformer-based models on medical datasets. In *Proceedings of the 17th Annual Workshop of the Australasian Language Technology Association*, pages 165–171, Sydney, Australia. Australasian Language Technology Association.

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.

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 2020. Inexpensive domain adaptation of pretrained language models: Case studies on biomedical NER and covid-19 QA. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1482–1490, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany.

Hoo-Chang Shin, Yang Zhang, Evelina Bakhturina, Raul Puri, Mostofa Patwary, Mohammad Shoeybi, and Raghav Mani. 2020. BioMegatron: Larger biomedical domain language model. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4700–4706, Online.

Yuqi Si, J. Wang, H. Xu, and Kirk Roberts. 2019. Enhancing clinical concept extraction with contextual embedding. *Journal of the American Medical Informatics Association : JAMIA*.

Chin Man Yeung Tai. 2019. Effects of inserting domain vocabulary and fine-tuning bert for german legal language. Master's thesis, University of Twente, Netherlands.

Wen Tai, H. T. Kung, Xin Dong, Marcus Comiter, and Chang-Fu Kuo. 2020. exBERT: Extending pre-trained models with domain-specific vocabulary under constrained training resources. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1433–1439, Online. Association for Computational Linguistics.

Daniel Toews and Leif Van Holland. 2019. Determining domain-specific differences of polysemous words using context information. In *Proceedings of the 25th International Working Conference on Requirement Engineering: Foundation for Software Quality*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Computing Research Repository*, abs/1706.03762.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. 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.

Hai Wang, Dian Yu, Kai Sun, Jianshu Chen, and Dong Yu. 2019b. Improving pre-trained multilingual model with vocabulary expansion. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 316–327, Hong Kong, China. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *Computing Research Repository*.

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. 2016a. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation.

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. 2016b. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *Computing Research Repository*, page arXiv:1609.08144.

Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. 2020. Large batch optimization for deep learning: Training bert in 76 minutes. In *Proceedings of the 7th International Conference on Learning Representations*, Online.

George Zipf. 1936. *The Psychobiology of Language*. London, Routledge.

benchmark	collection	dataset	value
BLUE	Wikipedia+PubMED	Chemprot	60.0
		DDI	72.7
		HoC	83.1
		MedNLI	71.5
	PubMED	Chemprot	51.7
		DDI	66.3
		HoC	81.7
		MedNLI	68.4
	Wikipedia	Chemprot	56.6
		DDI	69.3
		HoC	82.0
		MedNLI	70.9
GLUE	Wikipedia+PubMED	MRPC	66.9
		QNLI	81.2
		QQP	85.8
		RTE	52.7
		SST-2	84.0
		WNLI	43.7
	PubMED	MRPC	65.0
		QNLI	79.7
		QQP	86.4
		RTE	51.3
		SST-2	85.0
		WNLI	56.3
	Wikipedia	MRPC	70.8
		QNLI	79.1
		QQP	86.3
		RTE	50.5
		SST-2	82.5
		WNLI	56.3

Table 5: Expanded results from Table 3.

8 Determining vocabulary size

Prior to pretraining, when building the wordpiece tokenizer. We estimated the upper limit of unique vocabulary tokens based on the assumptions that: (1) each corpora is english; (2) each corpora shares no tokens; and, (3) the corpora's token frequency follows a zipf distribution (Zipf, 1936). From, (EllegaRd, 1960) (Table 5), we calculated the upper limit for the vocabulary for each corpus given our second assumption and summed the result which gives a combined vocabulary size of approximately 90,000. We extend the vocabulary by an extra 10,000 to determine if our vocabulary size was sufficient.

8.1 Pretraining Data Experiments

We train separate models for each corpora, namely Wikipedia, PubMed and the combined corpora of Wikipedia and PubMed. We use the same training procedure as in our main experiments, but at a fixed vocabulary size of 40,000.

8.2 Downstream Tasks

We use the standard GLUE benchmark tasks and the BLUE language understanding tasks. We describe the BLUE tasks as follows:

Relation Extraction DDI (Herrero-Zazo et al., 2013), is a medical corpus consisting of texts from the Drugbank database and MeEDLINE abstracts annotated by experts for drug-drug interactions.

Chemprot (Krallinger et al., 2017), a classification task for five different chemical-protein interaction categories from PubMed abstracts.

Multilabel classification Hallmarks of Cancers (HoC) (Baker et al., 2015), a corpus of PubMed abstracts labeled with one or more of ten cancers.

Inference For inference-based tasks, we use Medical Natural Language Inference (MedNLI) (Johnson et al., 2016a) created from MIMIC-III (Johnson et al., 2016b) and annotated by radiologists with entailment, neutral and contradiction labels for each premise-hypothesis pair.

Metrics Generally, for the BLUE tasks, we use macro averaged F1-score, except for HoC where we report the micro averaged F1-score similar to that described in Peng et al. (2019). Evaluation of the GLUE benchmark is based on GLUE's official metrics (Wang et al., 2019a): F1-score for QQP and MRPC, Pearson and Spearman correlation for STS-B, Matthew's Correlation for CoLA, which measures binary agreement between prediction and observed from -1 (total disagreement) and +1 (perfect prediction), and accuracy for the remaining tasks.

9 Minimizing Sub-word overlap

We describe the intuition behind the reduction in sub-word overlap in more detail here and discuss some results.

9.1 Definitions

Sub-word overlap is a phenomena wherein tokens in a sub-word model will exhibit a polysemous, though it is closer to homonymy, effect where sub-words will be shared by words that have different meanings. To combat this, we scale the vocabulary size, such that fewer sub-words are shared by different words. Chalkidis et al. (2020) also notes that in specialized contexts, general domain vocabularies tend to over-segment specialized terminology, such as diseases or medications.

$\|V\|$	Jaccard Similarity	Num. Overlaps	% Vocab Used	Num. Tokens used in GLUE tasks	Num. Tokens used in BLUE tasks
5000	94.6	4708	99.5	4970	4713
10000	87.8	8733	99.5	9893	8786
20000	73.8	14609	99.0	19418	14989
30000	62.8	18490	98.2	28457	19498
40000	54.6	21193	97.1	37057	22980
50000	48.2	23083	95.8	45239	25726
60000	43.0	24359	94.4	53109	27888
70000	38.9	25226	92.7	60545	29549
80000	35.6	25858	90.8	67563	30961
90000	32.8	26287	89.1	74369	32118
100000	30.4	26593	87.3	80842	33095

Table 6: Detailed results from Table 2, including statistics for the number of unique tokens used the BLUE and GLUE tasks.

9.2 Measuring Sub-word Overlap

We used Jaccard Index (Jaccard, 1912), to measure the set overlap between the GLUE and BLUE tasks. We found a decreasing trend in overlap when increasing vocabulary size, which was correlated with an increase in downstream task performance. We found that as vocabulary size increased, more vocabulary elements were used in terms of absolute quantities for both the GLUE and BLUE tasks. This could be attributed to fewer words being *broken up* into sub-word units as vocabulary size increases (Chalkidis et al., 2020).

9.3 Task Vocabulary Sizes

For each task, we used tokenized based on whitespace to approximate the vocabulary size needed to represent all words in at task (Table 7).

Task	Unique Vocabulary Elements
CoLA	1948
MNLI	13693
MRPC	3858
QNLI	17837
QQP	38260
RTE	4510
sst-2	4293
sts-b	7073
WNLI	592
biosses	362
Chemprot	12385
DDI	3280
HoC	8288
MedNLI	2840

Table 7: Unique vocabulary elements (whole words delimited by spaces)

9.4 Discussions

By expanding the vocabulary dimension, fewer overlaps will occur which is shown in Table 2 as a proportion of the overall vocabulary size and Jaccard Index. Though, in absolute terms the number of overlaps increase, suggesting that some overlap between domains does exist and the overlap percentage being approached is similar to the one found in Beltagy et al. (2019). This is further reflected in Table 3 where the GLUE tasks perform similarly when pretrained on either PubMED or Wikipedia. Suggesting that the pretraining data on its own has enough data to pretrain a general domain model.

Although this not hold true for the specialized domain, which requires both the general domain and specialized domain. Our intuition for pretraining on both Wikipedia and PubMED simultaneously is to reduce the catastrophic forgetting effect (McCloskey and Cohen, 1989), which may be present in models such as BioBERT (Lee et al., 2019), and ClinicalBERT (Alsentzer et al., 2019) given that the models are trained sequentially with medical corpora.