exKidneyBERT: A Language Model for Kidney Transplant Pathology Reports and the Crucial Role of Extended Vocabularies

by

Tiancheng Yang

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Statistics

Waterloo, Ontario, Canada, 2022

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Tiancheng Yang was the main author of the thesis. Professor Matthias Schonlau and Dr. Ilia Sucholutsky revised various parts of the thesis. Dr. Kuang-Yu Jen was the author of Table 2.1 and Table 2.2 and he improved the wording of parts of the abstract, and parts of Chapter 1.1, and Chapter 2.1.

Abstract

Background: Pathology reports contain key information about the patient's diagnosis as well as important gross and microscopic findings. These information-rich clinical reports offer an invaluable resource for clinical studies, but data extraction and analysis is often manual and tedious given their unstructured texts. Thus, an automated data extraction method from pathology reports would be of significant value and utility. Language modeling is useful for classifying and extracting information from natural language reports. Released in 2018, Bidirectional Encoder Representations from Transformers (BERT) achieved state-of-the-art performance on several natural language processing (NLP) tasks. Pre-training BERT to the task-specific domain usually improves the model performance. BioBERT was pre-trained with large biomedical corpora on BERT and outperformed BERT on biomedical NLP tasks. Clinical BERT pre-trained with clinical data on BioBERT achieved better results than BioBERT on clinical NLP tasks. It is not clear, however, whether pre-training on ever smaller training data sets is worthwhile.

Objective: to develop a language model for renal transplant-pathology reports to extract the answers for two pre-defined questions.

Methods: The study aimed to answer two pre-defined questions: 1) "What kind of rejection does the patient show?"; and 2)"What is the grade of interstitial fibrosis and tubular atrophy (IFTA)?". First, we followed the conventionally recommended procedure and pre-trained Clinical BERT further with the corpus which contains 3.4K renal transplant-reports and 1.5M words using Masked Language Modeling to obtain the Kidney BERT. Second, we hypothesize that the conventional pre-training procedure fails to capture the intricate vocabulary of narrow technical domains. We created extended Kidney BERT (exKidneyBERT) by extending the six words to the tokenizer of Clinical BERT and pre-trained with the same corpus as Kidney BERT on Clinical BERT. Third, all three models were fine-tuned with QA heads for the questions.

Results: For the first question regarding rejection, the overlap ratio at word level for exKidneyBERT (83.3% for antibody-mediated rejection (ABMR) and 79.2% for T-cell mediated rejection (TCMR)) beats that of both Clinical BERT and Kidney BERT (46.1% for ABMR, and 65.2% for TCMR). For the second question regarding IFTA, the exact match rate of exKidneyBERT (95.8%) beats that of Kidney BERT (95.0%) and Clinical BERT (94.7%),

Conclusion: When working in domains with highly specialized vocabulary, it is essential to extend the vocabulary library of the BERT tokenizer to improve model performance. In this case, pre-training BERT language models for kidney pathology reports improved model performance even though the training data were relatively small.

Keywords: natural language processing, NLP, transformer, BERT, QA, renal, kidney, pathology, deep learning

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Professor Matthias Schonlau. During my first co-op, he taught and guided me to open the gate of machine learning. In his senior class, he deepened my understanding of machine learning with lively and well-organized lectures. After I graduated, he helped me with the application for the master's program and supervised me on this project. Without his supervision and precious help, I could not have gone so far.

I would like to say many thanks to Dr. Ilia Sucholutsky. He can always give me so many impressive ideas and suggestions on coding, experimental design, and phrasing for the project.

I would like to say many thanks to Dr. Kuang-Yu Jen, who provided the precious real world clinical data for the project and revised the paper to make it professional and rigorous in the medical domain.

I would like to say many thanks to my family for supporting and encouraging me throughout my seven-year overseas study. Your love is the biggest motivation for me to move on.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Renal pathology reports contain crucial diagnostic information, often in an unstructured text format. With the development of deep learning, pre-trained language model such as Bidirectional Encoder Representations from Transformers (BERT) [5] was successfully applied to many different language domains. This project's goal was to develop a pre-trained language model for clinical reports of kidney transplants called Kidney BERT. This model was to be used for classifying and/or building a question-answering system to query reports provided by the pathology laboratory at the University of California, Davis. The reports contained diagnostic information as well as descriptive information regarding the light, immunofluorescence, and electron microscopy findings. A comment section that summarizes and interprets the findings and how they justify the diagnoses was also present in some cases. Two questions were of particular interest: "What kind of rejection does the patient show?" and "What is the grade of interstitial fibrosis and tubular atrophy (IFTA)?"

Pre-training involves training a language model on a large amount of text prior to considering the specific application of interest. Pre-trained language models have had tremendous success in recent years. In late 2018, an attention based pre-trained NLP model, deep bi-directional transformer model (BERT) [5], was released by Google. BERT has achieved state-of-the-art performance in a number of NLP GLUE tasks [16], which includes named entity recognition (NER), question and answering (QA)and sentiment classification. For general purpose NLP tasks, BERT is a leading choice.

For tasks in a specific domain of application, researchers usually pre-train BERT on a task-specific corpus to improve the prediction performance on the task. For example, clinical BERT [1] is a language model for texts in the medical domain, and it has been shown that Clinical BERT achieved better results on biological NLP tasks compared to the so-called vanilla BERT [1].

When data are available for a specific clinical subdomain of interest, we can pre-train clinical BERT further to adapt to that specific clinical subdomain. Cabernet [12] is a question-and-answer (QA) system based on Clinical BERT based on Moffitt [12] pathology reports which contains 276K reports with 196M words from Moffitt Cancer Center. The authors demonstrate that Cabernet is superior to clinical BERT on Cancer pathology reports.

We also pre-trained Clinical BERT further for the subdomain of renal pathology reports. However, compared to Cabernet, we had less data available to do so.

We noticed that both BioBERT and Clinical BERT use the default tokenizer of BERT, which will parse the out-of-bag (OOB) words into subwords. We tried to extend the six keywords in the two pre-defined questions into the tokenizer of Clinical BERT and pre-trained it to obtain a new model extended Kidney BERT (exKidneyBERT).

We found that exKidneyBERT outperformed both Clinical BERT and Kidney BERT.Thus we conclude that for the QA tasks, extending the keywords to the tokenizer will improve the model performance.

1.2 Prior Work

At the beginning, people often exploited rule-based system for QA, which parses the natural language input by semantic rules, and then matches the parsed output with some pre-stored answers. In 1995, [2] proposed an architecture for QA tasks by using deep NLP system combined with a rule-based semantic parser of natural language input and a database query and management system. The semantic parser was made of a parse tree and a semantic interpreter. The semantic interpreter took the output of the parse tree and generated a logical query which will be treated as an input of the NLP database. Moreover, a lexicon and a world model offered domain knowledge to the parser system, while the lexicon contained the domain specific word knowledge and the world model described the structure of the domain classes. Finally, the database could take the logical query from the parser as an input and return the query result as a potential answer for the natural language input. A similar system [14] called ExtrAns was applied to the medical data. The input questions

about genomics was parsed to semantic representations and then being matched with the most proper documents.

Later, deep neural network was more and more popular, and as an efficient method to extract features from the natural language document, recurrent neural network (RNN), especially long short-term memory (LSTM) [6] was generally used in different kinds of NLP tasks, includes QA. Based on LSTM, a novel architecture called match-LSTM was proposed and being exploited with a Pointer Net [17]. The model was made of two separate LSTM layers for processing inputs, and a Pointer Net to limit the output as a span of the given input. Similar approach was applied on medical corpus as well. In 2018, [20] designed an NLP framework named SeaReader for the MedQA dataset, which was created as a medical QA task based on real-world clinical medicine text materials. The SeaReader is consist of a input layer, context layer, dual-path attention layer, reasoning layer and integration and decision layer. Bi-directional LSTM was used in the context layer and reasoning layer for extracting the important feature from input and attention results. The architecture achieved a better accuracy on the MedQA task compare to several other models.

In recent years, with the development of attention and transformer [15], BERT [5] was proposed in 2019 by Google, and achieved state-of-the-art performance on a bunch of natural language understanding (NLU) tasks of General Language Understanding Evaluation (GLUE) benchmark [16], which includes QA task. BERT stacked multiple layers of transformer encoder and applied two unsupervised pre-trained tasks, masked language modeling (MLM) and next sentence prediction (NSP) to extract word and sentence level representation respectively. BioBERT [9] is a language model for biomedical language understanding. It was pre-trained on the PubMed abstracts with 18 billion words. BioBERT benefits from the pre-training process and beats BERT on multiple bio-NLP tasks such as biomedical NER, biomedical relation extraction (RE), and biomedical QA. Later, [1] proposed Clinical BERT, which is a language model for electronic medical records (EMR). Clinical BERT is tuned on the EMR notes of the Medical Information Mart for Intensive Care (MIMIC-III) dataset [7] which contains about 60,000 data points. A QA system for extracting data from cancer pathology reports, Cabernet [12], was built based on BERT. They first pre-trained Clinical BERT on 276k Moffitt pathology reports which contains 196M words and got a new model CancerBERT (caBERT), and then they tuned caBERT on QA task to retrieve key information from the pathology reports. Finally the extracted phrases were used as inputs of a classification net and being classified into different codes.

Chapter 2

Methodology

We prepared the data set (Section 2.1) and developed Kidney BERT by pre-training clinical BERT (Section 2.2). Then we extended the vocabulary used in the Clinical BERT by six keywords in the questions of QA tasks and pre-trained on Clinical BERT to obtained exKidneyBERT (Section 2.3). At the end, we fine-tuned Kindey BERT for question-answering and classification tasks (Section 2.4).

2.1 Data Set

The renal transplant-pathology reports were obtained from the electronic medical records of University of California, Davis. This study was determined to be exempt from the need for Internal Review Board approval since all information was de-identified at the source. The pathology reports were for transplant kidney biopsy cases, which consists of unstructured text for the diagnosis as well as light, immunofluorescence, and electron microscopy results as described by the pathologist. Each report contains following sections: Diagnosis, Tissues, Gross Description, and Microscopic Description.

Among all the information in the pathology reports, we were interested in the cases with rejection and the cases with IFTA. There are two major types of rejection for patients after kidney transplant, T-cell-mediated rejection (TCMR) and antibody-mediated rejection (ABMR) [11]. The pathology reports classify IFTA into 5 classes of severity: severe, moderate, mild, minimal, absent/insignificant. We define a sixth class as "unclassified", meaning the report contains no corresponding information. Our goal is to extract a part of sentence or phrases from the report which best describe the condition of rejection and IFTA.

Table 2.1: An illustrative example of the text used for the QA task. Bold texts are the expected answers. Italicized text were removed during training.

Comments: The biopsy shows interstitial inflammation (i2) consisting of mostly mononuclear leukocytes. Tubulitis (t2) is readily identified in the areas with infiltrating inflammatory cells. **These findings support the diagnosis of acute T-cell mediated rejection (IA)**. *Microscopic Description:* The following findings are based on hematoxylin and eosin (HE), periodic acid-Schiff (PAS), and Masson trichrome-stained sections. The specimen submitted for light microscopic evaluation consists of cortical tissue with at least 35 glomeruli. No segmentally or globally sclerosed glomeruli are seen. The glomeruli demonstrate focal mild mesangial widening. The glomerular capillary walls are of normal thickness and contours. Patchy moderate inflammation is noted associated with scattered moderate tubulitis. The inflammation consists predominantly of mononuclear leukocytes with coms plasma cells and only rare eosinophils. **Mild** interstitial fibrosis and tubular atrophy are present ($\tilde{1}0\%$). The arteries and arterioles show focal mild hyalinosis. No endotheliitis or peritubular capillaritis is identified.

Table 2.2: An illustrative example of the text used for the classification task. Italicized text were removed during training.

Microscopic Description: The following findings are based on hematoxylin and eosin (HE), periodic acid-Schiff (PAS), and Masson trichrome-stained sections. The specimen submitted for light microscopic evaluation consists of cortical tissue with at least 35 glomeruli. No segmentally or globally sclerosed glomeruli are seen. The glomeruli demonstrate focal mild mesangial widening. The glomerular capillary walls are of normal thickness and contours. Patchy moderate inflammation is noted associated with scattered moderate tubulitis. The inflammation consists predominantly of mononuclear leukocytes with coms plasma cells and only rare eosinophils. The arteries and arterioles show focal mild hyalinosis. No endotheliitis or peritubular capillaritis is identified.

For the classification task, we focused on the content in the Microscopic Description section since it includes the most detailed descriptions of the biopsy. We removed the text related to the task to avoid showing the correct answer in the input text. An example of the input text for the classification task is shown in Table 2.2. For the QA task, along with the text in the Microscopic Description section, we also added the section of the report comments as a part of the input text since they contain the description of the rejection cases explicitly and we expect the language model to retrieve the answer from the given text. Table 2.1 shows an example of the input text for the QA task.

2.2 BERT Pre-training and Kidney BERT

BERT stacks 12 layers (BERT-base model) and 24 layers (BERT-large model) of transformer encoder layers with bi-directional self-attention head inside [19]. BERT is pre-trained by two unsupervised tasks, masked language modeling and next sentence prediction, on the BooksCorpus [21] and English Wikipedia data. In the masked language modeling stage, 15% of the words in the text were replaced by a special token "[MASK]" to let the model learn and predict the masked word based on the context. More specifically, among the words selected for masking, only 80% of them were replaced by the special mask token. 10% of them are replaced with a random token and the rest 10% of them are remain the same. In order to let the model learn the relationshop of sentences, BERT introduced next sentence prediction as well. Two sentences are concatenated together by a special token "[SEP]". 50% of the time the second sentence is the actual next sentence, and the rest time it is chosen randomly. However, in the latest research [10], next sentence prediction was found not to be important.

Both BioBERT and Clinical BERT take advantage of the pre-training process. CaBERT further pre-trained on Clinical BERT with Moffitt pathology reports. They simply masked 15% of the words in the Moffitt dataset to a special token "[MASK]", and then trained the language model to predict these words [12]. While the performance of pre-trained CaBERT on the specific downstream tasks of interest was better than when just fine-tuning Clinical BERT, the performance on other tasks with more general corpora such as SQuAD and BioASQ had decreased. We suspect that there are tradeoffs in the pre-training process that depend on the available dataset and choice of downstream task. As a result, we use the pre-training process suggested by the caBERT authors on our renal pathology reports but conduct an ablation study to determine whether the additional pre-training step adds value. Also, our data set is much smaller than that used for caBERT: our data contain 3.4K reports with approximately 1.5M words; caBERT is based on 276K reports with 196M

words [12].

2.3 Extend Six Keywords for exKidneyBERT

Both BioBERT and Clinical BERT use WordPiece tokenization [19] to handle vocabulary not included among the approximately 30k words BERT trained on. This is called out-of-bag or OOB vocabulary. For example, the word "interstitial" will be parsed into frequent subwords "inter", "##st", "##iti", and "##ai" first, and then tokenized into vectors. Therefore we added the six key words which are parsed into subwords originally in the two pre-defined questions in the QA tasks "interstitial", "fibrosis", "tubular", "atrophy", "T-cell", and "antibody" to the tokenizer. Also, we needed to extend the embedding layer's dimension from 28996 to 29002 to match the newly added words. We decided to only extend the six keywords to the tokenizer because 1) these six words contain the most important information needed for the model to locate the answers; 2) extending a lot of words to the tokenizer may affect the pre-trained representative for the existing vocabulary. Since the model does not have any knowledge to the newly added six words, we did the same pre-trained procedure as Kidney BERT on Clinical BERT and obtain a new language model called extended Kidney BERT (exKidneyBERT).

2.4 Fine-tune BERT models for QA and Classification

Figure 2.1 shows the architecture we exploited for question answering (QA) by using BERT models. For each input, we concatenated "What kind of rejection does the patient show?" or "What is the grade of interstitial fibrosis and tubular atrophy?" to the microscopic description section of the reports together by the special token "[SEP]". We also added the special token "[CLS]" to the beginning of the concatenated text to follow the BERT usage convention. On top of each BERT model, we added a linear layer as a QA span classifier to the output embedding of BERT. The linear classifier layer will be fine-tuned with BERT simultaneously. During fine-tuning, the model will predict a start vector S and an end vector E. The probabilities of each word to be the start and end of the answer will be the outputs of vectors S and E after softmax [3] by the formula of $p_{S_i} = \frac{e^{S_i}}{\sum_j e^{S_j}}$ and $p_{E_i} = \frac{e^{E_i}}{\sum_j e^{E_j}}$. Next, we applied cross-entropy loss [4] to calculate the gradients:

$$-\sum_{\forall \hat{y}} 1(X, \hat{y}) \log(P(\hat{y}|X)),$$

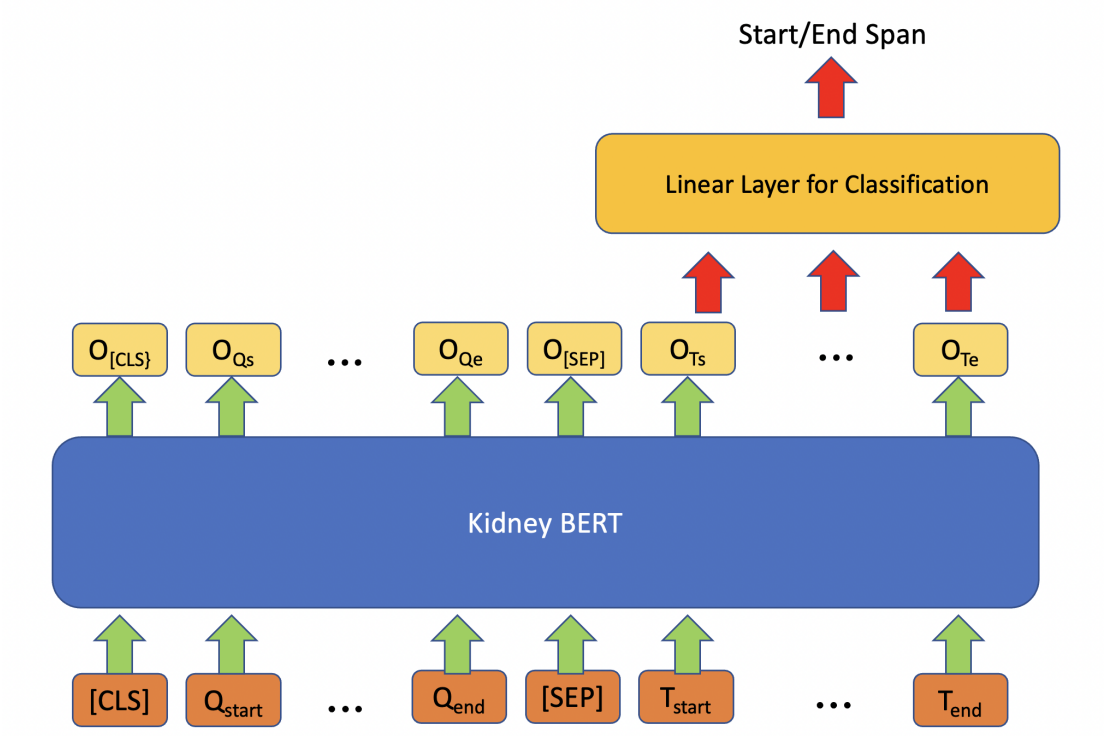

Figure 2.1: Architecture of Kidney BERT for the QA Task

where $1(X, \hat{y})$ is the binary indicator for whether or not the predicted label $\hat{y}$ matches the ground truth label for input X, and $P(\hat{y}|X)$ are the probabilities of the outputs from softmax. Then, we updated the parameters of BERT and the classification layer through backpropagation. The words with the maximum probability are chosen as the start and end of the answer text span. If the position of the end word is smaller than that of the start word, then "no information" will be predicted as an output.

In addition to QA, we also tried to use BERT models on the classification task for questions with multiple categories as expected answers. Figure 2.2 describes the architecture for it. Similar to that of QA, the classification model also exploits a linear classifier layer on top of the BERT models. However, this time we only use the output embedding corresponding to the special token "[CLS]" as the input of the classifier, and then the outputs of the classifier are converted into the probabilities through softmax. Cross-entropy loss is used as the loss function as well. We used huggingface transformer [18] as the BERT

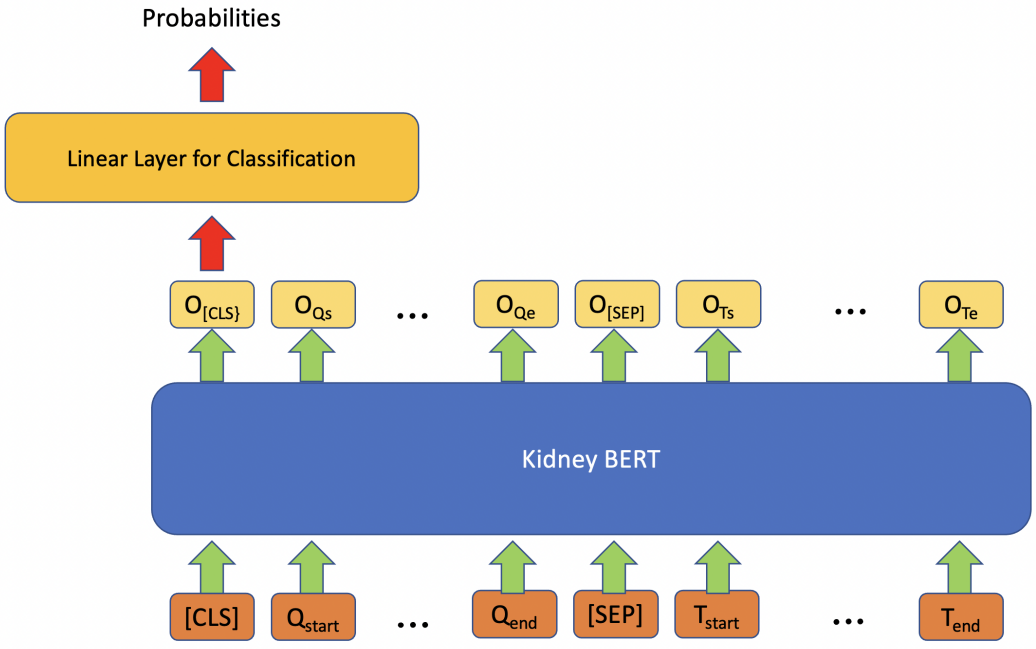

Figure 2.2: Architecture of Kidney BERT for the Classification Task

framework.

Chapter 3

Results

After introducing the metrics used for evaluating model performance (Section 3.1), we report on four results. First, we trained the models on rejection cases only (Section 3.2). Second, we trained the BERT models on all renal pathology reports for the classification tasks (Section 3.3). Third, we trained the BERT models on all the reports for the QA tasks (Section 3.4).

3.1 Metrics

For the first question, "What kind of rejection does the patient show?", we labelled the text span manually from the reports. A typical answer for the question is "No evidence of acute antibody-mediated rejection". Since the answers are quite long, we measured the overlap between the predicted text span and the ground truth answer. We calculated the overlap ratio of how much the two text spans overlap on a character level and word level respectively. For the second question, "What is the grade of interstitial fibrosis and tubular atrophy?", since the answers are one-word or two-word phrases, we only counted the prediction results which exactly matched the ground truth phrases. In this case F1-score was used as a measurement metric.

3.2 Training on a portion of reports - rejection cases

At the beginning, we focused only on the 242 reports with the rejection cases. Of these, 87 contain positive examples for TCMR. For simplicity, we converted the QA problem into

Table 3.1: Classification results for freezing Clinical BERT vs. fine-tuning Clinical BERT on the small TCMR sample. Acc. means accuracy. Log.Reg. means logistic regression. DNN means dense neural network.

Model	Overall Acc.	F1 of Positive
Frozen Clinical BERT+Log.Reg.	0.78	0.35
Frozen Clinical BERT+DNN	0.88	0.77
Fine-tuned Clinical BERT+DNN	0.92	0.85

a binary classification task (rather than predicting a text answer). The task is to predict whether or not the patient shows TCMR in the report. For the two baseline models, we froze the parameters of Clinical BERT and used (separately) logistic regression and linear neural network as a classifier to the embedding of the output sentence of BERT. Next, we fine-tuned a third model of Clinical BERT with a single layer dense neural network. Table 3.1 shows the results of the three models. We can see that by fine-tuning the classifier and Clinical BERT together, both overall accuracy and F1-score of the positive samples increased a lot.

3.3 Training on all reports - Classification

Next, we extended the data set to all 3.4K reports. Similar to the transfer learning process used for caBERT [12], we randomly selected and masked 15% of the words in all the reports and trained the Clinical BERT to predict those replaced words. After the pre-training process, we obtained Kidney BERT, our language model for renal pathology reports. Then we extended the six keywords to the tokenizer of Clinical BERT and redo the same pre-training prcedure as Kidney BERT on the 3.4k reports to obtain the exKidneyBERT. We fine-tuned all the BERT models in the pre-training chain includes the vanilla cased base BERT, BioBERT, Clinical BERT, Kidney BERT, and exKidneyBERT on the rejection classification tasks. Also, we added a second task, grade classification of IFTA. The results are shown in Table 3.2.

Table 3.2: Classification results for fine-tuning BERT models on the full data set. 'CLS' means classification task.

Model	Task	Overall Acc.	Positive F1-score
BERT	Rej. CLS	0.945	0.000
BioBERT	Rej. CLS	0.953	0.515
Clinical BERT	Rej. CLS	0.977	0.750
Kidney BERT	Rej. CLS	0.977	0.765
exKidneyBERT	Rej. CLS	**0.978**	**0.800**
Model	Task	Overall Acc.	Weighted F1-score
BERT	IFTA CLS	0.768	0.764
BioBERT	IFTA CLS	**0.788**	**0.789**
Clinical BERT	IFTA CLS	**0.788**	0.788
Kidney BERT	IFTA CLS	0.785	0.785
exKidneyBERT	IFTA CLS	0.782	0.780

3.4 Training on all reports - QA

After exploring classification tasks, we next considered question answering (QA). We manually tagged the desired answer phrases of ABMR and TCMR in each report for the question "What kind of rejection does the patient show?". For the question of IFTA, "What is the grade of interstitial fibrosis and tubular atrophy?", we tagged any mention of the six outcome classes as expected answers. We fine-tuned all the BERT models again and each model is attached with a QA head. For the QA tasks, question and text are concatenated as the model input. Table 3.3 and Figure 3.1 to Figure 3.3 show the results.

Table 3.3: QA results for fine-tuning BERT models on the full data set. Overlap Ratio Char and Overlap Ratio Word means the overlap length between the prediction answer and the expected answer divided by the length of the expected answer, at character level and word level, respectively. Exact Match Rate means a perfect match between the prediction answer and the expected answer.

Model	Task	Overlap Ratio Char	Overlap Ratio Word
BERT	ABMR QA	0.442	0.616
BioBERT	ABMR QA	0.519	0.667
Clinical BERT	ABMR QA	0.363	0.461
Kidney BERT	ABMR QA	0.363	0.461
exKidneyBERT	ABMR QA	**0.604**	**0.833**
BERT	TCMR QA	0.494	0.653
BioBERT	TCMR QA	0.494	0.653
Clinical BERT	TCMR QA	0.494	0.653
Kidney BERT	TCMR QA	0.494	0.653
exKidneyBERT	TCMR QA	**0.664**	**0.792**
Model	Task	Exact Match Rate	
BERT	IFTA QA	0.942	
BioBERT	IFTA QA	0.956	
Clinical BERT	IFTA QA	0.947	
Kidney BERT	IFTA QA	0.950	
exKidneyBERT	IFTA QA	**0.958**	

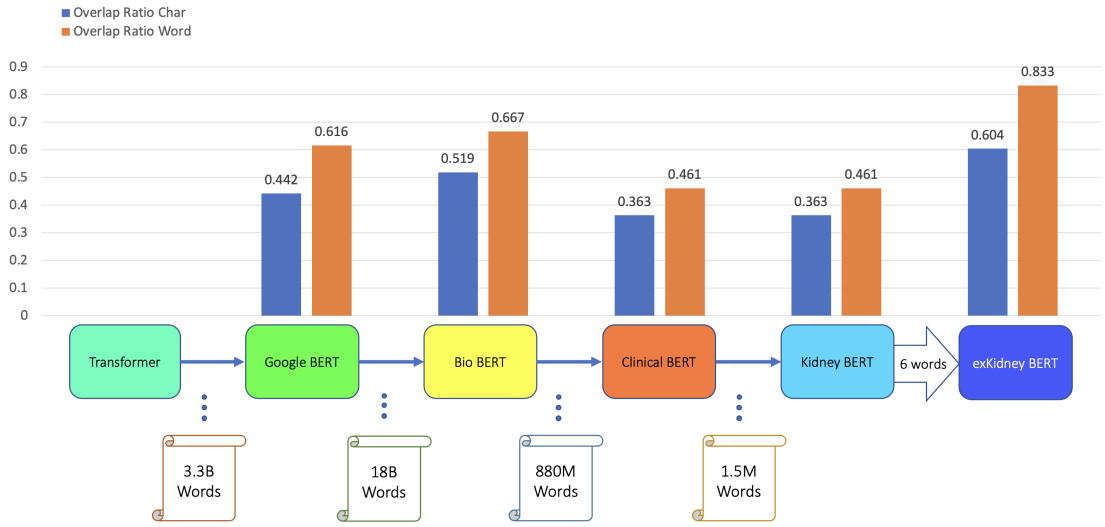

Figure 3.1: BERT Models' Results of the QA task for ABMR

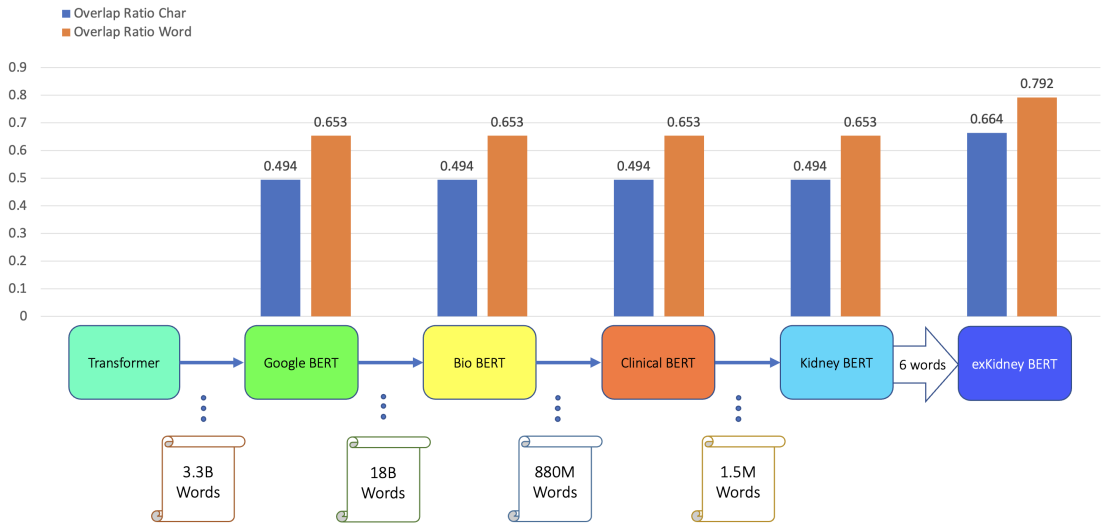

Figure 3.2: BERT Models' Results of the QA task for TCMR

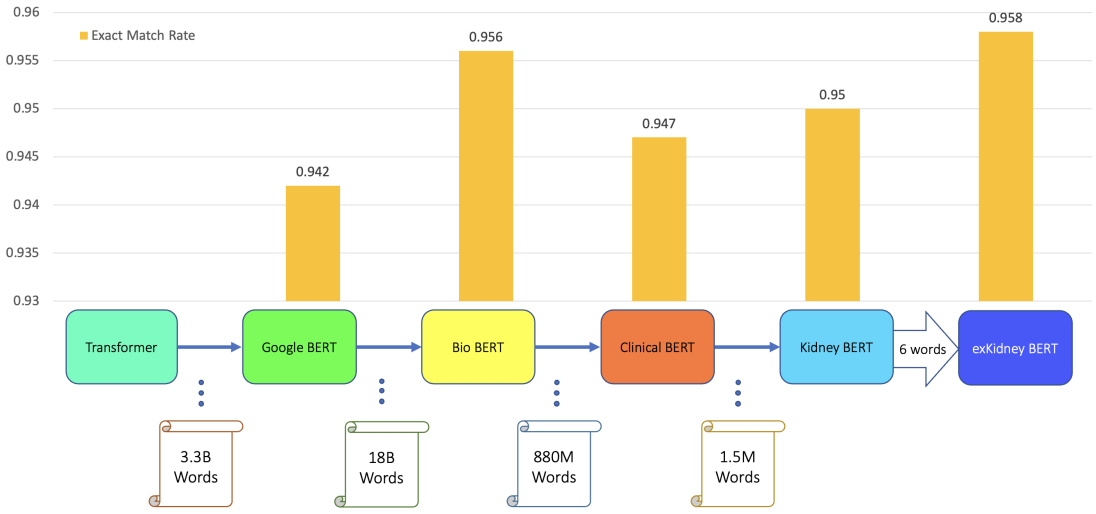

Figure 3.3: BERT Models' Results of the QA task for IFTA

Chapter 4

Discussion

4.1 Principal Results

First, we found that by extending the six keywords in the questions of the QA tasks, exKidneyBERT outperforms the other BERT models. We compared five BERT models in total. BioBERT was pre-trained with the PubMed corpus on the cased base BERT model. Clinical BERT was pre-trained with the MIMIC-III dataset on BioBERT. We created Kidney BERT by pre-training with our renal pathology data on Clinical BERT and we developed exKidneyBERT by extending the tokenizer of Clinical BERT with six keywords in the two questions in our QA tasks and pre-training with our data on Clinical BERT. We compared the five BERT models' performance on classification tasks and QA tasks. Recall that we removed the words related to the target which are the explicit answers in the input text of the classification tasks, and so that the input text does not contain the six words extended to the exKidneyBERT. In the classification task of rejection case, the exKidneyBERT performed the best on both overall accuracy and F1-score of positive samples. But for the classification task of IFTA, the exKidneyBERT performs the second worst and the result of BioBERT beats others. The results for the classification tasks could be a baseline to measure how much the exKidneyBERT benefits from word-extension on QA tasks. For the QA tasks of ABMR and TCMR, the exKidneyBERT outperforms other four BERT models on both overlap ratio at character level and word level. Notice that in the TCMR case, the other four BERT models were stuck at 0.494 of characters' overlap ratio and 0.653 of words' overlap ratio while exKidneyBERT broke the barrier and achieved 0.664 and 0.792 on overlap ratio at character level and word level, respectively. For the QA tasks on IFTA, unlike the classification case which exKidneyBERT performed

the second worst, this time exKidneyBERT achieved the best result among all the five BERT models. This is evidence that the the contribution comes from the six extended words being present in the vocabulary. This is also consistent with our hypothesis, that extending the vocabulary is what improves performance, because when the data was pre-processed for the IFTA classification task, we striped out any sentences that contain the six words, which means the model was fitted to the extended words.

Second,we performed an ablation study to determine which modeling components contributed to the performance increase. We found that the masked language modeling pre-training on an increasingly small domain-specific text corpus without extending the vocabulary did not improve the performance in our domain. Previous language models like Clinical BERT and Cancer BERT suggest that when adapting BERT to a particular domain, the BERT model will benefit from pre-training with the domain-specific corpus. We tried this approach in a comprehensive ablation study and found that pre-training on a small domain-specific corpus for renal pathology reports is ineffective. By comparing the result in Table 3.2 and Table 3.3, we can see that on the classification task, the results for the model based on Kidney BERT is the same as that on Clinical BERT in overall accuracy, and only 0.015 higher in F1-score of positive samples. On the QA tasks, the results for Clinical BERT and Kidney BERT are same on the rejection tasks, and the exact match rate of IFTA task with Kidney BERT is only 0.003 higher than Clinical BERT. In addition, fine-tuning was beneficial based on the results shown in Table 3.1.

Third, we found that in the domain of our dataset for renal pathology reports, BioBERT was better than Clinical BERT. We fine-tuned the BERT models on five tasks in total, except the classification task of rejection case, the results of BioBERT is better than that of Clinical BERT on other four tasks. A possible reason is that the Clinical BERT was pre-trained on a different domain than our dataset while BioBERT was pre-trained on a more general domain.

4.2 Limitations

First, exKidneyBERT were designed to answer the two pre-defined questions only. For exKidneyBERT, we extended the six keywords in the two questions of the QA tasks we wanted to resolve. As always, if we desired to solved other QA tasks we need to train new models for them.

Second, the dataset we used is small compared to the other BERT models we compared to. Google BERT was pre-trained on 3.3 billion words, BioBERT was pre-trained on 18

billion words, Clinical BERT was pre-trained on 880 million words, and the dataset we used for pre-training only contains 196 million words. However, in order to investigate pre-training when data are scarce, we have to work with a small data set.

4.3 Comparison with Prior Work

We followed the exact same unsupervised pre-training procedure as the Cancer BERT did [12] to develop Kidney BERT, which initialize the model parameters from Clinical BERT and randomly selected 15% of the words and replaced them with a special token "[MASK]" and then train the model to predict the masked tokens. In addition, we tried to extend the six keywords in the questions of the QA tasks to the BERT tokenizer and repeat the same pre-training procedure as Cancer BERT on our dataset to create exKidneyBERT. We found that exKidneyBERT performs better than that of Kidney BERT, which is an improvement compare to the procedure of Cancer BERT.

4.4 Conclusion

We have made three primary contributions. First, we developed exKidneyBERT, a language model with an extended vocabulary of six keywords and specific to renal pathology reports. ExKidneyBERT outperformed in the QA tasks. Second, we conducted an ablation study and found that BERT model performance does not benefit from pre-training on our dataset, which is a small amount of renal pathology reports by comparing the results of Kidney BERT and Clinical BERT. Third, we found that in our renal pathology dataset, BioBERT performed better than the Clinical BERT on the five NLP tasks.

4.5 Conflicts of Interest

None declared.

References

[1] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

[2] Ion Androutsopoulos, Graeme D Ritchie, and Peter Thanisch. Natural language interfaces to databases–an introduction. *Natural language engineering*, 1(1):29–81, 1995.

[3] John Bridle. Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. *Advances in neural information processing systems*, 2, 1989.

[4] David R Cox. The regression analysis of binary sequences. *Journal of the royal statistical society: series B (methodological)*, 20(2):215–232, 1958.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[6] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. 9(8):1735–1780, 1997.

[7] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[8] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[9] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[10] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[11] A Loupy, M Haas, K Solez, Racusen L, D Glotz, D Seron, BJ Nankivell, RB Colvin, M Afrouzian, E Akalin, N Alachkar, S Bagnasco, JU Becker, L Cornell, C Drachenberg, D Dragun, H de Kort, IW Gibson, ES Kraus, C Lefaucheur, C Legendre, H Liapis, T Muthukumar, V Nickeleit, B Orandi, W Park, M Rabant, P Randhawa, EF Reed, C Roufosse, SV Seshan, HK Sis, Ba nd Singh, C Schinstock, A Tambur, A Zeevi, and M. Mengel. The banff 2015 kidney meeting report: Current challenges in rejection classification and prospects for adopting molecular pathology. *American journal of transplantation : official journal of the American Society of Transplantation and the American Society of Transplant Surgeons, 17(1), 28–41.*, 2017.

[12] Joseph Ross Mitchell, Phillip Szepietowski, Rachel Howard, Phillip Reisman, Jennie D Jones, Patricia Lewis, Brooke L Fridley, and Dana E Rollison. A question-and-answer system to extract data from free-text oncological pathology reports (CancerBERT network): Development study. *Journal of medical internet research*, 24(3):e27210, 2022.

[13] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: $100,000+$ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

[14] Fabio Rinaldi, James Dowdall, and Gerold Schneider. Answering questions in the genomics domain. 2004.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[16] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

[17] Shuohang Wang and Jing Jiang. Machine comprehension using match-LSTM and answer pointer. *arXiv preprint arXiv:1608.07905*, 2016.

[18] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020.

[19] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[20] Xiao Zhang, Ji Wu, Zhiyang He, Xien Liu, and Ying Su. Medical exam question answering with large-scale reading comprehension. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, pages 5706–5713, 2018.

[21] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27, 2015.