DePaul University
Via Sapientiae

College of Computing and Digital Media Dissertations

College of Computing and Digital Media

Winter 11-17-2020

BERT efficacy on scientific and medical datasets: a systematic literature review

Clayton Cohn
DePaul University, clayton@claytoncohn.com

Follow this and additional works at: https://via.library.depaul.edu/cdm_etd

Part of the Artificial Intelligence and Robotics Commons

Recommended Citation

Cohn, Clayton, "BERT efficacy on scientific and medical datasets: a systematic literature review" (2020). *College of Computing and Digital Media Dissertations*. 24.
https://via.library.depaul.edu/cdm_etd/24

This Thesis is brought to you for free and open access by the College of Computing and Digital Media at Via Sapientiae. It has been accepted for inclusion in College of Computing and Digital Media Dissertations by an authorized administrator of Via Sapientiae. For more information, please contact digitalservices@depaul.edu.

BERT EFFICACY ON SCIENTIFIC AND MEDICAL DATASETS: A SYSTEMATIC LITERATURE REVIEW

BY

CLAYTON COHN

A THESIS SUBMITTED TO THE SCHOOL OF COMPUTING, COLLEGE OF
COMPUTING AND DIGITAL MEDIA OF DEPAUL UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE
OF
MASTER OF SCIENCE IN COMPUTER SCIENCE

DEPAUL UNIVERSITY
CHICAGO, ILLINOIS
2020

DePaul University
College of Computing and Digital Media

MS Thesis Verification

This thesis has been read and approved by the thesis committee below according to the requirements of the School of Computing graduate program and DePaul University.

Name: **Clayton Cohn**

Title of dissertation: **BERT Efficacy on Scientific and Medical Datasets: A Systematic Literature Review**

Date of Dissertation Defense: 11/17/2020

Name (not signature)

Peter Hastings

Name (not signature)

Noriko Tomuro

Name (not signature)

Roselyne Tchoua

** A copy of this form has been signed, but may only be viewed after submission and approval of FERPA request letter.*

© 2020 Clayton Cohn

ALL RIGHTS RESERVED

ABSTRACT

Bidirectional Encoder Representations from Transformers (BERT) [Devlin et al., 2018] has been shown to be effective at modeling a multitude of datasets across a wide variety of Natural Language Processing (NLP) tasks; however, little research has been done regarding BERT's effectiveness at modeling domain-specific datasets. Specifically, scientific and medical datasets present a particularly difficult challenge in NLP, as these types of corpora are often rife with technical jargon that is largely absent from the canonical corpora that BERT and other transfer learning models were originally trained on. This thesis is a Systematic Literature Review (SLR) of twenty-seven studies that were selected to address the various methods of implementation when applying BERT to scientific and medical datasets. These studies show that despite the datasets' esoteric subject matter, BERT can be effective at a wide range of tasks when applied to domain-specific datasets. Furthermore, these studies show that the addition of domain-specific pretraining, either through additional pretraining or the utilization of domain-specific BERT derivatives such as BioBERT [Lee et al., 2019], can further augment BERT's performance on scientific and medical texts.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Professor Peter Hastings, for guiding me throughout the thesis process, continually inspiring and motivating me, and being a member of my defense committee. I would also like to thank Professor Hastings’ doctoral student, Keith Cochran, whose many insights and collaborations aided my writing process significantly. I would like to thank my advisor’s former doctoral student (now PhD), Simon Hughes, for his work that became the basis for my research and also for his contributions on an additional collaborative effort. I would also like to thank Professor Noriko Tomuro for serving on my defense committee and for being a willing wealth of knowledge (NLP and otherwise) both inside and outside of the classroom. I would like to thank Professor Roselyne Tchoua for serving on my defense committee. Lastly, I would like to thank DePaul University and the College of Computing and Digital Media for providing a safe, fun, and (above all) enlightening atmosphere for the years I spent as both an undergraduate and graduate student at DePaul University.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

With the advent of contextual word embeddings, transfer learning has become the status quo for modeling many NLP tasks. Like Computer Vision's ImageNet [Deng et al., 2009], models such as ELMo (Embeddings from Language Models) [Peters et al., 2018], ULMFiT (Universal Language Model Fine-Tuning) [Howard and Ruder, 2018], GPT-3 (Generative Pre-trained Transformer) [Brown et al., 2020], and BERT allow users to leverage pretrained networks and fine-tune them toward a downstream task. BERT, specifically, achieved state-of-the-art (SOTA) results on eleven NLP tasks when it was first released in October 2018 [Devlin et al., 2018]. Although largely considered to be robust to variations in subject matter across datasets, fine-tuning BERT on documents whose terms are largely under-represented (or absent) from BERT's training corpora (Wikipedia (en.wikipedia.org) and BookCorpus [Zhu et al., 2015]) often yields results that leave considerable room for improvement [Peng et al., 2019] [Alsentzer et al., 2019]. As such, additional measures must often be taken to adapt BERT to domain-specific datasets.

For the purposes of this thesis, the term "domain-specific" refers to datasets whose subject matter is homogeneous. A corpus of medical diagnoses, for example, would be domain-specific, whereas a corpus of tweets would not typically be considered domain-specific due to the potentially variegated subject matter. The focus of this thesis is to analyze and evaluate current approaches to modeling domain-specific datasets with BERT in the realms of science and medicine and to compare the efficacies of the various approaches. Twenty-seven relevant studies were gathered via SLR. The results of each study were evaluated based on task, model type (i.e., the specific method of BERT implementation), and dataset language.

This review is meant to inform readers of the most effective methodologies of applying BERT to scientific and medical texts across a variety of tasks and languages. The methods herein are worth considering for any researcher aiming to model a dataset whose terms are largely absent in the vocabularies of more general language models. Furthermore, a strong language model (such as BERT) can go a long way in improving performance on esoteric datasets relative to other deep learning approaches. By experimenting with BERT and the methods described in this thesis, one has a reasonable expectation of augmenting his or her performance on tasks conducted on scientific and medical data.

1.1 Background

BERT is a novel architecture that performs a wide variety of NLP tasks. Due to the newness of the model, its current body of research is limited. This study seeks to answer questions pertaining to BERT and its performance that have so far been largely absent from the current body of BERT-related works. Most of the current research assesses BERT's efficacy at modeling datasets whose subject matter is in the general domain, and little research has been done with regard to how effective BERT is when applied to domain-specific data. Domain-specific texts in the fields of law (court reports), finance (earnings statements), and science (conference proceedings) are examples of texts currently outside the purview of most BERT-related research. To help remedy this, this review aims to ascertain whether or not BERT can be as effective on domain-specific data as it is on data in the general domain.

The reason that a domain-specific review is needed, as opposed to a review of all BERT-related research, is that the current body of BERT-related research is mostly limited to datasets whose vocabularies are similar to that of BERT. Most of the major benchmarks that BERT is measured against are comprised of datasets whose instances were generated from the same (or similar) corpora that BERT was initially trained on. Examples of these benchmarks include SQuAD (Stanford Question Answering Dataset) [Rajpurkar et al., 2016], GLUE (General Language Understanding Evaluation) [Wang et al., 2018], MultiNLI (Multi-Genre Natural Language Inference) [Williams et al., 2018], and SNLI (Stanford Natural Language Inference) [Bowman et al., 2015]. While researchers agree that BERT has been an impactful architecture, no one has yet undertaken a survey with the goal of determining BERT's robustness to datasets whose vocabularies differ significantly from that of its own training corpora. Therefore, it is the goal of this review to determine whether or not BERT is as effective at modeling esoteric texts as it is at modeling canonical ones.

In the limited body of domain-specific BERT research, some domain-specific BERT-derivatives have emerged in the realms of science and medicine. While these models have been shown to be effective when applied to datasets within their respective studies, no research has evaluated the efficacies of these domain-specific models relative to "standard" BERT (i.e., a base BERT model that has not undergone domain-specific pretraining) across a wide range of tasks and datasets. This review seeks to obtain an "apples-to-apples" comparison of standard BERT's performance relative to that of some of its domain-specific counterparts across a multitude of tasks and datasets.

Furthermore, little research has been conducted with the goal of specifically identifying the problems encountered when applying BERT (and its domain-specific derivatives) to domain-specific datasets. As such, this review will also address the most prevalent issues encountered by researchers when applying BERT to scientific and medical texts. Lastly, BERT's performance on non-English scientific and medical datasets will also be evaluated, as this has similarly not been explored in current research.

Overall, this review seeks to obtain a comprehensive determination as to whether BERT can be effective at modeling datasets that are outside the general domain. If BERT is shown to be able to be successfully adapted for scientific and medical datasets, it is possible that BERT could similarly be adapted to other non-general domains as well. Furthermore, identifying effective methods of applying BERT to scientific and medical datasets, as well as identifying the problems BERT encounters while being applied to datasets in these domains, will provide a basis from which to conduct further research.

1.2 Research Questions

The research questions (RQs) for this SLR were formulated based on this review's objective of evaluating the current BERT implementations for scientific and medical datasets. Because BERT is a novel architecture, there is a limited amount of research regarding BERT's specific adaptation to datasets whose data are similar in subject matter. The domains of science and medicine were chosen because these fields contain the largest number of usable, domain-specific studies. This thesis seeks to answer the following questions:

- *RQ1: How does the performance of BERT-based approaches for modeling scientific and medical datasets compare to the performances of other approaches?*
- *RQ2: How does BERT's performance on scientific and medical datasets compare to further pretrained, domain-specific BERT derivatives (e.g. BioBERT and SciBERT)?*
- *RQ3: What types of problems does BERT encounter when presented with scientific and medical datasets?*
- *RQ4: How well does BERT perform on scientific and medical datasets that are in languages other than English?*

The following chapter will discuss the current body of research pertinent to answering these questions. Additionally, the research questions are accompanied by corresponding hypotheses and justifications that can be found in Chapter 3.

CHAPTER 2

RELATED WORK

This section denotes the origins of BERT and the previous approaches for modeling datasets in NLP. Additionally, the advent of transfer learning in NLP is discussed, as is BERT's architecture and the current (literature) reviews that evaluate BERT's application to domain-specific datasets.

2.1 Origins

The following subsections discuss the origins of BERT, the influential technologies that comprise it, and the impact BERT has had on NLP research.

2.1.1 Deep Learning, LSTM (Long Short-Term Memory), and ELMo

Prior to 2018, most SOTA NLP benchmarks used deep learning models, particularly Recurrent Neural Networks (RNN). Considering how important word order is when determining the meaning of a sentence, RNNs appeared particularly well-suited for NLP. RNNs are designed to account for sequence order, as the hidden state at any time-step is dependent upon the hidden-states of all of the previous time-steps. Conversely, Standard feed-forward networks are blind to previous states and can only "feed forward." While RNNs are effective at many NLP tasks, they pose multiple problems. RNNs' sequential nature generally precludes them from working in parallel. As a result, training can become very computationally expensive, especially with long sequences of words. Additionally, long-term dependencies tend to get lost, as the model "forgets" what it previously learned (especially in the earliest parts of the sequence) as the gradients either explode or vanish [Hochreiter and Schmidhuber, 1997].

The LSTM [Hochreiter and Schmidhuber, 1997] architecture was developed, at least in part, to remedy the issue of vanishing and exploding gradients. Originally developed in 1997 as a gap-insensitive alternative to traditional RNNs [Hochreiter and Schmidhuber, 1997], LSTMs gained prominence in NLP due in large part to their effective handling of long-term dependencies. Google's Neural Machine Translation system [Wu et al., 2016], for example, consists of a network of LSTMs. LSTM's are a specific type of RNN where each time-step's hidden state is replaced by a "memory cell." Each LSTM unit is comprised of a memory cell and a series of gates:

> *A multiplicative input gate unit is introduced to protect the memory contents stored in* [the unit] *from perturbation by irrelevant inputs. Likewise, a multiplicative output gate unit is introduced which protects other units from perturbation by currently irrelevant memory contents stored in* [the unit]. [Hochreiter and Schmidhuber, 1997]

From these gates, the memory cell is able to discern what information it should remember and what information it can forget, making LSTMs better-capable of maintaining long-term dependencies in sequences than traditional RNNs. There have since been additions and variations to the original LSTM model, such as the addition of a "forget" gate [Gers et al., 2000] (which enables an LSTM to reset its own state), and the creation of the Gated Recurrent Unit (GRU) [Cho et al., 2014].

The bidirectional LSTM (BiLSTM) was by far the most prolific LSTM architecture found in the studies selected for inclusion in this review. The BiLSTM is simply the concatenation of two independent LSTMs. One LSTM reads the input sequence from front to back, while the second LSTM reads the input sequence from back to front. An encoded vector is then formed by the concatenation of both outputs.

ELMo is a language model that was developed by researchers at the Allen Institute of Artificial Intelligence by training a BiLSTM model on a billion-word corpus

[Chelba et al., 2013][Peters et al., 2018]. ELMo is considered the first deep, contextualized, and bidirectional language model. ELMo's bidirectionality allows it to contextualize its vector representations in a manner that largely mitigates the issue of polysemy in NLP. For instance, the word "club" has different vector representations in the phrases "I am a member of the club" and "he was struck by a club." Additionally, ELMo is able to better handle "out-of-vocabulary" (OOV) terms better than previous models, as ELMo's vector representations are character-based as opposed to word-based. This allows the model to represent inscrutable words as a representation of its most frequently occurring character combinations.

ELMo proved it was possible to leverage a pretrained model and fine-tune it toward a downstream task, similar to Computer Vision's ImageNet. When ELMo was published in early 2018, it set a new SOTA for six NLP tasks including SQuAD and SNLI [Peters et al., 2018]. This was a significant breakthrough, as previous SOTAs were mostly achieved via deep learning (as opposed to transfer learning). ELMo was the first model to demonstrate the potential of transfer learning to perform as well as (or outperform) deep learning in NLP.

2.1.2 Encoder-Decoder, Attention, Self-Attention and Transformer

Prior to transformers, the encoder-decoder RNN was the preferred architecture for modeling sequence-to-sequence tasks, especially the task of Machine Translation. The encoder would transform an input sequence from one language into a vector representation. The vector was then fed into the decoder, which would transform it back into an output sequence in a different language [Sutskever et al., 2014]. This was an important innovation in Machine Translation, as sequences rarely align in a one-to-one fashion. Consider the French phrase "il va faire chaud," which means "it is going to be hot" in English. The French sequence has four words, while the English equivalent has six words. A one-to-one alignment is not possible as the input and output are different lengths. The encoder-decoder architecture mitigates this problem by using an RNN to encode the French input into a vector and then uses another RNN to decode the vector and produce an output in English [Sutskever et al., 2014].

Largely adopted for the task of Machine Translation, the *attention* mechanism was an improvement to the original encoder-decoder architecture. The attention mechanism computes a weight vector that specifies how much "attention" should be paid to each of the other embeddings in the input at each time-step of the output. The output then uses these attention weights to determine which words in the input are most important when determining the correct word to generate in the output. Figure 2.1 shows an example of an attention mechanism during a Machine Translation task. One can see at each time-step in the output (English) where the most attention is being paid to in the input (French). Unsurprisingly, the most attention is paid (in this case) to the output's corresponding French equivalent in the input, represented by the diagonal white squares.

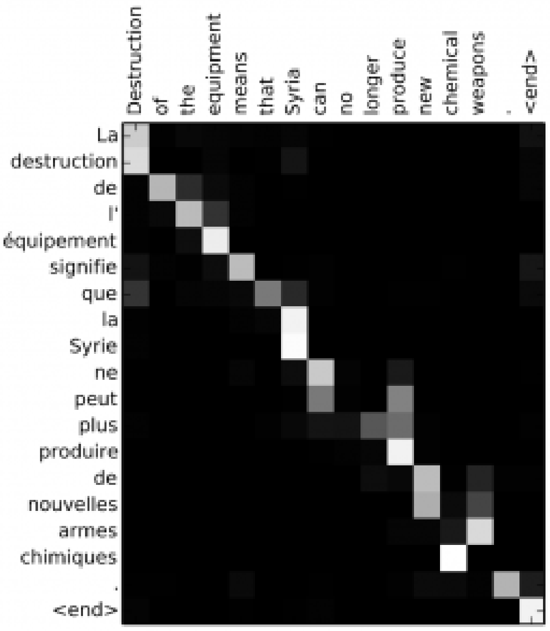
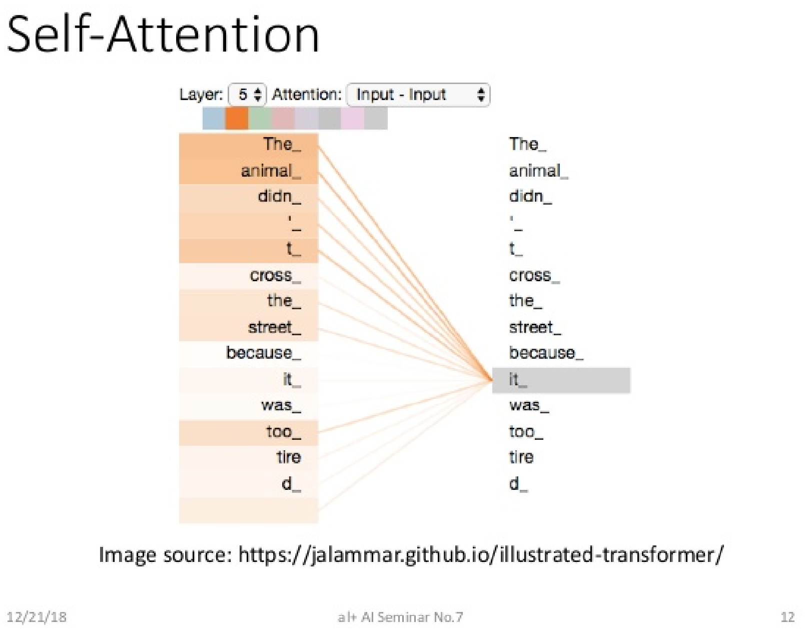


Figure 2.1: *Attention During Example Machine Translation Task (machinelearningmastery.com)*

Unlike the traditional attention mechanism, which communicates attention weights across layers, *self-attention* is only concerned with activations in the same layer in which the attention is being applied. While attention mechanisms are often used to transfer information from an encoder to a decoder, self-attention is only applied within a single layer. Self-attention is good at modeling dependencies between different parts of the same sequence, as opposed to the dependencies between two different sequences.

Transformers utilize self-attention for dependency modeling. Figure 2.2 illustrates a Transformer layer's self-attention mechanism. In this case (that of BERT), one can see that there is no output, only input. This is because self-attention is only concerned with within-sequence attention weights for the purposes of encoding. This example illustrates the words being attended to by the word "it."



Self-Attention

Layer: 5 Attention: Input - Input

Image source: https://jalammar.github.io/illustrated-transformer/

12/21/18 aI+ AI Seminar No.7 12

Figure 2.2: *Self-Attention of a Transformer Layer (jalammar.github.io/illustrated-transformer)*

Transformers eschew recurrence and convolution altogether and rely solely on self-attention to formulate dependencies [Vaswani et al., 2017]. Transformers consist of six encoders and six decoders. Each encoder has two layers: a self-attention layer and a feed-forward neural network layer. Each decoder consists of three layers: a self-attention layer, an attention layer for the encodings, and a feed-forward neural network layer [Vaswani et al., 2017]. The encoder generates encodings that indicate the relevant parts of the input (where the attention should be paid), and the decoder takes these encodings and uses them to generate an output.

The Transformer's ability to facilitate training in parallel by forgoing recurrence in favor of self-attention enabled pretrained models like BERT to be generated from the unsupervised training of massive corpora.

2.1.3 Issues With Previous Approaches

Prior to the advent of the Transformer, previous NLP approaches encountered a wide range of issues. Deep learning was the method of choice for many NLP tasks (usually via RNN); as a result, many NLP tasks were subject to the same difficulties encountered when applying deep learning to non-NLP tasks. Deep learning requires a massive amount of both data and computational power. As such, training deep neural networks is not always practical. In many instances, SOTA GPUs cost thousands of dollars and still take days to train large corpora. Additionally, many datasets are comprised of only a few thousand instances and are therefore less-than-ideal candidates to effectively train deep learning models. RNNs, specifically, are also prone to vanishing and exploding gradients due to their deep and sequential nature [Hochreiter and Schmidhuber, 1997]. LSTMs were developed to combat the exploding and vanishing gradient problem encountered by traditional RNNs and to preserve semantic dependencies across large sequences of text [Hochreiter and Schmidhuber, 1997]. While this was an improvement over the base RNN model, it was far from a panacea. LSTMs are still RNNs, and as such operate sequentially.

This makes them very difficult (if not impossible) to parallelize, which in turn precludes the training of large amounts of data in a reasonable amount of time—something that is absolutely necessary when generating effective transfer learning models in NLP.

Another problem was related to word embeddings. Algorithms like GloVe [Pennington et al., 2014], for example, do not handle OOV words gracefully, nor does GloVe have the ability to distinguish between homonyms: all instances of the word "club" have the same vector representation, regardless of context. Furthermore, many of the previous methods of generating word embeddings do not offer contextualization, as they are based on the frequency of the words' appearances (or co-occurrences) instead of the words' relative positional encodings. Poor word representation in the vector-space precluded transfer learning from being a viable alternative to deep learning until ELMo.

2.2 Transfer Learning

Transfer learning is the process of using knowledge gained from one task and applying it to other related tasks (towardsdatascience.com), allowing one to leverage pretrained models as opposed to training a model from scratch. This provides a salient alternative to deep learning, as transfer learning does not require the massive amounts of data and computation required by deep learning once the pretrained model is fully trained. A fully pretrained transfer learning model can be applied to a smaller dataset that would have otherwise been intractable with deep learning approaches. A transfer learning model can be fine-tuned for a specific task with a minimal amount of additional training and at a fraction of the computational cost of deep learning, thus solving many of the aforementioned issues that plague deep learning.

Transfer learning has had a tremendous impact on the field of Computer Vision in particular. In its original publication, ImageNet was a database consisting of roughly 3.2 million hand-annotated images across (over) five thousand classes [Deng et al., 2009] (although most models are trained on a distilled version consisting of one thousand classes). Today, the ImageNet database boasts a collection of roughly 14.2 million images across approximately 22,000 classes (image-net.org). ImageNet's power lies in its size: any Computer Vision algorithm has the opportunity to train on ImageNet's extensive database before being used to model a downstream task. The weights generated by pretraining on ImageNet are, as a result, able to serve as a platform from which to build off of when modeling the downstream task. This is often done by simply adding an additional layer (of classification, for example) on top of the pretrained model. Transfer learning is currently the status quo for modeling many Computer Vision tasks such as image recognition and image captioning.

The same approach can be taken to model datasets in NLP, as most of the computation power expended in deep NLP is used to gain an understanding of the underlying language. Considering that within-language tasks are all predicated on the same lexicon and semantic relationships (i.e., language), a pretrained English model, for example, should be able to be leveraged and fine-tuned for any task in English. However, until recently, this simply was not possible, as there did not exist a language model whose embeddings were well-represented enough in the vector-space to usurp previous deep learning approaches. It was not until ELMo's deep, bidirectional, and contextual encodings that transfer learning's efficacy was able to match deep learning's in NLP. Since ELMo's publication, other language models such as BERT, ULMFiT, and GPT-3 have been published and shown to outperform many of the deep learning approaches that previously set the SOTA on a variety of NLP tasks. Today, transfer learning is (nearly) as ubiquitous in NLP as it is in Computer Vision.

2.3 BERT

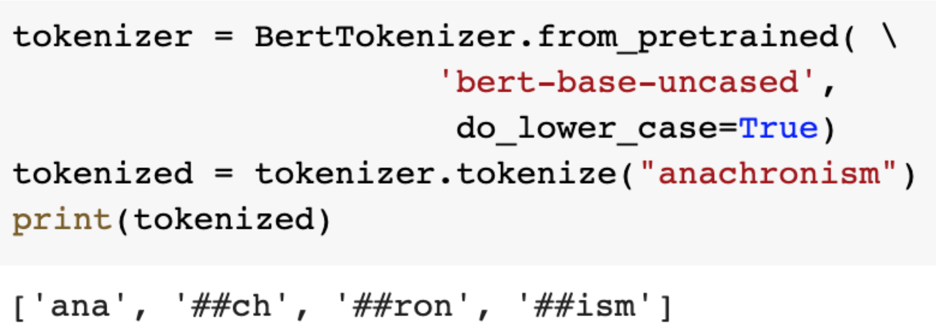
This section takes a deeper dive into the BERT model, examining its architecture and core components. Training, current advances, and issues that currently surround the model are also explored.

2.3.1 WordPiece Embeddings

Adopted from the Byte Pair Encoding (BPE) algorithm [Gage, 1994] originally created for data compression, BERT uses WordPiece tokenization [Schuster and Nakajima, 2012] to generate its 30,000-word vocabulary [Devlin et al., 2018]. These tokens consist of words, subwords, and characters. WordPiece tokenization allows for a much better handling of OOV words, as WordPiece tokenization allows obscure, OOV terms to be identified by the combination of their subword components [Schuster and Nakajima, 2012]. Subwords that are not prefixes are identified with '##' before the token, which can be seen in Figure 2.3. In the figure, the word *anachronism* (something out of place in time), for example, is broken down and represented by BERT as a combination of subwords instead of being labeled as OOV. The figure shows a real-world example of how BERT tokenizes the word "anachronism." Being able to avoid the use of OOV tokens is imperative when modeling texts largely composed of uncommon words and phrases.

2.3.2 Architecture

Previous RNN models were trained in one direction, either front-to-back or vice versa. BERT was born out of the desire for bidirectional contextualization in pretrained language models. In a sentence, words generate their contextual significance from the words that



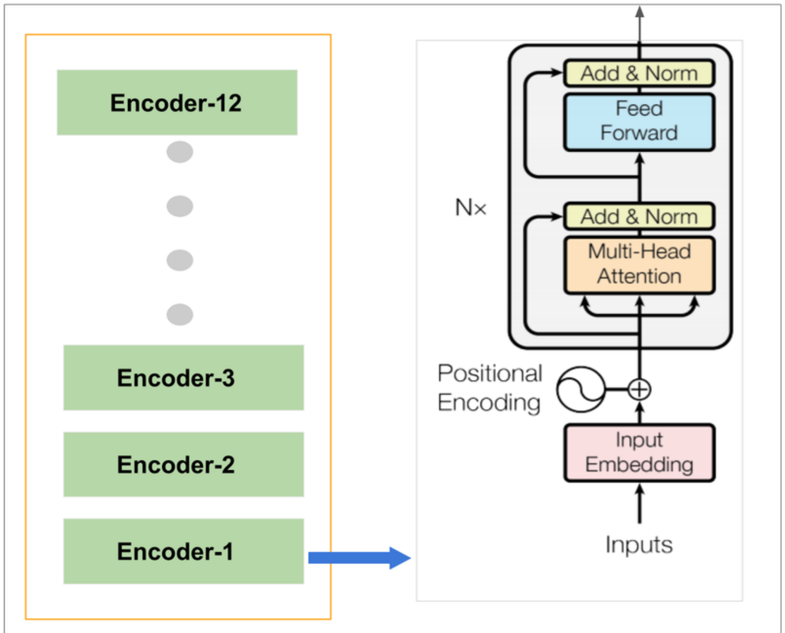
```
tokenizer = BertTokenizer.from_pretrained( \
                        'bert-base-uncased',
                         do_lower_case=True)
tokenized = tokenizer.tokenize("anachronism")
print(tokenized)
```

```
['ana', '##ch', '##ron', '##ism']
```

Figure 2.3: *BERT's Tokenized Representation of the Word "anachronism"*

both precede and succeed them. As such, unidirectional models are prone to inaccurate contextualizations—especially with polysemic words. Consider the phrase "Ernie went down to the bank..." If that phrase ends with "to go fishing," the word "bank" will have a different meaning (and therefore different embedding) than if the phrase ends with "to make a withdrawal." A unidirectional model would be unable to make this distinction (reading the text from front to back). Therefore, it is important that language models are able to contextualize bidirectionally. While ELMo achieved pseudo-bidirectionality by training in each direction and then concatenating the results, BERT was the first model to achieve true birdirectionality via the inclusion of Transformers in its architecture [Devlin et al., 2018].

BERT was released with two versions: BERT-base and BERT-large, and each has a cased and uncased iteration (there is also a Chinese BERT for Chinese and a Multilingual BERT that was originally trained on 102 different languages). BERT-base has twelve layers (Transformer blocks), each with twelve self-attention heads and 768 hidden neurons. BERT-base consists of approximately 110 million parameters: approximately 24 million from the embeddings, 85 million from the transformers, and one million from the pooler (BERT-large, comparatively, has roughly 340 million parameters) [Devlin et al., 2018]. Although not specifically stated, it is inferred that [Devlin et al., 2018] opted for the above parameters based on an ideal trade-off between model performance and memory requirements. BERT-large is over three times as large, but it has not been shown to outperform BERT-base by an equally significant margin. BERT-base's architecture can be seen in Figure 2.4, where twelve encoders are stacked sequentially. Each encoder is a transformer with its own attention heads. It is also important to note that only the encoder portion of the transformer (shown) is included in BERT's architecture, as BERT is not a generative model and does not implement a decoder. Figure 2.5 from the original BERT paper illustrates BERT's architecture compared to its peers GPT and ELMo. One can see by the arrows that BERT is the only truly bidirectional architecture. GPT is unidirectional (front-to-back), and ELMo achieves pseudo-bidirectionality by concatenating two separate, unidirectional LSTMs.



Source: https://arxiv.org/pdf/1706.03762.pdf

Figure 2.4: *BERT's Architecture.*
[Vaswani et al., 2017]

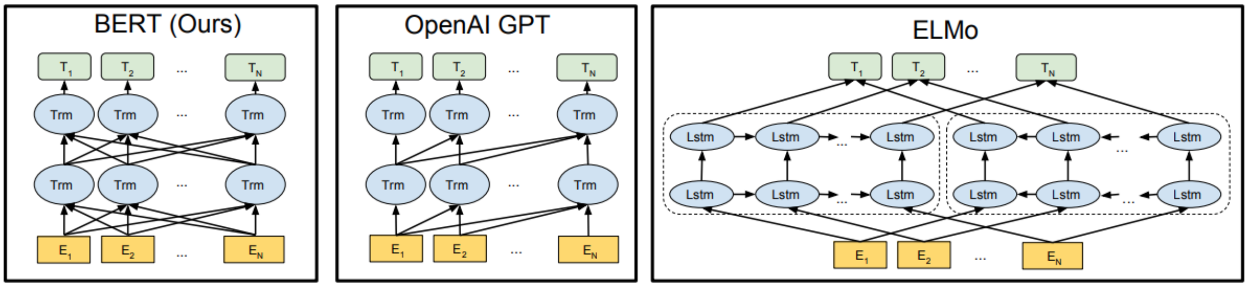


Figure 2.5: *Comparison of BERT's Architecture to GPT and ELMo* *[Devlin et al., 2018]*

2.3.3 Training

BERT was originally trained via multitask, unsupervised learning on Wikipedia and BookCorpus. Wikipedia consists of over 2.5 billion words [Devlin et al., 2018], and BookCorpus consists of nearly one billion words [Zhu et al., 2015]. [Devlin et al., 2018] employed two different methods during pretraining: masked language modeling (MLM) and next sentence prediction (NSP). MLM is the process of randomly omitting a word (or words) from a sentence and training the model to predict the missing words based on context. NSP jointly trains text-pairs by inputting two sentences and then training the model to discern whether or not the second sentence is a valid continuation of the first [Devlin et al., 2018]. MLM was done with a masking rate of 15%, and NSP was done with the correct sentence being present 50% of the time. The model was trained with a batch size of 256 sequences * 512 tokens per sequence. Training was done over one million steps, which equates to roughly forty epochs over 3.5 billion words. Training took four days to complete on four cloud TPUs (sixteen TPU chips) [Devlin et al., 2018].

The pretraining diagram from the original BERT paper can be seen in Figure 2.6. During pretraining, each training instance is comprised of a pair of sentences. During the MLM task, 15% of the words in each sentence are masked with either a **"[MASK]"** token (80% of the time) or a random token (10% of the time). For the remaining 10%, the token selected for masking remains unchanged. The addition of the random and unchanged tokens is done to mitigate the fact that the **"[MASK]"** token does not actually appear during the fine-tuning process [Devlin et al., 2018]. After the tokens are masked, the model then tries to discern the missing words. During the NSP task, the model looks at both sentences and decides whether or not the second sentence is a valid continuation of the first. The **"[CLS]"** token indicates the start of the input sequence, and **"[SEP]"** indicates the separation of the first and second sentences.

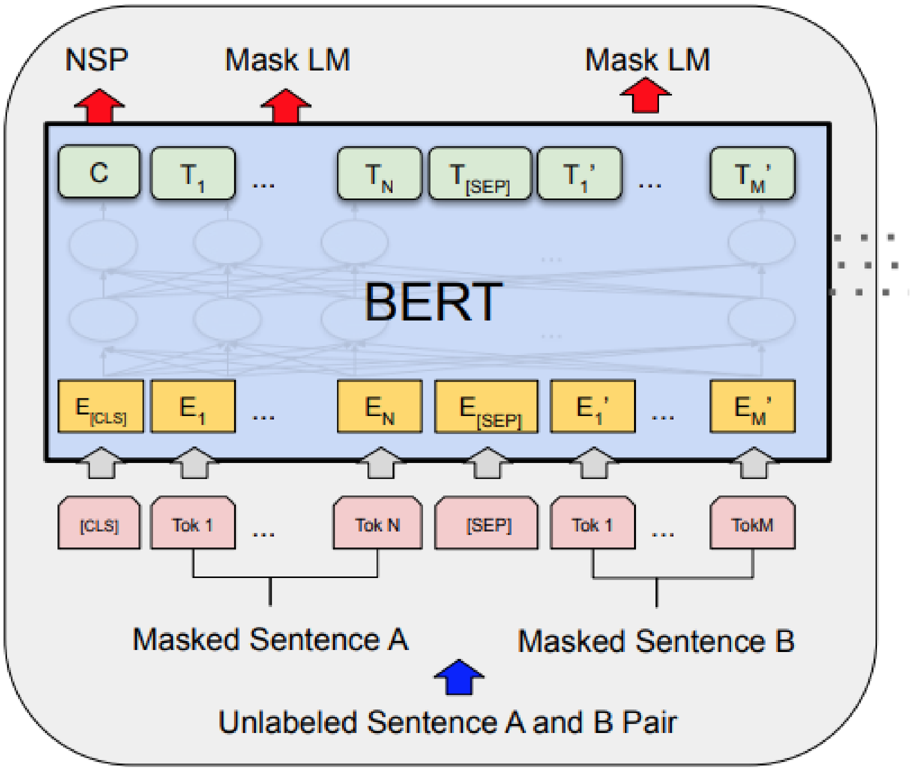


Figure 2.6: *Pretraining BERT [Devlin et al., 2018]*

2.3.4 Advances

BERT was designed as a pretrained model that could be fine-tuned for any Natural Language Understanding (NLU) task with a minimal amount of additional training.

When BERT was released in October 2018, it obtained a new SOTA on eleven separate NLP benchmarks including GLUE [Wang et al., 2018], MultiNLI [Williams et al., 2018], SQuAD 1.1, and SQuAD 2.2 [Rajpurkar et al., 2016]. This was significant, as it was the first time that a task-agnostic pretrained model was able to outperform its deep-learning peers on such a wide array of tasks. Furthermore, it led further credence to the idea that transfer learning models could supplant deep learning ones. As a result, transfer learning has emerged as the method of choice for modeling several different NLP tasks. The fine-tuning diagram from the original BERT paper is depicted in Figure 2.7. [Devlin et al., 2018] recommend fine-tuning BERT for two, three, or four epochs; in batch sizes of either sixteen or thirty-two; and with a learning rate of either 5e-5, 3e-5, or 2e-5 [Devlin et al., 2018]. The figure demonstrates that the same BERT model is being used to model three different tasks. In the SQuAD example, the BERT model takes two sentences as the input: a question and a block of Wikipedia text. The model searches the Wikipedia passage to find the correct sequence of words that answers the posed question. If the passage cannot answer the question, the model labels the instance as unanswerable.

Fine-tuning BERT has become a popular approach for modeling many downstream tasks such as Named Entity Recognition (NER), question answering, sentiment analysis, relation extraction, intent classification, coreference resolution, and many more. Google has even started incorporating BERT in its search engine (blog.google.com). Many benchmarks continue to experience new SOTA results with BERT and transfer learning.

Additionally, BERT has since evolved into further pretrained, domain-specific models that are designed for application toward specific subject matter. BioBERT [Lee et al., 2019] and SciBERT [Beltagy et al., 2019], for example, were further pretrained with clinical and medical corpora to give BERT a better understanding of scientific and medical terminology (SciBERT also released an additional domain-specific model trained from scratch [Beltagy et al., 2019]). Many of the words that pervade scientific and medical corpora are

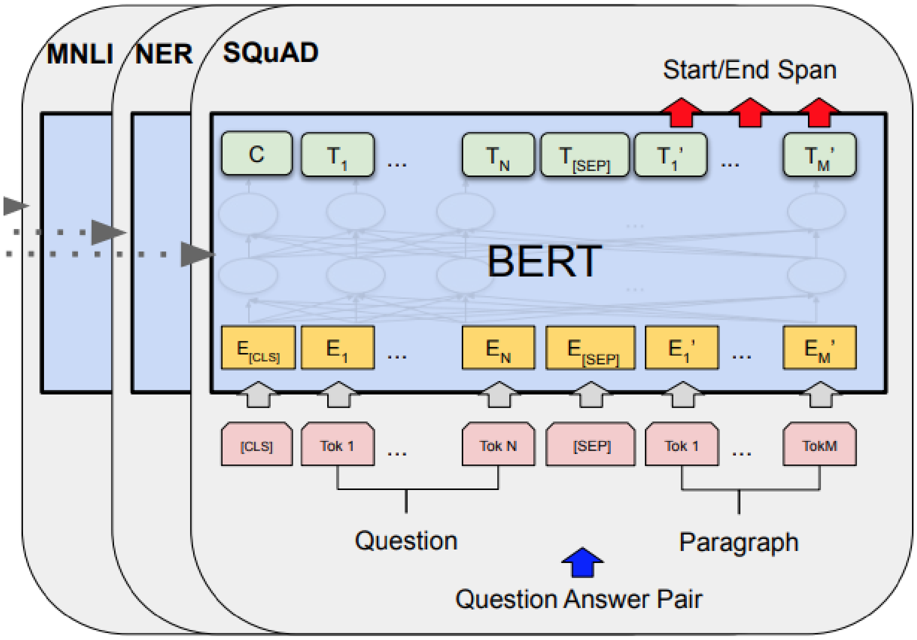


Figure 2.7: *Fine-Tuning BERT [Devlin et al., 2018]*

technical terms and not likely to be well-represented in canonical texts such as Wikipedia or BookCorpus. This additional pretraining allows BERT to improve the contextual vectorization of terms that it was previously unfamiliar with.

BioBERT was further pretrained with PubMed (a corpus of thirty million citations and abstracts of biomedical literature) (pubmed.ncbi.nlm.nih.gov) and PMC (a corpus of 6.5 million full-text biomedical and life sciences journal articles) (ncbi.nlm.nih.gov), while SciBERT was further pretrained with a random sample of 1.14 million papers from Semantic Scholar (semanticscholar.org). While BioBERT chose to utilize BERT's lexicon, SciBERT opted to create its own vocabulary (called SciVocab) based on the most frequently occurring words and subwords in scientific research papers. The overlap between BERT's vocabulary and SciVocab is only about 40%, illustrating the drastic difference in the frequency of words in scientific versus canonical corpora [Beltagy et al., 2019]. As a result, scientific terms need to be tokenized by subword components more often in BioBERT than in SciBERT. WordPiece tokenization, the process by which BERT tokenizes words that are OOV, is illustrated in Figure 2.3. The performance of these domain-specific BERT derivatives versus standard BERT is one of the concerns of this thesis and is subsequently addressed in RQ3.

2.3.5 Issues With BERT

While BERT is well-suited for many NLP tasks, it is not without its limitations. BERT is a series of encoders and is not a generative model. As such, it cannot be used as-is for Natural Language Generation. Furthermore, BERT's input is limited to sequences of 512 tokens (or fewer) comprising one or two sentences [Devlin et al., 2018], making long-term relationships between words difficult to identify. This was by design, as the self-attention mechanism in the Transformer architecture is specifically designed to only consider in-sequence embeddings when calculating attention weights; still, it is worth noting that it can serve as a limitation when using BERT to discern long-distance relationships across multiple sentences or pages.

Memory can also be an issue. Even with the 512-token input limit, I regularly encountered out-of-memory (OOM) issues on the dedicated Nvidia Tesla K80 GPU provided by Colab when using batch sizes higher than six with BERT-base. It was not until the input maximum was reduced to 128 tokens that the OOM issues completely subsided and batch sizes of sixteen or thirty-two were able to be used without concern for memory. Lastly, BERT is still subject to many of the same issues that torment other deep learning approaches. Sparsely populated and imbalanced datasets, in particular, can be difficult to apply BERT to in a meaningful and effective manner.

2.4 Other Literature Reviews

To the best of my knowledge, there are currently no SLRs specific to BERT implementations on domain-specific datasets. The queries “allintitle: bert review” and “allintitle: bert survey” were posed to Google Scholar. The searches yielded a total of twelve results published in or after 2018 (the year BERT was published), five of which were specific to BERT. No BERT-related results were SLRs. An additional Google (full search engine, not just Scholar) search was conducted to see if there were any outlying reviews. Multiple reviews containing BERT were found; however, nearly all these reviews concerned BERT’s performance relative to its peers (primarily ELMo and GPT-2). Like Scholar, Google did not yield any SLRs addressing domain-specific adaptation with BERT.

Although transfer learning is pervasive in NLP, BERT is a relatively new technology and was only made public in 2018. While BERT has enjoyed a wide adaptation across many NLP tasks, little research has been done on BERT’s efficacy on specific domains and whether the base model can be improved upon for application toward specific subject matter. Specific domains often contain obscure words and phrases that are not prevalent in Wikipedia or BookCorpus, and as such are not likely to be well-represented in BERT’s vector-space. BERT has been shown to generalize very well across task and subject matter [Devlin et al., 2018], but further research needs to be done to evaluate BERT’s performance on domain-specific data. The purpose of this literature review is to evaluate the most effective practices of modeling domain-specific scientific and medical datasets with BERT.

CHAPTER 3

METHOD

The SLR protocol for this review is primarily based on the guidelines set forth by [Kitchenham, 2004]. Other sources such as [Heckman and Williams, 2011], [Otter et al., 2019], [Kitchenham et al., 2009], [Hughes, 2019], and [VanLehn, 2011] were also utilized. This section presents the SLR protocol, which includes hypotheses, search strategy, study selection, and data synthesis.

3.1 Hypotheses

Each research question (except RQ3) is accompanied by a corresponding hypothesis and justification. The hypotheses are listed below:

- Hypothesis (RQ1): BERT outperforms other approaches when applied to scientific and medical datasets.

One of the contributing factors to BERT's robustness is its use of WordPiece embeddings in the place of OOV tokens. While WordPiece embeddings may not be as effective as having the words themselves present in the underlying model's vocabulary, they are nevertheless unique representations of otherwise unidentifiable words. Therefore, it is believed that BERT should be able to outperform other approaches when applied to scientific and medical datasets just as it has outperformed other models on general-domain datasets.

- Hypothesis (RQ2): Domain-specific pretraining and modeling of BERT outperforms standard BERT when applied to scientific and medical datasets.

BERT owes its success to the contextualized word representations in its vector-space. Considering that many of the terms in scientific and medical datasets are rare in the canonical texts on which BERT was trained, it stands to reason that BERT's further pretraining on domain-specific data would help improve the vector representations of words that were largely underrepresented during BERT's initial training. As such, it is predicted that exposing BERT to scientific and medical corpora before the fine-tuning process will augment the model's performance on these same types of datasets. WordPiece embeddings allow otherwise-OOV words to have unique representations in BERT's vector-space. As such, these embeddings are trainable. While the embedding representations of OOV terms consist of multiple tokens (subwords), as opposed to a single token, the fact that these rare words can be vectorized by BERT indicates that they are trainable and thus able to be better contextualized through additional pretraining. It is possible that further pretraining is more effective on words that are present in BERT's vocabulary (as a single token), but it is predicted that further pretaining the subword representations of domain-specific words not present in BERT's lexicon will nevertheless provide better contextualization than the original embeddings that were generated via pretraining on canonical corpora.

- Hypothesis (RQ4): BERT is not the best-performing approach for modeling scientific and medical datasets in languages other than English.

BERT's contextualization is word-based (as opposed to character-based, like ELMo [Peters et al., 2018]). Unlike English BERT, Chinese BERT is character-based. Because of this, there is skepticism regarding its potential for contextualizing words as well as English BERT. For all datasets that are not in English or Chinese, multilingual BERT is used. Unfortunately, multilingual BERT was trained on so many languages that it is doubtful that it can be particularly effective on any single language. For these reasons, it is surmised that BERT will not be the preferred approach for modeling datasets in languages other than English.

Answering the research questions and evaluating these hypotheses will go a long way in determining how robust BERT is to texts whose vocabularies are not well-represented in BERT's lexicon. Furthermore, BERT's performance will be assessed across multiple tasks, datasets, and languages. This will provide insight into the most effective methods of applying BERT to scientific and medical datasets. RQ3 did not have an accompanying hypothesis, as it is merely a survey of prevalent issues affecting researchers when applying BERT to scientific and medical datasets.

3.2 Search Strategy

This section illustrates the process by which papers were gathered for this review. It includes the terms and databases used for the search. The search window was from BERT's original publication date in October 2018 through the end of May 2020.

3.2.1 Search Terms

Because BERT is a novel architecture whose published body of work is limited in size, all BERT-related research was initially considered during the search. Thus, the only term searched for across all databases was "bert." All papers containing "bert" in the title that were present in one of the databases listed in Table A.1 in Appendix A were considered for inclusion in this review. Each paper was initially inspected to ensure the following inclusion criteria were met:

- Paper must be relevant to BERT (as opposed to an individual named Bert, Bert and

Ernie from Sesame Street, the plant Stevia rebaudiana Bert., etc.).

- Paper must be available in English.
- Paper must not be a press release.

Accounting for these criteria, the initial search yielded 550 total papers, 425 of which were unique. Duplicates were merged based on the latest version number. If the version numbers were the same (or unavailable), papers were merged based on the most recent publication date.

3.2.2 Databases

Table A.1 in Appendix A lists the databases that were searched for BERT-related studies. The number of studies initially retrieved from each (that also adhered to the above criteria) is listed next to each database. The list is arranged by the total number of papers found in each database.

This list of databases was compiled by examining the most prevalent databases in the realm of computer science. Additionally, databases were added based on recommendations from the Thesis Advisor, Professor Peter Hastings. Relevant databases outside of computer science (i.e., computational linguistics) were also included, as there is a considerable overlap in NLP between computer science and computational linguistics.

3.3 Study Selection

This section describes the process taken for selecting studies for the review and also explains the manner in which data was synthesized.

3.3.1 Study Selection Process

The process for selecting studies for inclusion in this review was a three-step process that involved accepting papers based on evaluating each study's title, abstract, and finally, the entire paper. At each step, papers were discarded for not meeting the criteria specific to that stage of quality control. 425 unique BERT-related studies were initially considered for inclusion based on the criteria listed in Section 3.2.1. Forty-nine papers remained after the first round of quality control based on title. Twenty-nine papers remained after the second round of quality control based on abstract. Twenty-seven papers remained after the third round of quality control based on the reading of the entire paper and were subsequently included in this review. The number of studies evaluated at each quality control stage is listed in Table 3.1.

Table 3.1: *Number of Papers Accepted and Rejected at Each Stage of Quality Control*

Quality Control Stage	**Accepted**	**Rejected**
Initial Search	425	
Quality Control 1 (Title)	49	376
Quality Control 2 (Abstract)	29	20
Quality Control 3 (Full Paper)	27	2
Total	**27**	

The specific criteria used for evaluation at each stage of quality control are listed in the subsections that follow.

3.3.1.1 Quality Control 1 (Title)

In the first round of quality control, each paper was evaluated based on its title. The criteria used for inclusion in the next stage of quality control, Quality Control 2 (Abstract), were:

- Paper must be relevant to applying BERT to scientific or medical corpora.
- Paper must be freely available through DePaul's library of databases.

376 papers were rejected in Quality Control 1. Quality Control 1 winnowed the field considerably by requiring each BERT paper selected be specific to applying BERT to scientific or medical datasets. Furthermore, some qualifying papers were excluded based on their inaccessibility. This was mostly due to certain papers being only available on smaller, foreign servers that were not accessible through DePaul's library of databases.

3.3.1.2 Quality Control 2 (Abstract)

In the second round of quality control, each paper was evaluated based on its abstract. The criteria used for inclusion in the next quality control, Quality Control 3 (Full paper), were:

- Paper must be in the realm of NLP and not multimodal (e.g. no image captioning).

Additionally, the paper must meet at least one of the following criteria:

- Paper is published in a peer-reviewed journal or conference, workshop, symposium, or congress proceedings specified in the list of databases in Table A.1 in Appendix A (not including arXiv or Google Scholar).
- Paper has not yet been published, but has been accepted for publication to a journal, conference, workshop, symposium, or congress.

- Paper has not been published or accepted for publication but is widely referenced across the industry (having at least 20 citations on Google Scholar, at least one of which is from a peer-reviewed source).

Twenty papers were rejected in Quality Control 2. Certain papers were multimodal and were discarded, as they were outside the scope of this work. Additionally, because BERT is a new technology and its body of work is limited, open access databases (such as arXiv) were included for consideration. It is also becoming more and more commonplace for computer science researchers to post their work to open access sites, so eschewing them in their entirety in favor of a purely peer-reviewed body of papers would have omitted some key works (the original BERT paper, for example, was published to arXiv and is not located in any peer-reviewed journal). Unfortunately, many of these types of papers are not peer-reviewed and can therefore be apocryphal. For that reason, the criteria above were implemented to ensure that all relevant, but only quality, papers were considered for inclusion in the review.

3.3.1.3 Quality Control 3 (Full Paper)

In the third round of quality control, each paper was evaluated based on the paper in its entirety. All papers that passed Quality Control 3 were included in the review. The criteria used for inclusion were:

- Paper must make relative performance to other teams available if the paper accompanies a competition.
- Paper must use a variation of the original BERT model.

Two papers were rejected in Quality Control 3. One was rejected due to the unavailability of the competition's results. Without those results, it is not possible to determine the relative efficacy of the paper's proposed BERT implementation. This is important, as this thesis seeks to identify the most effective methods of applying BERT to scientific and medical datasets, and a model's efficacy cannot be evaluated if it cannot be compared to the other models that it competed against. The other paper was rejected because it reduced BERT's dimensionality before implementing it. This review is only concerned with BERT variations that stem from the original BERT-base or BERT-large models.

3.3.2 Data Synthesis

Following Quality Control 3, papers were reread to gather pertinent data. For each paper, the following data was extracted and logged:

- Title
- Author(s)
- URL
- Publication date
- BERT models
- Dataset language
- Publication/Proceedings
- Tasks
- Model structure
- Dataset type
- Approaches taken
- Performance metric
- Additional notes

Reasons for inclusion or exclusion at each quality control stage were also tracked. Additionally, a separate spreadsheet was maintained as a visual representation of papers where each sheet in the workbook compared all papers across a single attribute. The sheets in this spreadsheet were: tasks, languages, BERT models, and publication type. After the data was collected, studies were reread to extract information relevant to answering each research question. These notes were maintained as well.

CHAPTER 4

QUALITATIVE RESULTS

Chapter 4 examines the composition of the body of works that comprise this review. Studies are aggregated by publication type, model type, task, and dataset language. This is done in order to provide insight into the different types of tasks that BERT is being applied to (and with what models) in the domains of science and medicine. The next chapter, Chapter 5, contains the performance metric comparisons of the studies' various models (both BERT and non-BERT). Chapter 4 is intended to answer the question, *"what types of studies are found in this review?,"* while Chapter 5 seeks to answer the research questions and evaluate the hypotheses.

4.1 Publications

Of the twenty-seven studies selected for this review, twelve were selected from workshops, ten were selected from conferences, three were selected from journals, one was selected from a symposium, and one was selected from a congress. Table B.1 in Appendix B shows the publication sources for the studies and the number of studies included from each source.

4.2 Models

Of the twenty-seven studies selected for this review, twelve used English BERT (either BERT-base or BERT-large), seven used Chinese BERT, and five used multilingual BERT. Twenty-two studies experimented with standard BERT, ten studies experimented with a BERT model pretrained on custom corpora, eleven studies tried applying BioBERT, one study tried applying SciBERT, and sixteen studies experimented with at least one non-BERT model. Some studies used multiple BERT models. In total, sixteen different studies experimented with some sort of further pretraining, either via a precompiled model such as BioBERT or SciBERT, or via additional pretraining on custom corpora. The breakdown of studies by the type of model applied (across all languages) can be see in Figure 4.1 (the three right-most bars in Figure 4.1 are the further pretrained models). Table 4.1 shows the various models used for each study.

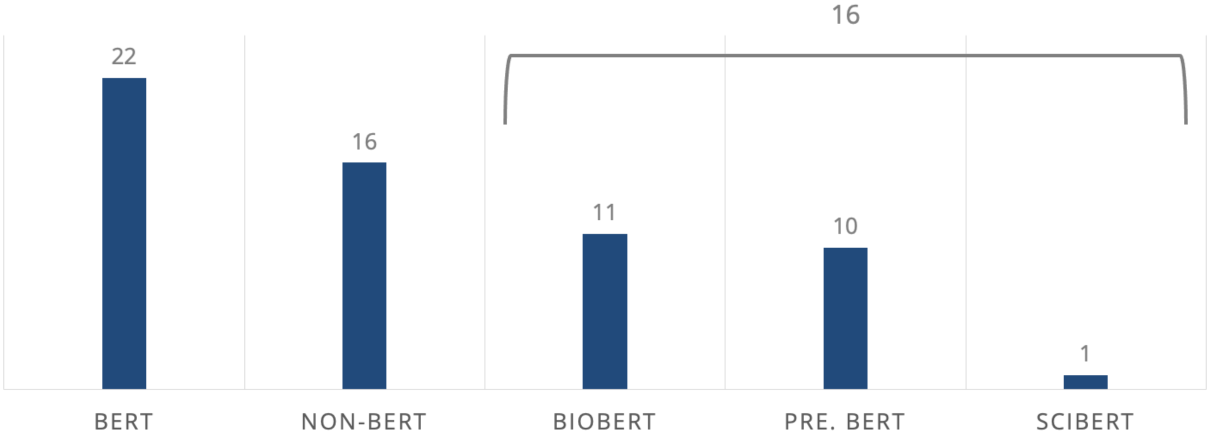


Figure 4.1: *BERT Models Applied by Study*

4.3 Tasks

Ten different tasks were modeled across the twenty-seven studies selected for this review. Of those studies, thirteen performed NER, nine performed classification, six performed relation extraction, three performed normalization, three performed text inference, two performed coreference detection, two performed sequence labeling, two performed anonymization, one performed question answering, and one performed ellipsis detection. Some models performed multiple tasks. Figure 4.2 shows the number of studies that performed a particular task. Table 4.2 shows the tasks performed by each study. Additionally, the various tasks that were modeled are enumerated in the following sub-

Table 4.1: *Models Used*

Study	**BERT**	**Pre. BERT**	**BioBERT**	**SciBERT**	**Non BERT**
[Akhtyamova, 2020]	X	X			X
[Alsentzer et al., 2019]	X		X		
[Dai et al., 2019]	X				X
[Ding et al., 2019]			X		
[García-Pablos et al., 2020]	X				X
[Hakala and Pyysalo, 2019]	X				
[Lee et al., 2019]	X		X		
[Li et al., 2019a]	X		X		X
[Li et al., 2019b]	X		X		
[Li et al., 2020]	X	X			X
[Lin et al., 2019a]	X				X
[Lin et al., 2019b]	X	X	X		
[Liu et al., 2019]	X	X			
[Miftahutdinov et al., 2019]	X		X		X
[Peng et al., 2020]		X	X		
[Peng et al., 2019]		X	X		
[Phongwattana and Chan, 2019]			X		X
[Sänger et al., 2019]	X	X			X
[Song et al., 2019]	X				X
[Sun and Yang, 2019]	X		X		
[Sung et al., 2019]	X	X			
[Trieu et al., 2019]	X				X
[Wang et al., 2019]		X			X
[Xue et al., 2019]	X				X
[Yu et al., 2019]	X	X		X	X
[Zhang et al., 2019a]	X				X
[Zhang et al., 2019b]	X				X
Total	22	10	11	1	16

Pre. BERT = Experimented with a pretrained BERT model

Non BERT = Experimented with a non-BERT model

sections. The results for the performances of the various models across the different tasks and languages is located in Table 5.1 in section 5.

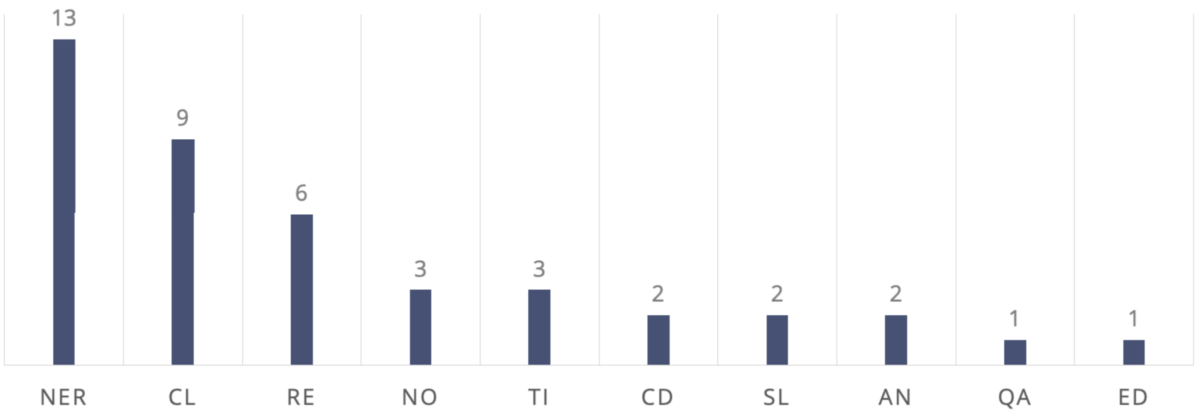


Figure 4.2: *Number of Studies That Performed Each Task*

NER = Named Entity Recognition RE = Relation Extraction NO = Normalization TI = Text Inference CL = Classification
CD = Coreference Detection SL = Sequence Labeling AN = Anonymization QA = Question Answering ED = Ellipsis Detection

4.3.1 Named Entity Recognition

NER is the process of identifying key information in a body of text, i.e., named entities. It is used to identify the important elements in a text and can also help sort unstructured data by identifying important terms, phrases, and entities (en.wikipedia.org). An example of NER is identifying all of the different characters (persons, not letters of the alphabet) present in a Harry Potter book. NER is particularly difficult with scientific and medical datasets, as it is often the esoteric, domain-specific jargon that needs to be extracted. Named entities such as medical conditions, chemical compounds, and clinical symptoms are not likely to be well-represented in canonical corpora. As such, NER is a commonly pursued task in the realms of science and medicine.

4.3.2 Classification

Classification is the process of labeling an instance as belonging to a particular group. An example of a classification task is determining whether a lump present in a mammogram is benign or malignant. Classification is particularly important in NLP, as many

Table 4.2: *Tasks Performed*

Paper	NER	RE	No	TI	Cl	CD	SL	An	QA	ED
[Akhtyamova, 2020]	X									
[Alsentzer et al., 2019]	X			X				X		
[Dai et al., 2019]	X									
[Ding et al., 2019]	X	X	X							
[García-Pablos et al., 2020]							X	X		
[Hakala and Pyysalo, 2019]	X									
[Lee et al., 2019]				X						
[Li et al., 2019a]	X	X								
[Li et al., 2019b]			X							
[Li et al., 2020]	X									
[Lin et al., 2019a]						X				X
[Lin et al., 2019b]		X								
[Liu et al., 2019]					X					
[Miftahutdinov et al., 2019]			X		X		X			
[Peng et al., 2020]	X	X		X						
[Peng et al., 2019]	X	X							X	
[Phongwattana and Chan, 2019]	X									
[Sänger et al., 2019]					X					
[Song et al., 2019]					X					
[Sun and Yang, 2019]	X				X					
[Sung et al., 2019]					X					
[Trieu et al., 2019]						X				
[Wang et al., 2019]					X					
[Xue et al., 2019]	X	X								
[Yu et al., 2019]					X					
[Zhang et al., 2019a]					X					
[Zhang et al., 2019b]	X									
Total	13	6	3	3	9	2	2	2	1	1

NER = Named Entity Recognition RE = Relation Extraction No = Normalization TI = Text Inference Cl = Classification
CD = Coreference Detection SL = Sequence Labeling An = Anonymization QA = Question Answering ED = Ellipsis Detection

NLP tasks involve extracting semantic meaning from sequences of words. Tasks such as sentiment analysis, intent classification, and topic labeling are all types of classification that are used in NLP to discern meaning from text. In this review, classification tasks are grouped together for analytical purposes, as each type of classification task is concerned with predicting a label from an instance.

4.3.3 Relation Extraction

Relation Extraction is the process of detecting semantic relationships between different entities in a body of text. It can be used to identify correlative pairs of entities, such as the relationships between various diseases and the gene mutations that cause them. Relation extraction can also be used educationally to discern the degree to which a student's essay demonstrates his or her understanding of the course material. [Cochran et al., 2020] addressed this in their work with detecting causal relations in student short-answer essays. Relation extraction is a key component of NLP, as it involves detecting how different entities are related to each other; these relationships can often escape human detection.

4.3.4 Additional Tasks

Seven additional tasks were addressed in the studies that comprise this review, but each of these tasks was modeled by three or fewer studies. As such, these seven tasks were combined to form this section.

Normalization. Text normalization is the process of transforming text into a specific, consistent format (en.wikipedia.org). Normalization is often done in order to aid another downstream task such as classification. An example of normalization is the handling of different date formats: 1/1/1970 may be classified as a different date than 1-1-1970 if the two dates are not first normalized. This is important in the domains of science and medicine in NLP because many scientific terms often correspond to terms in common usage. The chemical compound NaCl, for example, is also referred to as ”sodium chloride” and is referred to colloquially as ”salt.” Without normalization, a model may overlook the fact that all three of these terms refer to the same entity.

Text Inference. Text inference, or Natural Language Inference (NLI), is the task of determining whether a given hypothesis is true (entailment), false (contradiction), or undetermined (neutral) based on a given premise (nlpprogress.com). An example would be inferring whether or not the hypothesis ”the Doors are a rock band” can be gleaned from the premise ”Jim Morrison was the lead singer of The Doors” (in this case the answer is no, and the result is ”neutral”). NLI is important in NLP because it is used to determine whether or not a piece of text (the premise) is relevant to answering a hypothesis. Tasks like question answering can use NLI to decide which parts of a passage are important in determining an answer to the posed question.

Coreference Detection. Coreference detection is the task of identifying all expressions in a text that refer to the same entity (en.wikipedia.org). This can be incredibly difficult, especially if there is ambiguity present in the text. Consider the phrase ”the dog liked the rain, and John hated it.” It is not clear whether John’s hatred is directed at the dog or the rain. Coreference detection aims to resolve this conundrum by identifying all references to a particular entity, which enables one to recognize parts of text that could be problematic due to ambiguity.

Sequence Labeling. Sequence labeling assigns a label to each member of a given sequence (en.wikipedia.org). An example of this is part-of-speech (POS) tagging, where each word in a sequence is labeled with its corresponding part of speech (noun, verb, adjective, etc.). Sequence labeling is important in NLP because the information learned from the sequence labels can then be used in conjunction with other models (such as language models) to augment performance.

Anonymization (Deidentification). Deidentification is a subset of NER that recognizes sensitive information that needs to be redacted. An example of anonymization is training a model to automatically redact patient names from hospital records. This is of particular concern in medicine, as much of the information in medical records is sensitive and governed by stringent privacy laws. Anonymization can also be used in other capacities, such as deidentifying a job applicant's gender and race to prevent bias during the hiring process.

Question Answering. Question answering is the process of building a model capable of answering questions by querying a base of knowledge (en.wikipedia.org). Examples of question answering models are Apple's Siri, Google's Google Assistant, and Amazon's Alexa. Each of these models queries a database (or the Internet) to answer questions posed to them by humans. Question answering is also heavily utilized in customer service applications as well, as an increasing number of companies are opting for automated assistance in lieu of in-person operators.

Ellipsis Detection. Ellipsis detection is the process of detecting an omission from a clause that is nevertheless understood in the context of the elements that are present. For instance, in the clause, ”John can play the guitar, Mary the violin,” the phrase ”can play” is understood to be in between the words ”Mary” and ”the violin” (en.wikipedia.org). Ellipsis detection, in this example, seeks to identify the missing phrase ”can play.” Ellipsis detection is important, as it seeks to identify specific sequences of words that language models may overlook due to a particular idea not having been explicitly stated. In the previous example, a language model may interpret the phrase ”Mary the violin” as a reference to a violin whose name is Mary, as opposed to a reference to a person named Mary who plays the instrument the violin. Ellipsis detection aims to resolves these types of conflicts.

4.4 Languages

Of the twenty-seven studies selected for this review, fifteen used English datasets, seven used Chinese datasets, four used Spanish datasets, and one used a German dataset. The papers whose datasets were in Spanish and German used multilingual BERT. Figure 4.3 shows the number of studies that modeled each dataset language. All five studies that modeled Spanish and German datasets applied the multilingual version BERT. The breakdown of studies by dataset language is illustrated in Table 4.3.

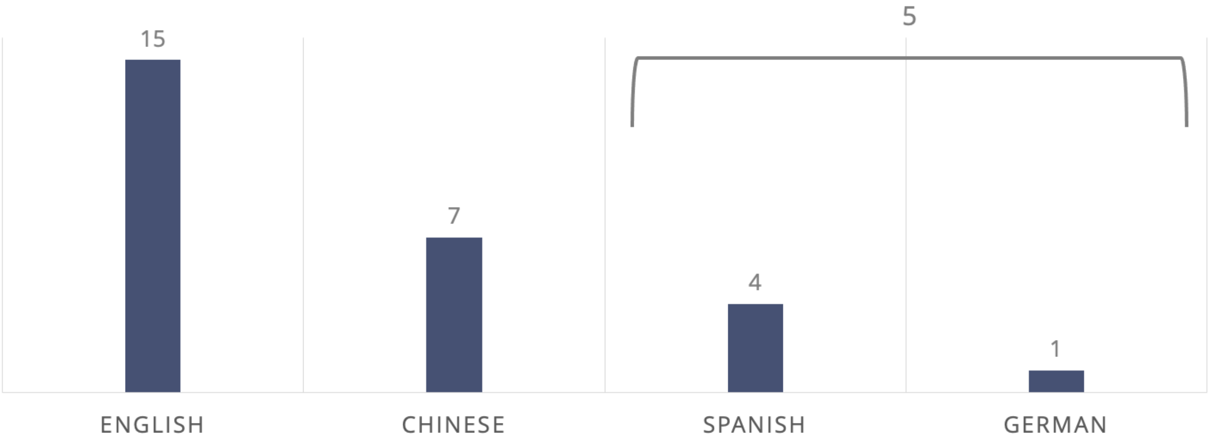


Figure 4.3: *Number of Studies That Modeled Each Dataset Language*

Table 4.3: *Dataset Languages by Study*

Language	**Quantity**
English	15
Chinese	7
Spanish	4
German	1

CHAPTER 5

QUANTITATIVE RESULTS

Unlike Chapter 4, which sought to examine the composition of the studies included in this review, Chapter 5 is provided specifically to answer the research questions and evaluate the hypotheses. This is done by comparing the performance metrics for each study across a variety of models, tasks, datasets, and languages. For each study, the following information is included: tasks modeled, dataset language, subject matter or type of dataset, BERT-based approaches taken, standard BERT performance (if attempted), pretrained BERT performance (if attempted), BioBERT performance (if attempted), SciBERT performance (if attempted), non-BERT model performance (if attempted), and non-BERT model type (if attempted). These performance metrics are compared to each other in order to answer the research questions and evaluate the hypotheses. Table 5.1 details the findings. The aggregated data is presented in section 5.1, followed by sections corresponding to each research question.

5.1 Aggregated Data From All Studies

The data was extracted and aggregated in order to answer the research questions and evaluate the hypotheses. Specifically, the performance metrics were included to make the following comparisons between models applied to the same task and dataset:

- BERT's performance versus the best-performing non-BERT model (RQ1).
- Domain-specific pretrained BERT's performance versus that of standard BERT (RQ2).
- BERT's performance versus the best-performing non-BERT model on datasets in languages other than English (RQ4).

- Domain-specific pretrained BERT's performance versus that of standard BERT in languages other than English (RQ4).

The performance metric used across different rows varies and is dependent upon the choice of each study's researchers (although each study used either *accuracy* or *F1-score*); however, all metrics within each row are the same type, making it possible to obtain a true apples-to-apples comparison. Abbreviations were used in order to limit the horizontal span of the table to one page. As such, a legend can be found below the table.

Table 5.1: *Results From Each Study Included in This Review*

Study	Task	Lang	Data Type	Approaches	BERT	Pre. BERT	BioBERT	SciBERT	Non Perf.	Non Type
[Akhtyamova, 2020]	NER	Spa	Clinical	Pretraining	84.00	**89.00**			87.00	FastText Embeddings
[Alsentzer et al., 2019]	NER	Eng	Biomedical	BioBERT + Pretraining	*79.70		***83.35**			
	TI	Eng	Biomedical	BioBERT + Pretraining	77.60%		**82.70%**			
	An	Eng	Biomedical	BioBERT + Pretraining	*93.35		***93.90**			
[Dai et al., 2019]	NER	Chi	Clinical	BERT-BiLSTM-CRF	***74.55**				*67.21	BiGRU - CRF
[Ding et al., 2019]	NER	Eng	Drug Labels	BioBERT + Ensemble			**62.91**			
	RE	Eng	Drug Labels	BioBERT			**46.77**			
	No	Eng	Drug Labels	BioBERT			**62.39**			
[García-Pablos et al., 2020]	An	Spa	Clinical	Fine-Tuning	**96.50**				95.10	spaCy
	Cl	Spa	Clinical	Fine-Tuning	**95.00**				89.50	spaCy
[Hakala and Pyysalo, 2019]	NER	Spa	Clinical	Fine-Tuning	**88.24**					
[Lee et al., 2019]	TI	Eng	Clinical	BioBERT + BiLSTM + Attention	80.90%		**82.40%**			
[Li et al., 2019a]	NER	Eng	Scientific	Multi-Task Learning, BioBERT	47.10		**51.90**		44.00	BILSTM + CRF
	RE	Eng	Scientific	Multi-Task Learning, BioBERT	81.80		**84.70**			
[Li et al., 2019b]	No	Eng	Clinical	BioBERT + Pretraining	*72.81		***73.49**			
[Li et al., 2020]	NER	Chi	Clinical	BERT-BiLSTM-CRF	90.50	**91.60**			87.90	BiLSTM + CRF
[Lin et al., 2019a]	CD	Chi	Clinical	Fine-Tuning	87.03				**89.95**	Rule-based Model

Table 5.1: *Results From Each Study Included in This Review*

Study	Task	Lang	Data Type	Approaches	BERT	Pre. BERT	BioBERT	SciBERT	Non Perf.	Non Type
	ED	Chi	Clinical	Fine-Tuning	63.54				**70.61**	DNN
[Lin et al., 2019b]	RE	Eng	Clinical	Pretraining, BioBERT	*61.80	*61.85	***62.45**			
[Liu et al., 2019]	Cl	Eng	Clinical	Pretraining	*76.00	***84.50**				
[Miftahutdinov et al., 2019]	No	Eng	Tweets	Fine-Tuning	**43.20**					
	Cl	Eng	Tweets	Fine-Tuning	**57.38**				51.64	SVM
	SL	Eng	Tweets	BioBERT + CRF			**65.80**			
[Peng et al., 2020]	NER	Eng	Biomedical, Clinical	Pretraining, Multi-Task Learning		***87.57**	*86.57			
	RE	Eng	Biomedical, Clinical	Pretraining, Multi-Task Learning		***76.97**	*76.07			
	TI	Eng	Clinical	Pretraining, Multi-Task Learning		**84.60%**	83.20%			
[Peng et al., 2019]	BLUE	Eng	Scientific	Pretraining, BioBERT		**82.30****	80.50**			
[Phongwattana and Chan, 2019]	NER	Eng	Scientific	Multi-Task Learning, BioBERT, Pre-training			***85.91**		*78.47	LSTM
[Sänger et al., 2019]	Cl	Ger	Scientific	Pretraining	77.80	**78.20**			72.50	SVM
[Song et al., 2019]	Cl	Chi	Diagnoses	BERT + CNN	**92.04%**				90.67%	One-Hot Character Encodings
[Sun and Yang, 2019]	NER	Spa	Biomedical	BioBERT	**89.24**		89.02			
[Sung et al., 2019]	Cl	Eng	Scientific	Pretraining	*80.03	***80.76**				
[Trieu et al., 2019]	CD	Eng	Biomedical	Parse Tree Filtering + Fine-Tuning	**45.50**				37.93	LSTM
[Wang et al., 2019]	Cl	Eng	Chemical	Pretraining		**75.89%**			70.38%	Seq3Seq

Table 5.1: *Results From Each Study Included in This Review*

Study	Task	Lang	Data Type	Approaches	BERT	Pre. BERT	BioBERT	SciBERT	Non Perf.	Non Type
[Xue et al., 2019]	NER	Chi	Biomedical	Multi-Task Learning	**96.89**				95.24	BiLSTM
	RE	Chi	Biomedical	Multi-Task Learning	**88.51**				87.29	BiLSTM
[Yu et al., 2019]	Cl	Eng	Scientific	Masked Sentence Model, Pretraining	86.19	91.15		86.81	**92.60**	BiLSTM + CRF
[Zhang et al., 2019a]	Cl	Chi	Clinical, Scientific	Fine-Tuning, Average Pooler	***77.15**				*65.50	Sequence Generation Model
[Zhang et al., 2019b]	NER	Chi	Clinical	BERT-BiLSTM-CRF	**88.45**				86.84	CNN-BiLSTM-CRF

* Metrics were generated by taking the average of multiple trials or datasets

** The BLUE Benchmark consists of ten datasets across five tasks and uses multiple evaluation metrics. Three datasets are used to perform NER, all of which use F1-score. Three datasets are used to perform relation extraction, two of which use micro F1-score and one of which uses macro F1-score. Two datasets are used to perform sentence similarity, all of which use the Pearson Correlation Coefficient. One dataset is used to perform document classification and uses the accuracy metric. One dataset is used to perform text inference and uses F1-score.

% Accuracy metric used in lieu of F1-score

Boldfaced score indicates best-performing model

Non Perf. = Performance of non-BERT model   Non Type = Type of non-BERT model   Pre. BERT = Metric from pretrained BERT model

NER = Named Entity Recognition   RE = Relation Extraction   No = Normalization   TI = Text Inference   Cl = Classification

CD = Coreference Detection   SL = Sequence Labeling   An = Anonymization   BLUE = BLUE Benchmark   ED = Ellipsis Detection

5.2 BERT Approaches Versus Non-BERT Approaches for Modeling Scientific and Medical Datasets (RQ1)

BERT was designed to be used as-is (with minimal fine-tuning) [Devlin et al., 2018], thus providing a salient alternative to more computationally expensive deep learning approaches. This paper seeks to answer the following question about how BERT is being used to model scientific and medical datasets:

- *RQ1: How does the performance of BERT-based approaches for modeling scientific and medical datasets compare to the performances of other approaches?*

The original hypothesis for RQ1 set forth at the beginning of this thesis was that BERT would outperform other approaches when modeling scientific and medical datasets. This hypothesis was supported: BERT-based approaches outperformed non-BERT approaches significantly. Fifteen studies compared the performance of at least one BERT-based model to a non-BERT model across nineteen different task-dataset combinations. Of the nineteen instances, fifteen of the comparisons yielded BERT as the better-performing model. If multiple BERT models were used during the comparison, the worst-performing model was used when calculating the increase (or decrease) in performance. This was done in an effort to, as much as possible, isolate the effect of using BERT-based models versus non-BERT models. ***On average, performance was enhanced 2.61% when applying the worst-performing BERT-based model compared to the best-performing non-BERT model.*** This was statistically significant, as a two-tailed paired t-test yielded $p = 0.0318$ which was less than the critical value of $p = 0.05$ needed to reject the null hypothesis that BERT-based and non-BERT models performed equally.

Of the four instances where BERT was not the best-performing model, two instances were deep learning models, one instance was a rule-based model, and one instance used

FastText embeddings (fasttext.cc). Interestingly, the study that used FastText embeddings applied BERT to a Spanish dataset, and a pretrained version of multilingual BERT resulted in a greater F1-score than the FastText embeddings [Akhtyamova, 2020] despite standard BERT's underperformance.

The prevalent approaches to applying BERT to scientific and medical datasets include fine-tuning, further pretraining (either manually or by using a domain-specific model such as BioBERT), ensembling, or a combination thereof. The most frequent method of modeling scientific and medical datasets through alternative means was via a BiLSTM approach. Figure 5.1 illustrates the averages of the performance metrics (accuracy of F1-score) when comparing BERT-based models to non-BERT models when measured on the same task-dataset combination. Table 5.2 depicts the individual comparisons of BERT's performance to non-BERT models applied to the same tasks and datasets.

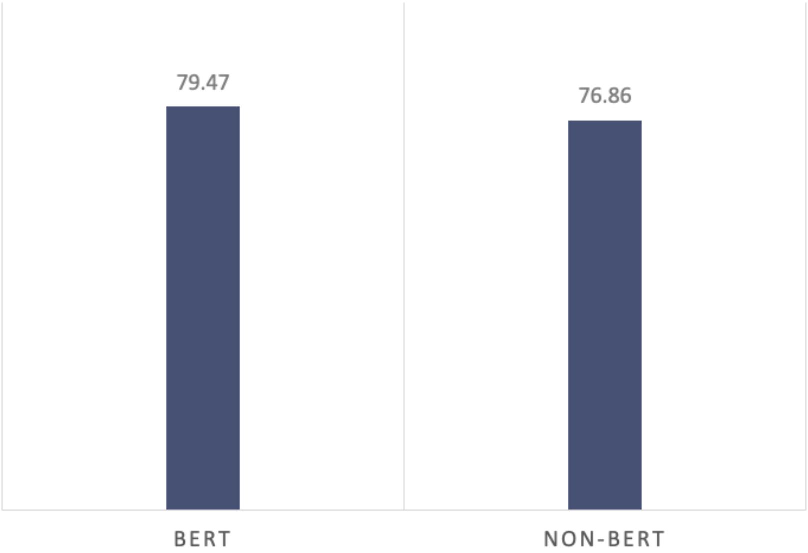


Figure 5.1: *Comparison of Performance of BERT-Based Approaches to Non-BERT Approaches*

Table 5.2: *BERT Versus Non-BERT Approaches*

Study	**BERT**	**Pre. BERT**	**Bio-BERT**	**Sci-BERT**	**Non Perf.**	**Non Type**
[Dai et al., 2019]	**74.55**				67.21	BiGRU - CRF
[García-Pablos et al., 2020]	**96.50**				95.10	spaCy
	95.00				89.50	spaCy
[Li et al., 2019a]	47.10		**51.90**		44.00	BILSTM + CRF
[Li et al., 2020]	90.50	**91.60**			87.90	BiLSTM + CRF
[Lin et al., 2019a]	87.03				**89.95**	Rule-based Model
	63.54				**70.61**	DNN
[Phongwattana and Chan, 2019]			**85.91**		78.47	LSTM
[Sänger et al., 2019]	77.80	**78.20**			72.50	SVM
[Song et al., 2019]	**92.04%**				90.67%	One-Hot Character Encodings
[Trieu et al., 2019]	**45.50**				37.93	LSTM
[Wang et al., 2019]		**75.89%**			70.38%	Seq3Seq
[Xue et al., 2019]	**96.89**				95.24	BiLSTM
	88.51				87.29	BiLSTM
	57.38				51.64	SVM
[Yu et al., 2019]	86.19	91.15		86.81	**92.60**	BiLSTM + CRF
[Zhang et al., 2019a]	**77.15**				65.50	Sequence Generation Model
[Zhang et al., 2019b]	**88.45**				86.84	CNN-BiLSTM-CRF

% Accuracy metric used in lieu of F1-score

Boldfaced score indicates best-performing model

Non Perf. = Performance of non-BERT model

Non Type = Type of non-BERT model

Pre. BERT = Metric from pretrained BERT model

5.3 BERT Efficacy Versus Pretrained BERT Models (RQ2)

Several researchers have taken BERT and further pretrained it with large, domain-specific scientific and medical corpora. These models were then released with the intention of being used as-is, just as BERT was. Of these models, BioBERT is currently the most used with 538 citations on Google Scholar. SciBERT trails in a distant second with 178 citations. This paper seeks to answer the following question with regard to standard BERT's efficacy relative to its further-pretrained peers:

- *RQ2: How does BERT's performance on scientific and medical datasets compare to further-pretrained, domain-specific BERT derivatives (e.g. BioBERT and SciBERT)?*

The original hypothesis for RQ2 set forth at the beginning of this thesis was that BERT models pretrained on scientific and medical corpora would outperform standard BERT on datasets consisting of similar subject matter as the data seen during pretraining. This hypothesis was supported: pretrained BERT models significantly outperformed their non-pretrained counterparts. Twelve studies compared standard BERT to a BERT model that had undergone domain-specific pretraining. These twelve studies compared BERT models across fifteen different task-dataset combinations. Of the fifteen instances, fourteen reported that further pretraining on domain-specific data augmented BERT's performance. If multiple pretrained BERT models were used during the comparison, the best-performing pretrained model was used when calculating the increase (or decrease) in performance. This was done in an effort to identify the maximum performance enhancement that was achieved when using a pretrained BERT model in lieu of standard BERT. ***On average, performance was enhanced 2.69% when applying the best-performing pretrained BERT model compared to standard BERT.*** This was statistically significant, as a two-tailed paired t-test yielded *p = 0.0011* which was less than the critical value of *p = 0.05* needed to reject the null hypothesis that pretrained BERT and standard BERT models performed equally. This indicates that, in the reviews selected for this study, further pretraining BERT on domain-specific data was just as impactful as using BERT over another model altogether. This is evidenced in Table 5.3, where one can see that nearly all of the best-performing models from each row incorporated domain-specific pretraining.

Additionally, Figure 5.2 illustrates the averages of pretrained BERT's performance versus that of standard BERT models when applied to the same task-dataset combinations.

Table 5.3: *Pretrained BERT Versus Standard BERT*

Study	**BERT**	**Pre. BERT**	**BioBERT**	**SciBERT**
[Akhtyamova, 2020]	84.00	**89.00**		
[Alsentzer et al., 2019]	79.70		**83.35**	
	77.60%		**82.70%**	
	93.35		**93.90**	
[Lee et al., 2019]	80.90%		**82.40%**	
[Li et al., 2019a]	47.10		**51.90**	
	81.80		**84.70**	
[Li et al., 2019b]	72.81		**73.49**	
[Li et al., 2020]	90.50	**91.60**		
[Lin et al., 2019b]	61.80	61.85	**62.45**	
[Liu et al., 2019]	76.00	**84.50**		
[Sänger et al., 2019]	77.80	**78.20**		
[Sun and Yang, 2019]	**89.24**		89.02	
[Sung et al., 2019]	80.03	**80.76**		
[Yu et al., 2019]	86.19	**91.15**		86.81

% Accuracy metric used in lieu of F1-score

Boldfaced score indicates best-performing model

Pre. BERT = Metric from pretrained BERT model

In the one study where further pretraining was less effective than standard BERT, the researchers compared BioBERT to multilingual BERT on a Spanish dataset [Sun and Yang, 2019]. Although BioBERT underperformed multilingual BERT, BioBERT attained an F1-score only 0.22% lower than multilingual BERT. This is interesting, considering that BioBERT is only trained on English corpora. The authors of the study ([Sun and Yang, 2019]) surmised that this is likely due to the fact that many chemicals and proteins have the same names in Spanish as they do in English, but they did not provide evidence to corroborate this as being an explanation for comparable performance

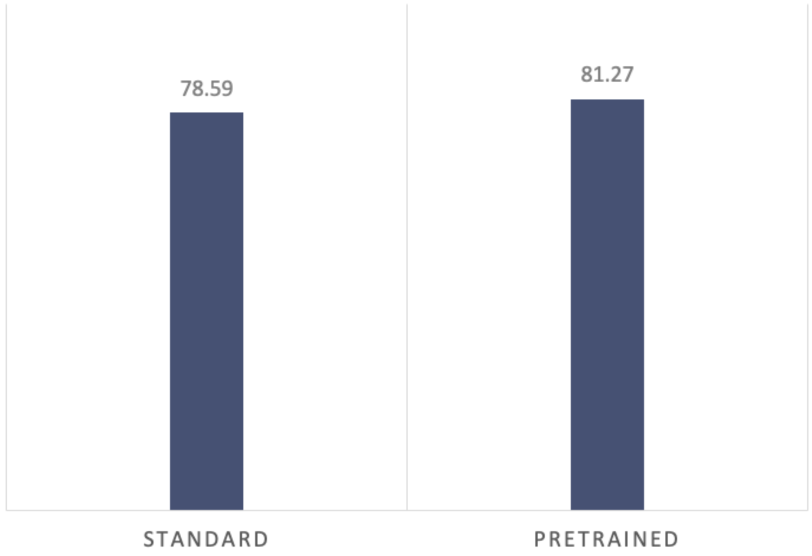


Figure 5.2: *Comparison of Performance of Pretrained BERT to Standard BERT*

between the two models.

Of the twelve studies that incorporated pretrained BERT models (and compared them to standard BERT), only one study selected for this review ([Yu et al., 2019]) used SciBERT. SciBERT underperformed its manually pretrained counterpart by 4.34%, and both models underperformed the SOTA for the task—a BiLSTM whose performance bested the pretrained BERT model by 1.45%. Both pretrained models did, however, outperform standard BERT. BioBERT outperformed all other models (standard BERT and other pretrained BERT models) in all instances except for the study by [Sun and Yang, 2019].

5.4 Problems Encountered Applying BERT to Scientific and Medical Datasets (RQ3)

Much of the obscure verbiage present in scientific and medical datasets is not part of BERT's vocabulary. As a result, these words are separated by BERT into WordPiece tokens. While these WordPiece tokens are thought to be better than the alternative (OOV tokens), they nevertheless create a challenge in NLP for datasets riddled with obscure words and phrases. This paper seeks to answer the following question with regard to the different types of problems incurred while applying BERT to scientific and medical datasets:

- *RQ3: What types of problems does BERT encounter when presented with scientific and medical datasets?*

There was no corresponding hypothesis for this research question, as this research question merely seeks to document the most prevalent problems that are encountered when applying BERT to scientific and medical datasets. Of the twenty-seven studies selected for this review, nineteen of them identified specific problems that were encountered while applying BERT to scientific and medical datasets. The most common problems dealt with the data being imbalanced, the number of classes being numerous, and the dataset texts being too long and spanning too many input sequences. Figure 5.3 illustrates how many studies mentioned each particular problem encountered. Table 5.4 illustrates which studies reported which problems when applying BERT to scientific and medical datasets. Only the studies that reported problems are enumerated in the table. The specific problems encountered are outlined in the subsections that follow.

5.4.1 Imbalanced Data

Five of the nineteen studies that mentioned problems applying BERT to scientific and medical datasets referenced dataset imbalance. Even in cases where the dataset was relatively balanced, a single dominating class (a control group, for instance) could cause the model to perform preferentially in favor of the more prevalent group. As a result, the model's recall could drop substantially. While this issue is not specific to scientific and medical datasets, it occurs more frequently due to the domains' tendency to include a

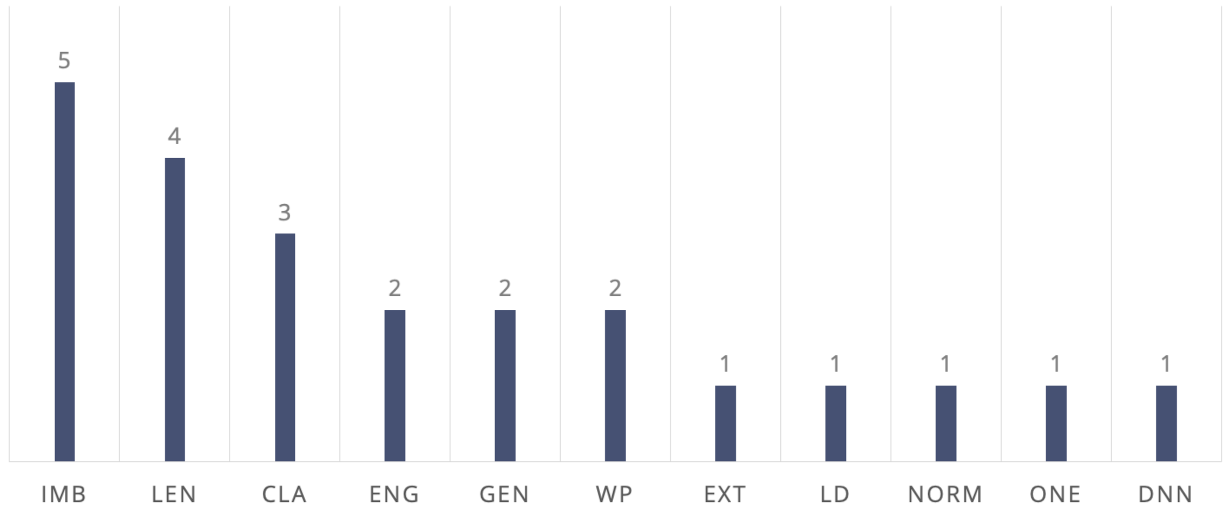


Figure 5.3: *Number of Studies Referencing Each Issue*

Ext = Model does not extrapolate to other datasets
LD = Large amounts of data needed for further pretraining
Len = Length of text spans multiple input sequences
Cla = Number of classes is too numerous
Imb = Data imbalance
Norm = Difficult to normalize terms consisting of multiple words

Eng = Difficult to find pretraining data in non-English languages
Gen = Hard to generalize domain-specific model to canonical text
WP = Problems with WordPiece embeddings
One = Models often only trained on one dataset
DNN = Deep learning still more effective at certain tasks than BERT

"control" group in studies. For instance, when extracting the side effects of a drug from a Randomized Controlled Trial's (RCT) clinical notes, the "asymptomatic" class would occur far more frequently than any other class (side effect). Because of this, scientific and medical datasets often have imbalanced classes, making it difficult to effectively model them.

5.4.2 Length of Text

Of the nineteen studies that mentioned problems applying BERT to scientific and medical datasets, four of them referenced the length of text as being an issue. Although BERT has been shown to effectively provide context within instances, this context does not extend across multiple sentences that span multiple inputs. BERT limits its input instances to 512 tokens, so contextualization outside of that parameter proves difficult. Tasks like relation extraction and coreference resolution become increasingly difficult the farther

Table 5.4: *Problems Reported by All Studies*

Title	Ext	LD	Len	Cla	Imb	Norm	Eng	Gen	WP	One	DNN
[Akhtyamova, 2020]							X				
[Alsentzer et al., 2019]								X			
[Ding et al., 2019]	X										
[Hakala and Pyysalo, 2019]									X		
[Li et al., 2019a]					X						
[Li et al., 2019b]				X	X	X					
[Li et al., 2020]			X								
[Lin et al., 2019a]											X
[Lin et al., 2019b]		X									
[Liu et al., 2019]									X		
[Miftahutdinov et al., 2019]					X						
[Peng et al., 2019]										X	
[Song et al., 2019]			X	X	X						
[Sun and Yang, 2019]							X				
[Sung et al., 2019]								X			
[Trieu et al., 2019]			X								
[Xue et al., 2019]					X						
[Yu et al., 2019]			X								
[Zhang et al., 2019a]				X							
Totals	1	1	4	3	5	1	2	2	2	1	1

Ext = Model does not extrapolate to other datasets
LD = Large amounts of data needed for further pretraining
Len = Length of text spans multiple input sequences
Cla = Number of classes is too numerous
Imb = Data imbalance
Norm = Difficult to normalize terms consisting of multiple words

Eng = Difficult to find pretraining data in non-English languages
Gen = Hard to generalize domain-specific model to canonical text
WP = Problems with WordPiece embeddings
One = Models often only trained on one dataset
DNN = Deep learning still more effective at certain tasks than BERT

away elements reside in the text. This is problematic in scientific and medical corpora, as many texts (scientific papers and clinical reports, for example) span multiple pages and often only mention a particular part of a relation one time, early in the text. For instance, a subject's clinical notes for an RCT could mention the drug in the first paragraph and then spend ten pages discussing its side effects. Trying to extract the relationship between the drug on page one and a side effect on page ten could prove difficult. [Yu et al., 2019] tried to address this issue via a Masked Sentence Model (as opposed to a word-level Masked

Language Model). Their result outperformed BERT, but it did not outperform the previous SOTA for the task which was a bidirectional (fully trained, not transfer learning) RNN.

5.4.3 Number of Classes

Of the nineteen studies that mentioned problems applying BERT to scientific and medical datasets, three of them referenced the large number of classes as being an issue. A large number of classes is common in scientific and medical corpora, as many of the tasks involve classifying diverse label arrays such as chemical compounds, clinical symptoms, and medical conditions. This is already an issue in NLP with canonical corpora, but the problem is magnified with scientific and medical texts. As mentioned earlier, these domain-specific corpora often have a null hypothesis class, and an increase in class number tends to further prejudice the model toward the prevalent class.

Furthermore, an increase in class size often yields sparsely populated datasets (one is presumably more likely to encounter "fatigue" as a drug side effect than "temporary blindness," for example). Because of this, many classes will inevitably have very few instances from which to train, making already-abstract concepts more difficult to predict. Additionally, one is likely to encounter many zero-shot instances, where the model is forced to try and predict a class that it was never exposed to during training. I found this challenge particularly difficult in a previous work with [Cochran et al., 2020]. BERT was applied to Hughes' [Hughes, 2019] work, which tried to detect causal relations in short-answer essays written by Chicago adolescents regarding the causes of coral bleaching and skin cancer (separate datasets). Because there were dozens of possible causal chains, and only about eight thousand training instances (sentences) per dataset (most of which were in the "no relation" class), evaluations of the test sets always encountered multiple instances of classes that had never been seen by the model during training.

5.4.4 Other Issues

Other problems were identified while applying BERT to scientific and medical datasets but were not as frequently reported. As such, the issues addressed by two or fewer papers were combined to comprise this subsection.

Two studies referenced the lack of availability of large scientific and medical corpora for conducting further pretraining in languages other than English. Both of these studies were referencing Spanish datasets, specifically. This is significant considering that there were only four studies that modeled Spanish datasets chosen for inclusion in this review and two of them mentioned this issue. This review has alluded to the effectiveness of domain-specific pretraining and domain-specific pretrained models, so lack of data for further pretraining on non-English datasets needs to be addressed in future works.

Two studies referenced difficulties generalizing domain-specific models to canonical texts. Although domain-specific models are effective at modeling their domains, they can "forget" their previous knowledge and become ineffective when applied to datasets even slightly outside of the domain of their pretraining. For example, multiple studies further pretrained BioBERT with clinical corpora. These studies would outperform standard BioBERT on clinical data, but would then underperform standard BioBERT when applied to scientific (non-clinical) data.

Two studies referenced difficulties with WordPiece embeddings. Because both BERT and BioBERT (not SciBERT) use canonical vocabularies, many of the more technical words in scientific and medical datasets are represented as WordPiece embeddings and not actual words. While further pretraining can improve the contextual vectorizations of these embeddings, it is likely not as effective as it would be had the words themselves been present in the underlying lexicon.

One study expressed difficulty reproducing its results with additional datasets. One study referenced the massive amount of data needed for additional pretraining. One study mentioned the difficulty normalizing instances of multiple words. One study referenced the fact that many studies evaluate their models on only a single dataset. One study mentioned that deep learning is still more effective for certain tasks.

5.5 BERT Performance on Non-English Datasets (RQ4)

As of this writing, Google has released three different versions of BERT (discriminated by language): English BERT, Chinese BERT, and multilingual BERT. The current iteration of multilingual BERT was trained on 104 languages, but it is the same size as the other BERT models in Chinese and English (twelve Transformer layers, 768 hidden units, twelve attention heads, 110 million parameters). Because of this, it was initially surmised at this thesis' outset that multilingual BERT would not be as effective on non-English datasets as BERT-base or BERT-large is on English datasets. Additionally, further pretraining BERT in languages other than English is difficult, as scientific and medical corpora large enough from which to conduct additional pretraining are much less readily available. Both BioBERT and SciBERT, for example, were only further pretrained on English corpora. It is also important to note that Chinese BERT implements character-level embeddings as opposed to word-level ones, as (unlike English characters) Chinese characters convey semantic meaning when expressed individually. This paper seeks to answer the following question with regard to applying BERT to scientific and medical datasets in languages other than English:

- *RQ4: How well does BERT perform on scientific and medical datasets that are in languages other than English?*

The original hypothesis for RQ4 set forth at the beginning of this thesis was that BERT would not perform as well as other approaches when modeling scientific and medical datasets in languages other than English. This hypothesis was not supported: BERT did not significantly underperform other methods when applied to non-English scientific and medical datasets. Twelve studies in this review applied BERT to non-English scientific and medical datasets: seven applied Chinese BERT to Chinese datasets, four applied multilingual BERT to Spanish datasets, and one applied multilingual BERT to a German dataset. These twelve studies modeled a total of fifteen task-dataset combinations.

Of the twelve studies that applied non-English BERT models to scientific and medical datasets, ten of them compared BERT's performance to a non-BERT method. These ten studies modeled a total of thirteen task-dataset combinations. Of the thirteen instances, eleven of them reported BERT as being the superior approach. On average, performance was enhanced 2.05% when applying standard BERT compared to a non-BERT model when modeling scientific and medical data in a language other than English; however, the performance gain was not as substantial as the 3.83% increase obtained by choosing BERT over a non-BERT model when modeling these same types of datasets in English. Additionally, a two-tailed paired t-test yielded $p = 0.1507$, which was not lower than the critical value $p = 0.05$. Thus, the null hypothesis that both BERT and non-BERT models performed equally on non-English scientific and medical datasets could not be rejected. However, it is possible that there does exist a statistically significant outperformance by BERT and that there were simply too few available instances from which to determine statistical significance. This possibility should be explored in future work. Figure 5.4 illustrates the average performance metrics broken down by language (including the av-

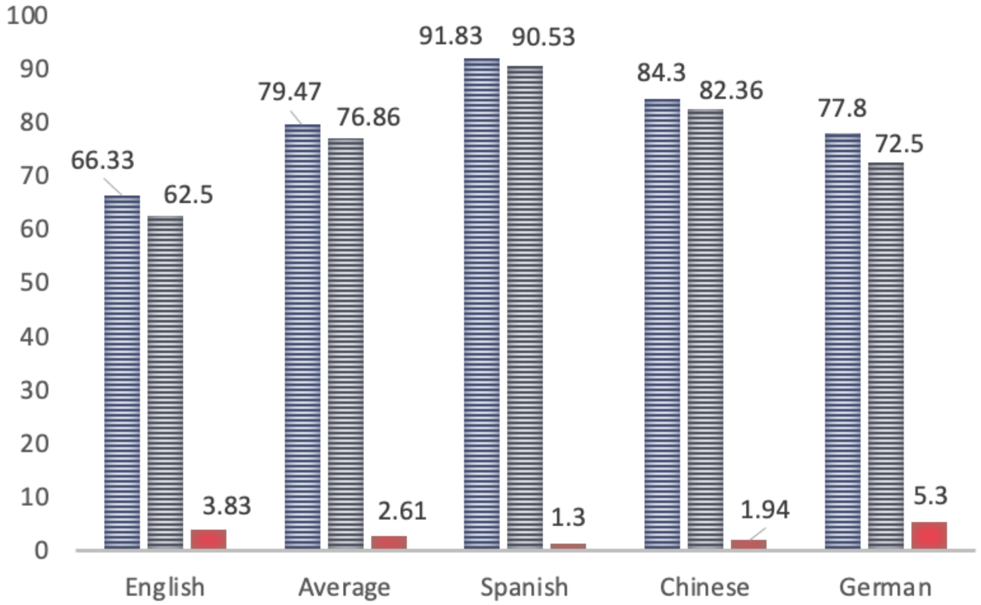


Figure 5.4: *Comparison of Performances of BERT-Based and Non-BERT Approaches Across Languages*

erage across all languages). For each language, the blue bar represents the BERT-based model, the gray bar represents the non-BERT model, and the red bar represents the performance gained from opting for the BERT-based model in lieu of the non-BERT one. The performance metrics for the twelve studies can be seen individually in Table 5.5. Both instances where BERT was not the superior model were part of the same (Chinese) study. The authors, [Lin et al., 2019a], achieved better results applying a rule-based model to the task of coreference detection and a deep learning model to the task of ellipsis detection.

Additionally, three studies (one in each non-English language) compared pretrained BERT models to standard BERT. All three reported improved results with domain-specific pretraining. On average, performance was enhanced 2.17% when applying a pretrained BERT model compared to standard BERT when modeling scientific and medical data in a language other than English; however, the performance gain was not as substantial as the 3.09% increase obtained when opting for pretrained BERT over standard BERT when modeling these same types of datasets in English. Additionally, the null hypothesis that

Table 5.5: *BERT Approaches Versus Non-BERT Approaches on Non-English Datasets*

Study	**Lang**	**BERT**	**Non Perf.**	**Non Type**
[Akhtyamova, 2020]	Spa	**84.00**	87.00	FastText Embeddings
[Dai et al., 2019]	Chi	**74.55**	67.21	BiGRU - CRF
[García-Pablos et al., 2020]	Spa	**96.50**	95.10	spaCy
	Spa	**95.00**	89.50	spaCy
[Li et al., 2020]	Chi	**90.50**	87.90	BiLSTM + CRF
[Lin et al., 2019a]	Chi	87.03	**89.95**	Rule-based Model
	Chi	63.54	**70.61**	DNN
[Sänger et al., 2019]	Ger	**77.80**	72.50	SVM
[Song et al., 2019]	Chi	**92.04%**	90.67%	One-Hot Character Encodings
[Xue et al., 2019]	Chi	**96.89**	95.24	BiLSTM
	Chi	**88.51**	87.29	BiLSTM
[Zhang et al., 2019a]	Chi	**77.15**	65.50	Sequence Generation Model
[Zhang et al., 2019b]	Chi	**88.45**	86.84	CNN-BiLSTM-CRF

% Accuracy metric used in lieu of F1-score

Boldfaced score indicates best-performing model

Non Perf. = Performance of non-BERT model

Non Type = Type of non-BERT model

both pretrained BERT and standard BERT models performed equally on non-English scientific and medical datasets was unable to be rejected due to a two-tailed paired t-test yielding p = *0.2692*, which was greater than the p = *0.05* value needed for rejection. It is possible that there does exist a statistically significant outperformance by pretrained BERT and that there were simply too few available instances from which to determine statistical significance. This possibility should also be explored in future work. One study applied BioBERT to a Spanish dataset and barely underperformed multilingual BERT. This instance was not counted toward the net effect of further pretraining, however, because BioBERT was trained only on English corpora. The performance metrics for the three studies can be seen in Table 5.6.

Table 5.6: *Pretrained BERT Versus Standard BERT on Non-English Datasets*

Study	Lang	BERT	Pre. BERT
[Akhtyamova, 2020]	Spa	84.00	**89.00**
[Li et al., 2020]	Chi	90.50	**91.60**
[Sänger et al., 2019]	Ger	77.80	**78.20**

Boldfaced score indicates best-performing model

Pre. BERT = Metric from pretrained BERT model

Chinese. Seven studies applied Chinese BERT to nine different task-dataset combinations. Seven of the nine instances (six of seven studies) reported that Chinese BERT outperformed other non-BERT approaches. One study that applied Chinese BERT to scientific and medical datasets conducted domain-specific pretraining, and that instance was the best-performing. On average, Chinese BERT outperformed non-BERT approaches on Chinese datasets by 1.94%. The study that pretrained BERT with Chinese domain-specific data saw an additional 1.10% increase in performance relative to standard BERT. However, neither of these numbers was statistically significant. A two-tailed paired t-test for BERT's comparison to non-BERT approaches in Chinese yielded $p = 0.3093$ which was not below the critical level of $p = 0.05$ needed to reject the null hypothesis that the two approaches are equal in performance. No two-tailed paired t-test was performed comparing pretrained BERT to standard BERT in Chinese, as there was only one instance of a pretrained BERT model being applied to a Chinese dataset. The performance metrics for the seven studies that modeled Chinese datasets can be seen in Table 5.7.

Spanish. Four studies applied multilingual BERT to five different task-dataset combinations. Of the three instances where BERT was compared to other approaches, BERT outperformed the other approaches in two instances. In the instance where BERT did not outperform, FastText embeddings were used; however, once BERT was pretrained on domain-specific data, the BERT model outperformed the FastText embeddings. One

Table 5.7: *Chinese BERT Performance Metrics*

Study	Lang	BERT	Pre. BERT	Non Perf.	Non Type
[Dai et al., 2019]	Chi	**74.55**		67.21	BiGRU - CRF
[Li et al., 2020]	Chi	90.50	**91.60**	87.90	BiLSTM + CRF
[Lin et al., 2019a]	Chi	87.03		**89.95**	Rule-based Model
	Chi	63.54		**70.61**	DNN
[Song et al., 2019]	Chi	**92.04%**		90.67%	One-Hot Character Encodings
[Xue et al., 2019]	Chi	**96.89**		95.24	BiLSTM
	Chi	**88.51**		87.29	BiLSTM
[Zhang et al., 2019a]	Chi	**77.15**		65.50	Sequence Generation Model
[Zhang et al., 2019b]	Chi	**88.45**		86.84	CNN-BiLSTM-CRF

% Accuracy metric used in lieu of F1-score

Boldfaced score indicates best-performing model

Pre. BERT = Metric from pretrained BERT model

Non Perf. = Performance of non-BERT model

study applied BioBERT to a Spanish dataset. While BioBERT did not outperform multilingual BERT, the results were similar (within 0.22%). This is interesting, as BioBERT is trained on only English corpora. On average, multilingual BERT outperformed non-BERT approaches on Spanish datasets by 1.30%. However, this was not statistically significant. A two-tailed paired t-test for BERT's comparison to non-BERT approaches in Spanish yielded $p = 0.6492$ which was not below the critical level of $p = 0.05$ needed to reject the null hypothesis that the two approaches are equal in performance. One study pretrained multilingual BERT on Spanish scientific and medical corpora before applying it to the dataset and realized a 5% performance increase over standard BERT and a 2% increase over FastText embeddings. Again, this was not statistically significant, as there was only a single study that further pretrained BERT and applied it to a Spanish dataset. The performance metrics for the four studies that modeled Spanish datasets can be seen in Table 5.8.

Table 5.8: *Multilingual BERT Performance Metrics on Spanish Datasets*

Study	**Lang**	**BERT**	**Pre. BERT**	**BioBERT**	**Non Perf.**	**Non Type**
[Akhtyamova, 2020]	Spa	84.00	**89.00**		87.00	FastText Embeddings
[García-Pablos et al., 2020]	Spa	**96.50**			95.10	spaCy
	Spa	**95.00**			89.50	spaCy
[Hakala and Pyysalo, 2019]	Spa	**88.24**				
[Sun and Yang, 2019]	Spa	**89.24**		89.02		

Boldfaced score indicates best-performing model

Pre. BERT = Metric from pretrained BERT model

Non Perf. = Performance of non-BERT model

German. One study applied multilingual BERT to a German dataset. The study reported that standard BERT outperformed a support vector machine (SVM) approach by 5.3%, and further pretraining BERT on domain-specific data yielded another 0.4% improvement over standard BERT. None of the performance increases were statistically significant, however, as there was only one study selected for inclusion in this review that applied BERT to a dataset in German. The performance metrics for the study that modeled the German dataset can be seen in Table 5.9.

Table 5.9: *Multilingual BERT Performance Metrics on German Datasets*

Study	**Lang**	**BERT**	**Pre. BERT**	**Non Perf.**	**Non Type**
[Sänger et al., 2019]	Ger	77.80	**78.20**	72.50	SVM

Boldfaced score indicates best-performing model

Pre. BERT = Metric from pretrained BERT model

Non Perf. = Performance of non-BERT model

Non Type = Type of non-BERT model

CHAPTER 6

DISCUSSION

This discussion will address the principal findings as they pertain to the research questions, the limitations of the review, the implications of the findings going forward, and topics to address in future works.

6.1 Principal Findings

For RQ1, BERT's performance relative to other approaches for modeling scientific and medical datasets, BERT outperformed the other approaches by an average of 2.61%, which was statistically significant. This was particularly pronounced in English, where BERT-based models outperformed non-BERT models by an average of 3.83%. Fifteen studies compared BERT-based approaches to other approaches across nineteen task-dataset combinations, and BERT was the best-performing model in fifteen of those cases. The prevalent methods of applying BERT to scientific and medical datasets included fine-tuning, further pretraining (either manually or by using a domain-specific model such as BioBERT), ensembling, or a combination thereof.

In the four instances where the BERT-based model was not the superior approach, two used deep learning approaches, one used FastText embeddings, and one used a rule-based model. Of theses same four instances where BERT was not the best-performing model, only one applied BERT to an English dataset. One study (two instances) applied Chinese BERT to a Chinese dataset, and one study applied multilingual BERT to a Spanish dataset. In the study where multilingual BERT underperformed on the Spanish dataset, further pretraining increased BERT's performance, and the pretrained BERT model ultimately eclipsed both the standard BERT and non-BERT methods. The most frequent method of modeling scientific and medical datasets through alternative (non-BERT) means was via a BiLSTM approach.

For RQ2, standard BERT versus domain-specific pretrained models, pretrained models outperformed standard BERT by an average of 2.69%. This was statistically significant and was particularly pronounced in English, where pretrained BERT models outperformed standard BERT by an average of 3.09%. Twelve studies compared the performance of standard BERT to at least one other pretrained BERT model across fifteen different task-dataset instances, where fourteen instances reported pretrained BERT models as being superior to standard BERT. This indicates that further pretraining BERT before fine-tuning was just as impactful as modeling a task with BERT in lieu of a non-BERT model.

The one study where further pretraining BERT had an adverse effect on the model's performance was when BioBERT was applied to a Spanish dataset. However, even on a Spanish dataset, BioBERT only underperformed standard BERT by a slim margin, which indicates that BioBERT has the potential to be effectively applied to scientific and medical corpora in languages other than, but alphabetically similar to, English.

For RQ3, problems encountered while applying BERT to scientific and medical corpora, a few issues were frequently mentioned. The most frequent problem identified by researchers was the data imbalance that pervades many scientific and medical corpora. Although not a problem specific to scientific and medical datasets, data imbalance is a significant issue in these domains due to these datasets' penchant for uneven data distribution between classes. RCTs for drug trials, for example, always have a "control" group, so it is reasonable to expect that a null hypothesis class will significantly outweigh each individual class that pertains to a specific drug side effect. Similar issues were encountered with [Cochran et al., 2020] while trying to use BERT to identify causal relations in short-answer essays written by high school students: most sentences did not contain any causal relation, thereby biasing the model in favor of the “null” class. Some studies were able to mitigate the issue of data imbalance by manually balancing the data during preprocessing.

Another issue plaguing researchers attempting to apply BERT to scientific and medical datasets is the large amount of classes endemic to scientific and medical texts. Classification becomes difficult when the number of classes is numerous—especially if the dataset itself is small in size. I experienced this problem first-hand as well during the same work with [Cochran et al., 2020]. Each of the datasets in that study contained roughly eight thousand to ten thousand instances split between fifty to one hundred classes. The sparse population of the datasets created problems during training, as many of the classes were woefully underrepresented. This precipitated a substantial amount of (largely unsuccessful) zero-shot learning.

The other issue that is prevalent when applying BERT to scientific and medical corpora is the issue concerning length of text. Tasks like relation extraction and coreference resolution become increasingly difficult when these connections must be identified across multiple pages of text. BERT has been shown to be astute at identifying relationships between sentences but only if those sentences are in the same input sequence. For sequences separated by a significant amount of text, BERT often struggles to identify relationships between them. One study, for example, reported that deep learning is still more effective than BERT-based approaches for coreference resolution in Chinese [Lin et al., 2019a]. This issue becomes more pronounced the farther apart the sequences are from each other.

For RQ4, BERT’s performance on non-English datasets, results were evaluated across three languages: Chinese, Spanish, and German. The Chinese studies were conducted with Chinese BERT, and the Spanish and German studies were conducted with multilingual BERT. On average, BERT’s application to scientific and medical datasets in languages other than English outperformed non-BERT models by 2.05%, but this result was not statistically significant. Additionally, while a noticeable improvement over the non-BERT approaches, the augmentation in performance was not as pronounced as BERT's 3.83% performance improvement over non-BERT models when applied to English datasets. Ten studies in this review compared BERT's performance (across thirteen different task-dataset combinations) to other approaches when applying BERT to non-English scientific and medical datasets. Of the thirteen instances, eleven of them reported BERT as being the superior approach. Both instances where BERT was not the best-performing model were from the same Chinese study ([Lin et al., 2019a]).

Additionally, three studies compared pretrained BERT models to standard BERT. All three reported improved results with domain-specific pretraining, indicating that domain-specific pretraining can be effective on scientific and medical datasets in languages other than English. On average, further pretraining increased BERT's performance on non-English datasets by an average of 2.17%, but this was not statistically significant. The difference was noticeable, but it was also less than the 3.09% performance increase seen when further pretraining BERT for English datasets.

Seven studies across nine task-dataset combinations applied BERT to scientific and medical datasets in Chinese. Seven of the nine reported the BERT-based approach as being the best approach. On average, Chinese BERT outperformed non-BERT approaches on Chinese datasets by 1.94%. Additionally, one study further pretrained Chinese BERT on domain-specific data. That study saw a 1.10% performance increase when compared to standard BERT. However, none of the results specific to Chinese datasets were statistically significant.

Three studies compared multilingual BERT with at least one other non-BERT approach when modeling scientific and medical datasets in Spanish. BERT outperformed the other approaches in two instances of the three. On average, multilingual BERT outperformed non-BERT approaches on Spanish datasets by 1.30%. In the instance where BERT did not outperform, [Akhtyamova, 2020] opted for FastText embeddings; however, once BERT was pretrained on domain-specific data, the BERT model outperformed the FastText embeddings. This study saw further pretraining precipitate a 5.0% performance improvement over standard BERT and a 2.0% improvement over FastText embeddings. However, none of the results specific to Spanish datasets were statistically significant.

One study applied multilingual BERT to a German dataset and reported that standard BERT outperformed an SVM approach by 5.3%. This same study also pretrained BERT on domain-specific corpora prior to fine-tuning and reporting a 0.4% increase in performance relative to standard BERT. Because there was only one study included in this review that applied BERT to a German dataset, none of the German-specific results were statistically significant.

6.2 Limitations

Because this paper is a thesis, and therefore completed by a lone researcher, I was not able to leverage [Kitchenham, 2004]'s guidance with regard to SLRs that calls for each author to rate a potential source for inclusion and then use the aggregation of the ratings (and the accompanying discourse) to ultimately deem a source worthy of inclusion. Furthermore, in an effort to include the latest research in a limited body of works, I opted to include papers in this review that had been accepted for publication but not necessarily published. As such, it is possible that a paper accepted for publication in a journal or inclusion in proceedings was subsequently withdrawn and therefore never peer-reviewed.

Additionally, the search for studies to include in this review defined a search window from BERT's release in October 2018 through May 2020 (when the search was conducted). 425 unique studies were found in the aforementioned databases (including arXiv and

Google Scholar) that included “bert” in the title. On October 18, 2020, a cursory search yielded over 1,100 such papers. It follows, then, that there is a large corpus of research that has been published since the initial search and is not included (or ever considered for inclusion) in this review.

As with any literature review, there exists the possibility of works being overlooked or unavailable for inclusion due to publication bias (i.e., the file-drawer problem en.wikipedia.org). BERT is currently a popular architecture for modeling many different NLP tasks. As such, it is possible that researchers may abstain from publishing results where BERT underperforms other approaches. This could lead to a collective bias toward BERT in published BERT-related research.

The scope of the review must also be considered. Statistical significance was determined across twenty-seven studies, and in no case was the number of instances (task-dataset combinations) used for comparison greater than nineteen. Furthermore, with regard to the instances where the null hypothesis could not be rejected (non-English BERT versus non-English non-BERT), it was not the case that BERT *underperformed* non-BERT models on non-English scientific and medical datasets. The null hypothesis was that the two model types performed equally, so failure to reject the null hypothesis merely indicates that the data were unable to prove otherwise. This means that I was unable to reject or confirm my own hypothesis that BERT would not be the best-performing model on non-English scientific and medical datasets and that additional research is still needed (with a greater number of sources) to effectively evaluate this hypothesis.

Lastly, while it is believed that the efficacy of domain-specific pretraining before the fine-tuning process will result in more effective models for additional domains other than science and medicine, this research was limited to scientific and medical datasets. Therefore, further research must be conducted to determine whether or not other domains can benefit from domain-specific pretraining and domain-specific modeling. Additionally, the results from this review are not meant to be interpreted as the “best” ways to apply BERT to scientific or medical datasets in an absolute sense; instead, this work merely seeks to: compare BERT-based models to non-BERT models, identify current approaches researchers are taking when applying BERT to scientific and medical datasets, and identify the methods that have so far proven to be the most effective.

6.3 Implications

This review demonstrates that BERT is a robust language model that can be effectively applied to texts whose words are largely outside of BERT’s vocabulary. Additionally, further pretrained domain-specific BERT models are often more effective than standard BERT when applied to scientific and medical datasets. This includes both additional pretraining on custom corpora and precompiled domain-specific models such as BioBERT. As a result, it is suspected that other domains outside of science and medicine could also benefit from domain-specific BERT implementations, and more research should be conducted to determine whether or not this is the case. Lastly, BERT’s robustness to esoteric verbiage in non-English datasets suggests that other languages could also benefit from further research regarding applications of BERT (both further-pretrained and standard) to non-English datasets.

6.4 Future Work

More research needs be done to create additional language-specific BERT models. While multilingual BERT is effective on other Latin-based languages such as Spanish, BERT’s effectiveness at modeling datasets in languages with alternative writing systems needs to be explored. Languages like English, Spanish, and Russian are alphabetic, while character-based languages like Chinese and Japanese are ideographic. While all of the aforementioned languages were present during multilingual BERT's pretraining, it is possible that the combination of multiple writing systems may hinder multilingual BERT's performance relative to a system-specific or language-specific model.

Additionally, this review compared twenty-seven studies across multiple tasks and languages. Because the scope of this review was limited by the availability of research evaluating BERT's application to scientific and medical datasets, aggregating the data proved difficult. No more than seven studies modeled datasets using the same non-English BERT model. Similarly, no single task was addressed by more than thirteen studies. This made it difficult to obtain meaningful results across both tasks and languages. As such, research evaluating BERT's performance on scientific and medical datasets specific to a single task and a single language should be conducted in the future.

Multiple studies in this review also referenced the lack of large scientific and medical corpora that are available in languages other than English. The development and publication of these corpora would go a long way toward improving BERT's efficacy on scientific and medical datasets in non-English languages. Further research should also be conducted to see if applying domain-specific BERT models to additional domains is as effective as applying these models to the domains of science and medicine. Other domains such as law, finance, and art are rife with equally esoteric jargon and as such could potentially benefit from domain-specific BERT adaptations either via further pretraining from canonical corpora or via training from scratch on purely domain-specific corpora.

A recent publication [Gu et al., 2020] by Microsoft researchers suggests that pretraining BERT from scratch solely on domain-specific corpora (as opposed to canonical training followed by domain-specific training) can be an even more effective means of applying BERT to scientific and medical corpora. PubMedBERT, as the researchers have named their model, utilizes a vocabulary generated from the frequencies of words' occurrences in the PubMed corpus. PubMedBERT achieves a new SOTA on a ”wide range of biomedical applications” [Gu et al., 2020], and its creators believe that their findings may precipitate similar successes in other domains as well [Gu et al., 2020].

APPENDIX A

DATABASES SEARCHED

Table A.1: *Databases Searched*

Database	**Quantity**
Google Scholar—arXiv	200
arXiv (not published anywhere else)	77
Google Scholar—other sources	62
IEEE Xplore	49
Google Scholar—Association for Computational Linguistics	48
SpringerLink	28
Google Scholar—educational domains (.edu)	20
Google Scholar—Semantic Scholar	18
Web of Science	13
Association for Computing Machinery Digital Library	9
Google Scholar—Ceur Workshop	8
Google Scholar—public entity domains (.gov)	5
Computers and Applied Sciences Complete	4
EBSCO Academic Search Complete	4
Google Scholar—private entity domains (.com)	3
ScienceDirect	2
CiteSeerX	0
Dissertations and Theses Full Text (ProQuest)	0
JSTOR	0
ProQuest Linguistics Database	0
Total	**550**
Total (unique, after merging duplicates)	**425**

APPENDIX B

PUBLICATIONS BY TYPE

Table B.1: *Publications by Type*

Publication	Type	Quantity
ACM International Conference on Bioinformatics, Computational Biology and Health Informatics	Conference	1
AMIA Annual Symposium Proceedings	Symposium	1
BioNLP Workshop and Shared Task	Workshop	2
Ceur Workshop Proceedings of the CLEF eHealth Challenge	Workshop	1
Clinical Natural Language Processing Workshop	Workshop	2
Conference of Open Innovations Association (FRUCT)	Conference	1
Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)	Conference	1
Conference on Language Resources and Evaluation (LREC 2020)	Conference	1
IEEE International Conference on Bioinformatics and Biomedicine (BIBM)	Conference	1
International Conference on Asian Language Processing (IALP)	Conference	1
International Conference on Information Reuse and Integration for Data Science (IRI)	Conference	1
International Conference on Intelligent Computation Technology and Automation (ICICTA)	Conference	1
International Conference on Neural Information Processing	Conference	1
International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)	Congress	1
JMIR Medical Informatics	Journal	1
Joint International Information Technology and Artificial Intelligence Conference (ITAIC)	Conference	1
Journal of Biomedical Informatics	Journal	1
Journal of Data and Information Sciences	Journal	1
SIGBioMed Workshop on Biomedical Language Processing	Workshop	1
Social Media Mining for Health Applications (SMM4H) Workshop & Shared Task	Workshop	1
Text Analysis Conference (TAC) Track on Drug-Drug Interaction Extraction from Drug Labels Workshop	Workshop	1
Workshop on BioNLP Open Shared Tasks	Workshop	4

APPENDIX C

INCLUDED STUDIES

Table C.1: *Studies Selected for Inclusion in This Review*

[Akhtyamova, 2020]
[Alsentzer et al., 2019]
[Dai et al., 2019]
[Ding et al., 2019]
[García-Pablos et al., 2020]
[Hakala and Pyysalo, 2019]
[Lee et al., 2019]
[Li et al., 2019a]
[Li et al., 2019b]
[Li et al., 2020]
[Lin et al., 2019a]
[Lin et al., 2019b]
[Liu et al., 2019]
[Miftahutdinov et al., 2019]
[Peng et al., 2020]
[Peng et al., 2019]
[Phongwattana and Chan, 2019]
[Sänger et al., 2019]
[Song et al., 2019]
[Sun and Yang, 2019]
[Sung et al., 2019]
[Trieu et al., 2019]
[Wang et al., 2019]
[Xue et al., 2019]
[Yu et al., 2019]
[Zhang et al., 2019a]
[Zhang et al., 2019b]

REFERENCES

[Akhtyamova, 2020] Akhtyamova, L. (2020). Named Entity Recognition in Spanish Biomedical Literature: Short Review and Bert Model. In *2020 26th Conference of Open Innovations Association (FRUCT)*, pages 1–7.

[Alsentzer et al., 2019] Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., and McDermott, M. B. A. (2019). Publicly Available Clinical BERT Embeddings. *arXiv:1904.03323 [cs]*. arXiv: 1904.03323.

[Beltagy et al., 2019] Beltagy, I., Lo, K., and Cohan, A. (2019). SciBERT: A Pretrained Language Model for Scientific Text. *arXiv e-prints*, page arXiv:1903.10676.

[Bowman et al., 2015] Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. *arXiv e-prints*, page arXiv:1508.05326.

[Brown et al., 2020] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language Models are Few-Shot Learners. *arXiv e-prints*, page arXiv:2005.14165.

[Chelba et al., 2013] Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., and Robinson, T. (2013). One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling. *arXiv e-prints*, page arXiv:1312.3005.

[Cho et al., 2014] Cho, K., van Merrienboer, B., Bahdanau, D., and Bengio, Y. (2014). On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. *arXiv e-prints*, page arXiv:1409.1259.

[Cochran et al., 2020] Cochran, K., Cohn, C., Hastings, P., and Hughes, S. (2020). Transformer Models for Identifying Causal Relations in Students' Exploratory Essays [Unpublished manuscript]. College of Computing and Digital Media, DePaul University.

[Dai et al., 2019] Dai, Z., Wang, X., Ni, P., Li, Y., Li, G., and Bai, X. (2019). Named Entity Recognition Using BERT BiLSTM CRF for Chinese Electronic Health Records. In *2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–5.

[Deng et al., 2009] Deng, J., Dong, W., Socher, R., Li, L., Kai Li, and Li Fei-Fei (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

[Devlin et al., 2018] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv e-prints*, page arXiv:1810.04805.

[Ding et al., 2019] Ding, L., Liang, L., Tong, Y., Jiang, S., and Dong, B. (2019). A BERT-based Model for Drug-Drug Interaction Extraction from Drug Labels. In *Text Analysis Conference*.

[Gage, 1994] Gage, P. (1994). A New Algorithm for Data Compression. *C Users J.*, 12(2):23–38.

[García-Pablos et al., 2020] García-Pablos, A., Perez, N., and Cuadros, M. (2020). Sensitive Data Detection and Classification in Spanish Clinical Text: Experiments with BERT. *arXiv:2003.03106 [cs]*. arXiv: 2003.03106.

[Gers et al., 2000] Gers, F. A., Schmidhuber, J. A., and Cummins, F. A. (2000). Learning to Forget: Continual Prediction with LSTM. *Neural Comput.*, 12(10):2451–2471.

[Gu et al., 2020] Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. (2020). Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *arXiv e-prints*, page arXiv:2007.15779.

[Hakala and Pyysalo, 2019] Hakala, K. and Pyysalo, S. (2019). Biomedical Named Entity Recognition with Multilingual BERT. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 56–61.

[Heckman and Williams, 2011] Heckman, S. and Williams, L. (2011). A systematic literature review of actionable alert identification techniques for automated static code analysis. *Information and Software Technology*, 53(4):363 – 387. Special section: Software Engineering track of the 24th Annual Symposium on Applied Computing.

[Hochreiter and Schmidhuber, 1997] Hochreiter, S. and Schmidhuber, J. (1997). Long Short-term Memory. *Neural computation*, 9:1735–80.

[Howard and Ruder, 2018] Howard, J. and Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. *arXiv e-prints*, page arXiv:1801.06146.

[Hughes, 2019] Hughes, S. (2019). *Automatic Inference of Causal Reasoning Chains From Student Essays*. PhD thesis, DePaul University.

[Kitchenham, 2004] Kitchenham, B. (2004). Procedures for Performing Systematic Reviews. *Keele, UK, Keele Univ.*, 33.

[Kitchenham et al., 2009] Kitchenham, B., Pearl Brereton, O., Budgen, D., Turner, M., Bailey, J., and Linkman, S. (2009). Systematic literature reviews in software engineering – A systematic literature review. *Information and Software Technology*, 51(1):7–15.

[Lee et al., 2019] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *arXiv e-prints*, page arXiv:1901.08746.

[Lee et al., 2019] Lee, L.-H., Lu, Y., Chen, P.-H., Lee, P.-L., and Shyu, K.-K. (2019). NCUEE at MEDIQA 2019: Medical Text Inference Using Ensemble BERT-BiLSTM-Attention Model. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 528–532.

[Li et al., 2019a] Li, D., Xiong, Y., Hu, B., Du, H., Tang, B., and Chen, Q. (2019a). DX-HITSZ at BioNLP-OST 2019: Trigger Word Detection and Thematic Role Identification via BERT and Multitask Learning. *EMNLP-IJCNLP 2019*, page 72.

[Li et al., 2019b] Li, F., Jin, Y., Liu, W., Rawat, B. P. S., Cai, P., and Yu, H. (2019b). Fine-Tuning Bidirectional Encoder Representations From Transformers (BERT)–Based Models on Large-Scale Electronic Health Record Notes: An Empirical Study. *Journal of Medical Internet Research*, 21(9):N.PAG–N.PAG. 00003 Place: Toronto, Ontario Publisher: JMIR Publications Inc.

[Li et al., 2020] Li, X., Zhang, H., and Zhou, X.-H. (2020). Chinese clinical named entity recognition with variant neural structures based on BERT methods. *Journal of Biomedical Informatics*, 107:103422.

[Lin et al., 2019a] Lin, C., Huang, C., and Wu, C. (2019a). Using BERT to Process Chinese Ellipsis and Coreference in Clinic Dialogues. In *2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 414–418.

[Lin et al., 2019b] Lin, C., Miller, T., Dligach, D., Bethard, S., and Savova, G. (2019b). A BERT-based universal model for both within-and cross-sentence clinical temporal relation extraction. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 65–71.

[Liu et al., 2019] Liu, H., Perl, Y., and Geller, J. (2019). Transfer Learning from BERT to Support Insertion of New Concepts into SNOMED CT. In *AMIA Annual Symposium Proceedings*, volume 2019, page 1129. American Medical Informatics Association.

[Miftahutdinov et al., 2019] Miftahutdinov, Z., Alimova, I., and Tutubalina, E. (2019).

KFU NLP Team at SMM4H 2019 Tasks: Want to Extract Adverse Drugs Reactions from Tweets? BERT to The Rescue. In *Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task*, pages 52–57.

[Otter et al., 2019] Otter, D. W., Medina, J. R., and Kalita, J. K. (2019). A Survey of the Usages of Deep Learning in Natural Language Processing. *arXiv:1807.10854 [cs]*. arXiv: 1807.10854.

[Peng et al., 2020] Peng, Y., Chen, Q., and Lu, Z. (2020). An Empirical Study of Multi-Task Learning on BERT for Biomedical Text Mining. *arXiv:2005.02799 [cs]*. arXiv: 2005.02799.

[Peng et al., 2019] Peng, Y., Yan, S., and Lu, Z. (2019). Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. *arXiv:1906.05474 [cs]*. arXiv: 1906.05474.

[Pennington et al., 2014] Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

[Peters et al., 2018] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv e-prints*, page arXiv:1802.05365.

[Phongwattana and Chan, 2019] Phongwattana, T. and Chan, J. H. (2019). Development of Biomedical Corpus Enlargement Platform Using BERT for Bio-entity Recognition. In Gedeon, T., Wong, K. W., and Lee, M., editors, *Neural Information Processing*, Lecture Notes in Computer Science, pages 454–463, Cham. Springer International Publishing. 00000.

[Rajpurkar et al., 2016] Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. *arXiv e-prints*, page arXiv:1606.05250.

[Schuster and Nakajima, 2012] Schuster, M. and Nakajima, K. (2012). Japanese and Korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152.

[Song et al., 2019] Song, Z., Xie, Y., Huang, W., and Wang, H. (2019). Classification of Traditional Chinese Medicine Cases based on Character-level Bert and Deep Learning. In *2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, pages 1383–1387.

[Sun and Yang, 2019] Sun, C. and Yang, Z. (2019). Transfer Learning in Biomedical Named Entity Recognition: An Evaluation of BERT in the PharmaCoNER task. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 100–104.

[Sung et al., 2019] Sung, C., Dhamecha, T., Saha, S., Ma, T., Reddy, V., and Arora, R. (2019). Pre-Training BERT on Domain Resources for Short Answer Grading. 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 6073–6077.

[Sutskever et al., 2014] Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. *arXiv e-prints*, page arXiv:1409.3215.

[Sänger et al., 2019] Sänger, M., Weber, L., Kittner, M., and Leser, U. (2019). Classifying German Animal Experiment Summaries with Multi-lingual BERT at CLEF eHealth 2019 Task. *CLEF (Working Notes)*.

[Trieu et al., 2019] Trieu, H.-L., Nguyen, A.-K. D., Nguyen, N., Miwa, M., Takamura, H., and Ananiadou, S. (2019). Coreference Resolution in Full Text Articles with BERT and Syntax-based Mention Filtering. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 196–205.

[VanLehn, 2011] VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4):197–221.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention Is All You Need. *arXiv e-prints*, page arXiv:1706.03762.

[Wang et al., 2018] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2018). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. *arXiv e-prints*, page arXiv:1804.07461.

[Wang et al., 2019] Wang, S., Guo, Y., Wang, Y., Sun, H., and Huang, J. (2019). SMILES-BERT: Large Scale Unsupervised Pre-Training for Molecular Property Prediction. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB '19, pages 429–436, Niagara Falls, NY, USA. Association for Computing Machinery. 00000.

[Williams et al., 2018] Williams, A., Nangia, N., and Bowman, S. (2018). A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Compu-*

tational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1112–1122. Association for Computational Linguistics.

[Wu et al., 2016] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Łukasz Kaiser, Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *CoRR*, abs/1609.08144.

[Xue et al., 2019] Xue, K., Zhou, Y., Ma, Z., Ruan, T., Zhang, H., and He, P. (2019). Fine-tuning BERT for Joint Entity and Relation Extraction in Chinese Medical Text. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 892–897.

[Yu et al., 2019] Yu, G., Zhang, Z., Liu, H., and Ding, L. (2019). Masked Sentence Model Based on BERT for Move Recognition in Medical Scientific Abstracts. *Journal of Data and Information Science*, 4(4):42–55. WOS:000504871800004.

[Zhang et al., 2019a] Zhang, K., Liu, C., Duan, X., Zhou, L., Zhao, Y., and Zan, H. (2019a). BERT with Enhanced Layer for Assistant Diagnosis Based on Chinese Obstetric EMRs. In *2019 International Conference on Asian Language Processing (IALP)*, pages 384–389.

[Zhang et al., 2019b] Zhang, W., Jiang, S., Zhao, S., Hou, K., Liu, Y., and Zhang, L. (2019b). A BERT-BiLSTM-CRF Model for Chinese Electronic Medical Records Named Entity Recognition. In *2019 12th International Conference on Intelligent Computation Technology and Automation (ICICTA)*, pages 166–169.

[Zhu et al., 2015] Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books. *arXiv e-prints*, page arXiv:1506.06724.