Are we there yet? Exploring clinical domain knowledge of BERT models

Madhumita Sushil[1] and **Simon Šuster**[2], and **Walter Daelemans**[1]
[1] Computational Linguistics and Psycholinguistics Research Center (CLiPS),
University of Antwerp, Belgium
firstname.lastname@uantwerpen.be

[2] Faculty of Engineering and Information Technology, University of Melbourne
simon.suster@unimelb.edu.au

Abstract

We explore whether state-of-the-art BERT models encode sufficient domain knowledge to correctly perform domain-specific inference. Although BERT implementations such as BioBERT are better at domain-based reasoning than those trained on general-domain corpora, there is still a wide margin compared to human performance on these tasks. To bridge this gap, we explore whether supplementing textual domain knowledge in the medical NLI task: a) by further language model pretraining on the medical domain corpora, b) by means of lexical match algorithms such as the BM25 algorithm, c) by supplementing lexical retrieval with dependency relations, or d) by using a trained retriever module, can push this performance closer to that of humans. We do not find any significant difference between knowledge supplemented classification as opposed to the baseline BERT models, however. This is contrary to the results for evidence retrieval on other tasks such as open domain question answering (QA). By examining the retrieval output, we show that the methods fail due to unreliable knowledge retrieval for complex domain-specific reasoning. We conclude that the task of unsupervised text retrieval to bridge the gap in existing information to facilitate inference is more complex than what the state-of-the-art methods can solve, and warrants extensive research in the future.

1 Introduction

Transformers-based neural architectures (Vaswani et al., 2017) currently hold the state-of-the-art performance on several NLP tasks and domains. In the biomedical domain itself, there exist several versions of transformers-based BERT models (Devlin et al., 2019) that have been shown to be successful. However, an analysis of the availability of medical knowledge to these models is incomplete. To facilitate better understanding, in our research, we analyze a sample of errors made by BioBERT (v1.1) model (Lee et al., 2019a) on a clinical language inference task (Romanov and Shivade, 2018). We find that the errors related to domain knowledge-based reasoning, such as the knowledge of treatments administered for certain diseases, are dominant (40%).

To address this limitation, we analyze a broad range of state-of-the-art methods to integrate medical knowledge in BERT models from textual medical corpora. These methods have previously been shown to excel at evidence retrieval in the generic domain. The goal of our study is to understand whether these methods can be successfully applied for knowledge integration in the more complex setup of finding *missing medical information* for supporting *sentence-pair* inference.

We explore both *implicit* and *explicit* knowledge integration, where *implicit* refers to indirect access to this knowledge by further language model pretraining on medical corpora, and *explicit* knowledge integration refers to the setup where a relevant sentence from external corpora is appended to the premise to support inference. For explicit knowledge integration, as the baseline method, we make use of the traditional best match 25 (BM25) algorithm (Robertson and Zaragoza, 2009) for finding the most relevant sentence in the medical corpora. As a modification of this method, we additionally incorporate syntactic knowledge in the retrieval step. We do so by restricting the retrieved sentence to the one that contains at least one dependency relation between premise and hypothesis medical entities. In the third setup, instead of using BM25 scores and dependency paths, we train an end-to-end model to first find the most relevant text block from Wikipedia for a given instance, and then append it to the instance for classification.

In both knowledge integration setups, we do not see any significant performance difference due to access to additional knowledge. On inspecting the sentences retrieved by the BM25 and dependency

Proceedings of the BioNLP 2021 workshop, pages 41–53

relation-based methods, we find that these methods successfully shortlist sentences related to the topic, but it is difficult to then automatically rank the best candidate among the shortlisted options. This best candidate should fill the information gap between the sentence pairs to enable pairwise inference. We expect to overcome the ranking issue when we instead train an end-to-end model that learns to dynamically retrieve relevant supporting knowledge alongwith pairwise classification, as opposed to static heuristic-based retrieval. However, we find that although the blocks of text retrieved in the end-to-end setup provide medical context, they are often unrelated to the desired information and are insufficient for improving inference.

Although knowledge-integration methods are effective for evidence retrieval in open domain QA (Lee et al., 2019b), where the task is to retrieve a passage that mentions the correct entities, they are insufficient for the more complex task of augmenting missing information for pairwise domain knowledge-based reasoning in an unsupervised setup. Entity span-based supervision simplifies the problem statement in the first case, hence resulting in the documented success. However, the more realistic setup of retrieving the specific context that can fill the information gap between pairs of sentences without supervision is not yet solved.

2 Related work

Since the BERT models were found to be effective for a wide range of NLP tasks (Devlin et al., 2019), several efforts have been extended towards improving them by more efficient training strategies (Liu et al., 2019; Yang et al., 2019b; Sanh et al., 2019; Lan et al., 2019), training them for different domains (Beltagy et al., 2019; Lee et al., 2019a; Lee and Hsiang, 2019; Chalkidis et al., 2020; Gururangan et al., 2020) and languages (Devlin, 2018; de Vries et al., 2019; Le et al., 2020; Martin et al., 2020; Delobelle et al., 2020; Cañete et al., 2020). Within the clinical domain, different models include the BioBERT models pretrained on PubMed abstracts and PMC full-text articles (Lee et al., 2019a), SciBERT trained on scientific text (Beltagy et al., 2019), clinicalBERT models trained on patient notes from the MIMIC-III corpus (Johnson et al., 2016) (sometimes as a continuation of the BioBERT models) (Alsentzer et al., 2019), and BlueBERT models that also use Pubmed abstracts and MIMIC-III patient notes for training (Peng et al., 2019). These models hold promising performance for clinical language processing (Si et al., 2019; Lin et al., 2019) and have become a popular choice for several classification tasks that involve the medical data, spanning tasks such as literature search and question answering for assisting healthcare professionals (Jin et al., 2019; Wang et al., 2020; Möller et al., 2020), as well as patient outcome prediction such as diagnosis prediction (Franz et al., 2020; Rasmy et al., 2020). Despite being a popular choice, little is known about the medical knowledge of these models and their limitations when in-depth domain knowledge is required for correctly solving a task.

Much prior research has explored augmentation of background knowledge in neural models to make them more effective for downstream tasks. Most common approaches include adapting entity embeddings learned by models such as BERT by providing additional knowledge from different ontologies that define relations between entities. This can be done either by using templates to convert the relations to text before finetuning embeddings (Weissenborn et al., 2017; Lauscher et al., 2020; Chen et al., 2020), by combining relational information from knowledge graphs with text embeddings (Mihaylov and Frank, 2018; Chen et al., 2018; Zhang et al., 2019; Yang et al., 2019a; Liu et al., 2020), or by jointly learning knowledge graph and textual embeddings (Peters et al., 2019; Feng et al., 2020). These ontologies are either generic like WordNet (Miller, 1995), ConceptNet (Liu and Singh, 2004), and Wikidata (Vrandečić and Krötzsch, 2014), or more specific to a particular domain like the UMLS (Bodenreider, 2004). An advantage of using ontologies is that the semantics of entities gets encoded in the learned representations, thereby enhancing their effectiveness. However, they are expensive to construct and either are incomplete, or do not exist for specialized domains. Methods that make use of textual corpora for background knowledge integration are therefore more easily transferable to other domains. Talmor et al. (2020) have shown earlier that having explicit access to external information can often improve reasoning skills of the state-of-the-art models, which we investigate further.

Use of TF-IDF (Ullman, 2011) and BM25 scores has been frequently explored for evidence retrieval from Wikipedia for open domain QA (Chen et al., 2017; Wang et al., 2018; Glass et al., 2020). An-

other popular approach includes representation similarity-based evidence retrieval (Lee et al., 2018; Das et al., 2019). Recently, joint training of retriever for span identification and pretraining language models have also been analyzed by Hu et al. (2019); Lee et al. (2019b); Guu et al. (2020). Although the methods extensively explore QA, this line of work has not been explored much for language inference, especially in specialized domains.

Existing studies for augmenting medical knowledge for clinical language inference are limited to the use of UMLS knowledge graph embeddings (Sharma et al., 2019), interaction weighting between premise and hypothesis based on distance in the UMLS (Chopra et al., 2019), augmenting clinical concept definitions during representation learning (Lu et al., 2019) and adding domain knowledge by means of pretraining existing models further on different in-domain corpora and closely related tasks (Romanov and Shivade, 2018; Lee et al., 2019a; Alsentzer et al., 2019; Chopra et al., 2019). The closest work to ours is the contemporary work by He et al. (2020) that shows improvements when knowledge from Wikipedia is implicitly integrated by training BERT masked language models to predict disease names and their aspects (such as symptoms, treatments) given the corresponding context. In our work, we instead explore whether we can augment domain knowledge by dynamically fetching relevant context in an unsupervised manner to improve medical language inference.

3 Medical language inference

In medical language inference, given a pair of sentences, the goal is to describe a logical relation between them. We make use of the MedNLI dataset (Romanov and Shivade, 2018), where the premise is a sentence borrowed from patient notes in the MIMIC-III dataset (Johnson et al., 2016), and the hypothesis is written by medical experts such that the premise either entails or contradicts the hypothesis, or their relation cannot be established (neutral). Entailment refers to whether the meaning of the second sentence, also known as the 'hypothesis', is already contained in the first sentence called the 'premise'. We explore whether the BioBERT v(1.1) model encodes sufficient medical knowledge for this task. In the same manner as Peng et al. (2019), we model this task as a sentence pair classification task, where the final pooled BERT [CLS] representations of the premise and the hypothesis are processed through a dense neural layer to classify the correct class. We then perform manual analysis on a subset of 50 incorrectly classified instances in the development set to understand the type of errors made by the model. We eliminate ambiguity in the cause of errors by using an adversarial evaluation, where we modify an instance according to a potential cause of error, and monitor whether the output changes accordingly. In this manner, we obtain the distribution of errors presented in Table 2 and discussed in Section 5.1.

4 Medical knowledge augmentation

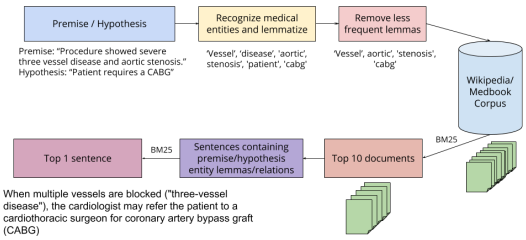
4.1 External medical corpora

Different versions of BERT that exist for biomedical tasks are either trained on journal abstracts and articles, or on patient notes. These articles and notes are written by and for an audience with an advanced level of domain knowledge. Fundamental domain-specific information, such as an understanding of domain terminology, commonly accepted clinical practices for specific medical conditions, human physiology and anatomy, etc. is often also required for clinical language understanding. We hypothesize that access to such fundamental domain knowledge during model training would complement training on more advanced information. To explore this, we create two corpora — one containing only the medical subset of Wikipedia (Wikimed), and one with contents of a popular medical textbook (Medbook). The Wikimed subset is parsed from the HTML sources of the medical Wikipedia dataset used in the Android app by the Kiwix team[1]. The medical subset of Wikipedia contains about 40 million tokens, and the medical textbook corpus contains nearly 3.6 million tokens.

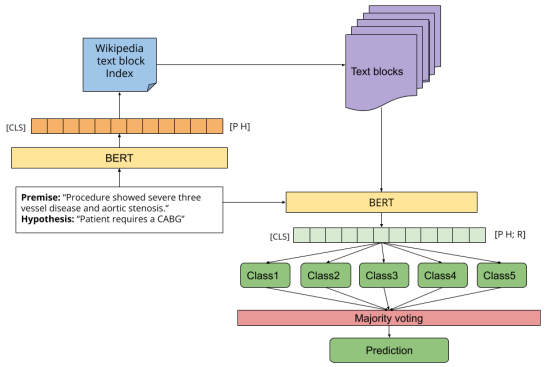
4.2 Implicit knowledge integration

Starting from an existing BioBERT checkpoint that is already pretrained on a combination of Google books, Wikipedia, biomedical abstracts and journal articles (Lee et al., 2019a), we continue to train BERT language models on the Medbook and the Wikimed corpora. Our goal is to explore whether further training on corpora that contain fundamental domain knowledge can implicitly improve medical knowledge-based reasoning in the medical language inference task. Since Wikimed is the

[1] https://play.google.com/store/apps/details?id=org.kiwix.kiwixcustomwikimed&hl=en_US&gl=US



(a) Lexical knowledge retrieval using BM25 score between a query formulated from premise-hypothesis pair and sentences in the external corpora. These sentences are restricted to either those that mention a premise *and* a hypothesis medical entity term, or hold a dependency relation between them.



(b) Relevant knowledge retrieval in an end-to-end manner by training weights that compute similarity between sentences in an external corpus and a premise(P)-hypothesis(H) query during classification.

Figure 1: Explicit domain knowledge integration for the MedNLI task.

medical-only subset of Wikipedia, it was also included in the first phase of training of BERT models. We do not expect to see a significant difference in the classification performance here due to this reason. However, since the Medbook corpus is quite different from other corpora used earlier, we expect bigger differences in classification results.

4.3 Explicit knowledge integration

We explore methods to explicitly augment medical knowledge to the instances in the MedNLI dataset by retrieving and appending relevant text blocks from either the Wikimed corpus or the Medbook corpus before processing it through our BERT models for finetuning, as described next. We illustrate the methods pipeline in Figure 1.

4.3.1 Lexical retrieval

We first explore the use of TF-IDF based techniques for retrieving evidence from external textual corpora to support inference. Although these methods are fairly simple, they have been shown to be effective for several open domain QA tasks (Lee et al., 2019b). Our goal is to investigate whether these simple methods are also effective at more complex information retrieval in our setup.

To this end, we construct a query from premise and hypothesis by retaining only the lemmas that are a part of infrequent medical entities, and then use the best match 25 (BM25) algorithm (Robertson and Zaragoza, 2009) to find the most relevant sentences. As the first step, we recognize premise and hypothesis medical entities with the help of Scispacy (Neumann et al., 2019). We lemmatize these entities and retain only those lemmas that occur less than a thousand times in the external corpus[2]. These lemmas jointly form the query. We first rank the documents in the external corpora according to their BM25 scores to retain the top 10 documents. The query is then used again to find the best matching sentences from these documents.

Due to the manner in which the MedNLI data has been annotated, premise is longer and more varied than the hypothesis. Hence, premise entities often dominate the BM25 retrieval at the cost of hypothesis entities. To overcome this, we prune the retrieved sentences if they do not mention at least one premise and one hypothesis entity lemma.

The highest ranking sentence retrieved in this manner is then appended[3] to the premise before classification. If none of the sentences satisfy either the constraint or the threshold score, then the use of explicit knowledge is skipped.

4.3.2 Lexical and syntactic retrieval

In our previous setup, we add an entity-presence constraint to ensure that the retrieved sentence is about both the premise and the hypothesis. In order to ensure that the retrieved knowledge also establishes an explicit relation between the two, we modify the previous approach to rank sentences based on dependency paths between premise and hypothesis lemmas. In this setup, we find the top documents in the same manner as earlier. Once the top documents are found, we restrict to the set of sentences in these documents that have a dependency relation between a premise and a hypothesis lemma. Once we have established the set of sentences that hold this relation, we rank them either

[2]The threshold was decided based on preliminary results on the development set, where retaining less frequent lemmas provides more specific matches.

[3]Separated by a space.

using the minimum dependency path length, or using the BM25 score between the query and a sentence. The sentence with the highest score above the threshold is then appended to the instance in the same manner as described earlier.

4.3.3 Joint retrieval and classification model

By using lexical and syntactic approaches that we have discussed earlier, we ensure that the candidate and the retrieved sentences would be related to both the premise and the hypothesis. However, when we are confronted with a high number of relevant candidate sentences, shortlisting one sentences becomes challenging. Adding multiple sentences is also infeasible due to the limited input sequence length in BERT models. In order to overcome this challenge, in our third setup, we instead train an end-to-end model, where the weights of the retriever are updated along with classification. Hence, the retriever learns to select the sentence that provides information that can improve classification. This approach has been previously shown to be quite successful in open domain QA via span identification (Lee et al., 2019b) and in language model pretraining (Guu et al., 2020), since it provides access to a wider evidence space compared to the limited number of retrieved blocks when using lexical approaches. However, the use of such an end-to-end retriever has not been explored for augmenting knowledge from textual corpora to support reasoning in NLI tasks. Since we do not have data annotated specifically for retrieval of supporting evidence for NLI tasks, training the retriever becomes much more complex compared to span identification. However, given the success of the end-to-end approaches earlier, we are interested in investigating its feasibility for our setup and we build upon existing methods for this.

Retriever pretraining: We reuse the pretrained retrieval model shared by Lee et al. (2019b), trained in an inverse cloze task (ICT) setup on complete Wikipedia, for our experiments. In this setup, a sentence in Wikipedia is treated as the query, and the retriever is trained to retrieve its context[4] in the original text. This retrieval is performed by computing a weighted dot product between the pooled BERT [CLS] embeddings of the query and the text block. In 10% of the cases, the query is not removed from the context to ensure that the model learns to retrieve lexical as well as semantic matches. Although it is trained on entire Wikipedia instead of only a subset, we reuse it due to resource constraints for retraining the retriever. Since the medical portion of Wikipedia is only a subset of this data, we expect to still be able to retrieve the sentences relevant for the MedNLI task.

End-to-end-classification: In an end-to-end setup, the retriever module first returns the k[5] most similar blocks of text given a BERT-encoded premise and hypothesis pair, in the same manner as described earlier. We add these k retrieved blocks to the input along with the premise and the hypothesis to obtain k inputs corresponding to each instance. We then encode these inputs with BERT to obtain k different [CLS] representations. All of these k [CLS] representations are then individually used for classification by adding a dense layer on the top in the finetuning phase. In this manner, we obtain k different outputs for a given instance. We then aggregate these k outputs together by retaining the most frequent output among the k options. We also experimented with average pooling and selecting the most peaked softmax output distribution, but majority pooling provided more promising results on the development set.

Classification loss: We use the categorical cross entropy loss (Murphy, 2012). The gradients are backpropagated jointly to both the classifier and the weights used to compute the similarity between the query and the blocks of Wikipedia text.

Retriever loss: In the span identification setup developed by Lee et al. (2019b), mention of the correct entity in the text provides the retriever with an explicit feedback. This makes their training easier compared to our setup where we do not have this supervised signal. To make the training more feasible, we experiment with an additional retrieval loss. This loss quantifies the difference between the model performance with and without the retrieved text block, and uses this difference to improve the retriever. The objective of this loss is to reward the model when it is better if a retrieved text block is used as opposed to when only the premise and the hypothesis are used for inference. We define this loss in terms of pairwise retrieval loss, i.e.,

$$R = max(0, m - (L_{(P,H)} - L_{(P,H,R)})),$$

where R is the retrieval loss, $L_{(P,H)}$ is the categorical cross entropy loss without using the retrieved

[4]Blocks of at most 288 wordpiece tokens (Wu et al., 2016)

[5]We use $k = 5$ in our experiments

text block, and $L_{(P,H,R)}$ is the categorical cross entropy loss after adding the retrieved text block to the given instance, and m is the margin value that we treat as a hyperparameter. We use $m = 0.1$ based on the results on the development set. To explain this loss, we consider three different cases:

1. The model performs equivalently with and without the retrieved text block: In this case, the model ignores the retriever and optimizes for classification without it. This is undesirable, and we set the retriever loss to the margin value, which refers to the minimum desired difference between the two sets of losses.

2. The model is worse after adding the retrieved text block: This behavior is again undesirable since the goal of retrieval is to improve the model. Hence, along with the margin, we also add the difference between the two losses to compute the retrieval loss.

3. The model improves after adding the retrieved text block: If the model becomes better due to retrieval, it could either be better by chance (when the difference is lower than the minimum margin), or the difference could be substantial. In the first case, we quantify the retrieval loss as the margin value. The latter behavior is the desired behavior of the model, and we set the retrieval loss to be zero.

Here, the final loss function is computed as the sum of the classification loss and the retrieval loss.

5 Results and Discussion

5.1 Availability of domain knowledge

In the top section of Table 1, we present the results when we finetune BERT models for medical language inference. Here we can see that the BERT model which has been trained on in-domain Pubmed abstracts for the largest number of optimization steps is consistently the best on both development and test sets. As expected based on prior research, all other models trained on in-domain data are also significantly better than the BERT models that are not trained on in-domain data.

We investigate the errors made by the best model, BioBert (v1.1). As discussed in Section 3, in Table 2, we present the distribution of the first 50 errors made on the development set of the MedNLI dataset. Examples of these errors are illustrated in Table 3. Although we present the distribution of errors for one specific run here, we also analyzed this distribution across 3 different runs of the model. We found that the average pairwise Cohen's kappa agreement (McHugh, 2012) between the predictions on the development set across 3 different runs is 0.9, and the distribution of errors across these runs is comparable. In Table 2, we can see that 40% of the errors happen due to insufficient domain information. Some of these errors happen because of missing factual domain knowledge, some lack advance reasoning based on factual domain knowledge, and some are incorrect potentially because of model biases due to limited size of the training dataset, such as assumption that a certain treatment is always administered for a specific condition, although the treatment might be more diverse. This highlights the potential to improve the BioBERT model by providing access to additional fundamental domain information.

Other dominant category of errors are related to spurious correlations, numeric inference, negation, and temporal reasoning. These categories are important for understanding patient condition in medical notes, since test results are often expressed in a numeric manner, patient conditions are often hedged and negated, and patient information is usually longitudinal in nature. We limit the focus of this work to the more frequent error category of integrating domain information.

5.2 Domain knowledge integration

In Table 1, we see marginal improvements on the test set between the BioBERT (v1.1) models with and without additional domain knowledge — both when the integration is done implicitly via additional language model pretraining, and when relevant sentences are retrieved using lexical and syntactic methods. Knowledge integration from the Medbook corpus — both implicit and explicit, does not show any improvement in the results. Despite marginal improvements using the Wikimed corpus, a lack of consistent pattern across both development and test sets suggests a random effect rather than significant differences. When we train an end-to-end retrieval model instead of further language modeling or pre-selecting the most relevant sentence, we again see a marginal improvement on the test set. However, this improvement is again not visible on the development set. Furthermore, we see that the pairwise loss for more aggressive

Model	MedNLI (% Acc.)	
	Dev	Test
BERT-base-uncased	82.1	77.8
BERT-base-cased	79.9	78.8
BERT-base-cased + PMC + PubMed (BioBERT v1.0)	84.3	82.5
BERT-base-cased + Pubmed 1M (BioBERT v1.1)	84.8	82.9
SciBERT-base-uncased (SciBERT vocab)	81.5	82.2
He et al. (2020): BioBERT v1.1 + disease	NA	82.2
Sharma et al. (2019)	NA	79.0
BERT-base-cased + Pubmed 1M (BioBERT v1.1)	84.8	82.9
BioBERT v1.1 + Wikimed MLM	84.2	**83.3**
BioBERT v1.1 + Medbook MLM	83.2	80.1
BioBERT v1.1 + Wikimed (lexical)	84.3	**83.2**
BioBERT v1.1 + Medbook (lexical)	83.8	82.6
BioBERT v1.1 + Wikimed (lexical+syntactic)	83.9	**83.1**
BioBERT v1.1 + Medbook (lexical+syntactic)	83.8	82.5
BERT-base-uncased (Wikipedia+BooksCorpus)	82.1	77.8
BERT-base-uncased + trained Wiki retriever	79.4	78.5
BERT-base-uncased + trained Wiki retriever + retrieval loss	79.1	77.9

Table 1: Classification accuracy of BERT models and explicit and implicit domain knowledge integration methods on MedNLI development and test sets. MLM here refers to masked language modeling.

Error type	Count
Insufficient domain knowledge	20
Spurious correlations / dataset bias	6
Difficult instance	5
Incorrect numeric inference	4
Incorrect negation	3
Incorrect tense resolution	2
Incorrect temporal sequence inference	2
Lexical (P,H) overlap trick	2
Modifier ignored	2
Incorrect abbreviation understanding	2
Insufficient commonsense knowledge	1

Table 2: Analysis of the first 50 errors of the BioBERT (v1.1) model on the MedNLI development set.

retriever training along with the classification cross-entropy loss does not have any significant impact. Despite this additional signal, the classifier continues to learn the task by ignoring the retrieved context, thus indicating that the penalty for incorrect retrieval is still not aggressive enough.

Our joint models use the complete Wikipedia as the source of knowledge, and the improvement patterns here are consistent with using the Wikimed corpus both implicitly and explicitly, but contrary to using the Medbook corpus. This suggests that Wikipedia, both complete and the medical-only subset, functions as a better source of information for the MedNLI task as compared to the medical textbook that contains more fundamental domain information. We believe that the difference in results of the two corpora emerges from a difference in their sizes, since the medical subset of Wikipedia is 10 times in size compared to the textbook corpus. We could not scale the Medbook corpus to larger sizes due to copyright limitations.

When we analyze the retrieved text blocks for one example in the development set and compare it to the gold standard retrieval by humans (presented in Table 4), we see that none of the retrieval algorithms are capable of finding the desired missing information to improve semantic inference. Although the 'lexical + syntactic' retriever finds a sentence related to the topic as well as to the premise and the hypothesis, it doesn't bridge the knowledge gap for correct inference. Moreover, the end-to-end model with a trained retriever retrieves text block that is unrelated to the topic, although in the medical genre.

Hence, we find that none of the explored methods provide better access to medical information for domain knowledge-based reasoning, although the desired factual information is present in these external corpora. One reason why we do not see further improvements on the BioBERT (v1.1) model

Error type	Example
Insufficient domain knowledge	**P:** ... she was treated with **Benadryl** ... **H:** Patient has had an **allergic reaction** ~~Entailment~~ Neutral
Spurious correlations / dataset bias	**P:** She spoke with her **oncology team** ... **H:** The patient has **cancer**. ~~Neutral~~ Entailment
Incorrect numeric inference	**P:** ... an **ejection fraction of 69%** with normal wall motion. **H:** patient has **normal cardiac output** ~~Entailment~~ Contradiction
Incorrect negation resolution	**P:** ... **no** identified sepsis risk factors. **H:** ... has **multiple** risk factors for sepsis ~~Contradiction~~ Entailment
Incorrect tense resolution	**P:** ... he **had a CT of the chest** and CTA of his coronary arteries ... **H:** patient **will** go for **coronary angiography** ~~Neutral~~ Entailment
Incorrect temporal inference	**P:** ... biopsy ... showed signs of rejection ... **subsequently did well**. **H:** The patient **had transplant failure** ~~Contradiction~~ Entailment
Lexical (P, H) overlap trick	**P: Pt denies** any **recent** chills ... **H:** The **patient denies recent** illness ~~Neutral~~ Entailment
Modifier ignored	**P:** Left common femoral dorsalis **pedis** bypass graft. **H:** Patient has **CAD** ~~Neutral~~ Entailment
Incorrect abbreviation understanding	**P:** Her ... **PO** intake have been normal. **H:** She has been **NPO** since midnigh ~~Contradiction~~ Neutral
Insufficient commonsense knowledge	**P:** ... status post high speed motor **vehicle crash** ... **H:** Patient has recent **trauma** ~~Entailment~~ Neutral

Table 3: One example of each category of errors made by the BioBERT (v1.1) model on the MedNLI development set. $a \not\to b$ refers to the fact that class a is the gold class, but the model predicts class b instead.

(that is a very strong baseline), despite the success of these methods in other tasks and domains, could be the complexity of the research question. Retrieval of relevant information for language inference demands a delicate balance between selecting a sentence that provides sufficient supporting information related to the given topic and instance to improve inference, and yet that is neither redundant nor superfluous. As we show in our results, in a limited computation setting as ours, current state-of-the-art methods are not capable of striking this balance in unsupervised setups and result in unreliable knowledge augmentation. He et al. (2020) also report similar results on the same task using the same BioBERT model. These results suggest that we either need more computation power to train these models for longer time to enable convergence, or we need to create large annotated corpora for retrieving missing facts to enable better performance of these algorithms with limited computation power. We need to direct our efforts towards investigating advanced evidence retrieval and knowledge integration setups such as this to improve knowledge-based reasoning of the current state-of-the-art models.

Method	Text
Example	**P**: Infusion stopped and she was treated with Benadryl 50 mg x 1, prednisone 40 mg x 1, ativan 1 mg. **H**: Patient has had an allergic reaction
Gold retrieval	Benadryl is a brand name for a number of different antihistamine medications used to stop allergies, including diphenhydramine, acrivastine and cetirizine.
Lexical retrieval	None
Lexical + syntactic retrieval	Prednisone is used for many different autoimmune diseases and inflammatory conditions, including asthma, COPD, CIDP, rheumatic disorders, allergic disorders, ulcerative colitis and Crohn's disease, granulomatosis with polyangiitis, adrenocortical insufficiency, hypercalcemia due to cancer, thyroiditis, laryngitis, severe tuberculosis, hives, lipid pneumonitis, pericarditis, multiple sclerosis, nephrotic syndrome, sarcoidosis, to relieve the effects of shingles, lupus, myasthenia gravis, poison oak exposure, Méniére's disease, autoimmune hepatitis, giant-cell arteritis, the Herxheimer reaction that is common during the treatment of syphilis, Duchenne muscular dystrophy, uveitis, and as part of a drug regimen to prevent rejection after organ transplant.
Trained Wiki retriever + retrieval loss	Gemeprost (16, 16-dimethyl-trans-delta2 PGE methyl ester) is an analogue of prostaglandin E. It is used as a treatment for obstetric bleeding. It is used with mifepristone to terminate pregnancy up to 24 weeks gestation. Vaginal bleeding, cramps, nausea, vomiting, loose stools or diarrhea, headache, muscle weakness; dizziness; flushing; chills; backache; dyspnoea; chest pain; palpitations and mild pyrexia. Rare: Uterine rupture, severe hypotension, coronary spasms with subsequent myocardial infarctions. **...**

Table 4: Text blocks retrieved by different methods from the (medical) Wikipedia corpus for one example in the development set that requires further domain knowledge for correct inference. Gold retrieval mentioned here is a manually retrieved sentence from Wikipedia, in presence of which the model corrects its output.

6 Conclusions and Future Work

On investigating the error categories of BioBERT (v1.1) models on the clinical language understanding task, we find that despite having a strong performance, the models still make several mistakes on examples that require medical domain knowledge. To this end, we explored multiple methods to improve access of these models to medical domain knowledge by implicit and explicit knowledge retrieval and augmentation. However, we see that these extensions do not show significant improvement on the test sets. We conclude that state-of-the-art solutions lead to unreliable knowledge augmentation for language inference, as is shown by a detailed analysis in our work. Future research should concentrate efforts towards developing methods to augment fundamental domain knowledge from textual corpora to solve the problem of advanced knowledge-based reasoning in these domains.

Acknowledgements

This research was carried out within the Accumulate strategic basic research project, funded by the government agency Flanders Innovation & Entrepreneurship (VLAIO) [grant number 150056]. It also received funding from the Flemish Government (AI Research Program). This research was conducted following an internship at Google Research in Zürich. The experience gained during the internship was instrumental in the research. We would like to thank everyone in the team, and particularly André Susano Pinto for several discussions related to BERT and Tensorflow, for exchange of ideas, and for feedback on the draft. We would like to additionally thank all the anonymous reviewers whose useful comments have ensured a better version of the paper.

References

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

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *to appear in PML4DC at ICLR 2020*.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Mingda Chen, Zewei Chu, Karl Stratos, and Kevin Gimpel. 2020. Mining knowledge for natural language inference from wikipedia categories. *arXiv preprint arXiv:2010.01239*.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Diana Inkpen, and Si Wei. 2018. Neural natural language inference models enhanced with external knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2406–2417, Melbourne, Australia. Association for Computational Linguistics.

Sahil Chopra, Ankita Gupta, and Anupama Kaushik. 2019. MSIT_SRIB at MEDIQA 2019: Knowledge directed multi-task framework for natural language inference in clinical domain. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 488–492, Florence, Italy. Association for Computational Linguistics.

Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, and Andrew McCallum. 2019. Multi-step retriever-reader interaction for scalable open-domain question answering. In *International Conference on Learning Representations*.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. Bertje: A dutch bert model. *arXiv preprint arXiv:1912.09582*.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. Robbert: a dutch roberta-based language model. *arXiv preprint arXiv:2001.06286*.

Jacob Devlin. 2018. Multilingual bert readme document.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yanlin Feng, Xinyue Chen, Bill Yuchen Lin, Peifeng Wang, Jun Yan, and Xiang Ren. 2020. Scalable multi-hop relational reasoning for knowledge-aware question answering.

Leopold Franz, Yash Raj Shrestha, and Bibek Paudel. 2020. A deep learning pipeline for patient diagnosis prediction using electronic health records. *arXiv preprint arXiv:2006.16926*.

Michael Glass, Alfio Gliozzo, Rishav Chakravarti, Anthony Ferritto, Lin Pan, G P Shrivatsa Bhargav, Dinesh Garg, and Avi Sil. 2020. Span selection pre-training for question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2773–2782, Online. Association for Computational Linguistics.

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

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*.

Yun He, Ziwei Zhu, Yin Zhang, Qin Chen, and James Caverlee. 2020. Infusing disease knowledge into bert for health question answering, medical inference and disease name recognition. *arXiv preprint arXiv:2010.03746*.

Minghao Hu, Yuxing Peng, Zhen Huang, and Dongsheng Li. 2019. Retrieve, read, rerank: Towards end-to-end multi-document reading comprehension.

In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2285–2295, Florence, Italy. Association for Computational Linguistics.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. 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 2567–2577, Hong Kong, China. Association for Computational Linguistics.

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

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

Anne Lauscher, Olga Majewska, Leonardo FR Ribeiro, Iryna Gurevych, Nikolai Rozanov, and Goran Glavaš. 2020. Common sense or world knowledge? investigating adapter-based knowledge injection into pretrained transformers. *arXiv preprint arXiv:2005.11787*.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Jieh-Sheng Lee and Jieh Hsiang. 2019. Patentbert: Patent classification with fine-tuning a pre-trained bert model. *arXiv preprint arXiv:1906.02124*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019a. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Jinhyuk Lee, Seongjun Yun, Hyunjae Kim, Miyoung Ko, and Jaewoo Kang. 2018. Ranking paragraphs for improving answer recall in open-domain question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 565–569, Brussels, Belgium. Association for Computational Linguistics.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019b. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.

Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. 2019. 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.

Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-BERT: enabling language representation with knowledge graph. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 2901–2908. AAAI Press.

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

M. Lu, Y. Fang, F. Yan, and M. Li. 2019. Incorporating domain knowledge into natural language inference on clinical texts. *IEEE Access*, 7:57623–57632.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica*, 22(3):276–282.

Todor Mihaylov and Anette Frank. 2018. Knowledgeable reader: Enhancing cloze-style reading comprehension with external commonsense knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 821–832, Melbourne, Australia. Association for Computational Linguistics.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Timo Möller, Anthony Reina, Raghavan Jayakumar, and Lawrence Livermore. 2020. Covid-qa: A question & answering dataset for covid-19.

Kevin P Murphy. 2012. *Machine learning: a probabilistic perspective*. MIT press.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

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

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. 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 43–54, Hong Kong, China. Association for Computational Linguistics.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2020. Med-bert: pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction. *arXiv preprint arXiv:2005.12833*.

Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596, Brussels, Belgium. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Soumya Sharma, Bishal Santra, Abhik Jana, Santosh Tokala, Niloy Ganguly, and Pawan Goyal. 2019. Incorporating domain knowledge into medical NLI using knowledge graphs. 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 6092–6097, Hong Kong, China. Association for Computational Linguistics.

Yuqi Si, Jingqi Wang, Hua Xu, and Kirk Roberts. 2019. Enhancing clinical concept extraction with contextual embeddings. *Journal of the American Medical Informatics Association*, 26(11):1297–1304.

Alon Talmor, Oyvind Tafjord, Peter Clark, Yoav Goldberg, and Jonathan Berant. 2020. Teaching pre-trained models to systematically reason over implicit knowledge. In *Advances in Neural Information Processing Systems 34*.

Jeffrey Ullman. 2011. *Mining of massive datasets*. Cambridge University Press.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, et al. 2020. Cord-19: The covid-19 open research dataset. *ArXiv*.

Shuohang Wang, Mo Yu, Jing Jiang, Wei Zhang, Xiaoxiao Guo, Shiyu Chang, Zhiguo Wang, Tim Klinger, Gerald Tesauro, and Murray Campbell. 2018. Evidence aggregation for answer re-ranking in open-domain question answering. In *International Conference on Learning Representations*.

Dirk Weissenborn, Tomáš Kočiskỳ, and Chris Dyer. 2017. Dynamic integration of background knowledge in neural nlu systems. *arXiv preprint arXiv:1706.02596*.

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

An Yang, Quan Wang, Jing Liu, Kai Liu, Yajuan Lyu, Hua Wu, Qiaoqiao She, and Sujian Li. 2019a. Enhancing pre-trained language representations with rich knowledge for machine reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2346–2357.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019b. XLNet: Generalized autoregressive pretraining for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 5753–5763. Curran Associates, Inc.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meet-*

ing of the Association for Computational Linguistics, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.