Ontology Mediated Attention for Biomedical Documents

Aditya Dave
Univ. of S. California
daveadit@usc.edu

Alexander Chang
Univ. of S. California
changam@usc.edu

Robert Steele
Univ. of S. California
rjsteele@usc.edu

Slava Zinevich
Univ. of S. California
zinevich@usc.edu

Abstract

The transformer (Vaswani et al., 2017) has improved the state-of-the-art across numerous sequence modeling tasks. However, its effectiveness comes at the expense of a quadratic computational and memory complexity with respect to the sequence length, hindering its adoption. Recently, researchers have directly addressed the Transformer's limitation by designing lower-complexity alternatives such as the Longformer (Beltagy et al., 2020), Reformer (Kitaev et al., 2020), Linformer (Wang et al., 2020), and Performer (Choromanski et al., 2020). In the biomedical domain language models, Clinical-Longformer and Clinical-BigBird (Li et al., 2022) which are based on Longformer (Beltagy et al., 2020) and BigBird (Zaheer et al., 2020) respectively have proven to outperform Clinical-BioBERT (Alsentzer et al., 2019) and other short-sequence-transformers in all downstream tasks. Our aim is to improve upon these sparse models in the Biomedical domain by using ontologies to inform the sparse attention. In this study we explore the utility of this ontology-mediated attention.

1 Introduction

The topic of solving seq2seq learning tasks using deep learning networks underwent a resurgence in 2014 when (Sutskever et al., 2014) published their breakthrough paper. The task discussed was machine translation, a problem yet unsolved at the time. They used LSTM networks, and those became the state-of-the-art. However, 3 years later a new architecture achieved better performance, using the Transformer model (Vaswani et al., 2017). Unlike LSTM, which was a combination of an RNN and a long-term context learner, the transformer used a mechanism of self-attention for the network to understand the context of words within a sentence. The concept of self-attention quickly popularized and the Transformer was soon superseded by the next generation model, a bi-directional architecture under the name BERT (Devlin et al., 2018).

There are two main issues that are present in the original proposed transformer based model: 1) the maximum length of the sequence was restricted to a sequence length of 512 tokens, and 2) the full self-attention mechanism of the transformer architecture leads to quadratic computational and memory complexity. There are many proposed methods and models to overcome these issues, some more successful than others. Such models are generally referred to as x-formers, built on top of transformers. The general idea was to limit the attention to a sub-scope of context, and various implementations arose such as the TransformerXL (Dai et al., 2019) which aimed at recurrence of local contexts, the sparse-transformer (Child et al., 2019) which was superlinear but still more efficient, and other techniques (Wu et al., 2019) (Ye et al., 2019).
The current state-of-the-art models Longformer (Beltagy et al., 2020) and BigBird (Zaheer et al., 2020) introduce sparsity into the attention matrix to decrease the time complexity and also increase the allowed sequence length to 4096 tokens.

When looking at specific domains another issue arises. As these models rely heavily on word distributions and derived meaning, tasks based on lingual sub-domains might see lesser performance than general-purpose tasks. With the transition of health services to the Electronic Health Record (EHR), a vast amount of data is constantly being generated for tracking and evaluating patients. However, much of the relevant information for patients is interweaved in unstructured data sources such as text and image. For instance, it has been shown that curated fields such as ICD codes are prone to error, decreasing their value to clinical research and intervention (Quan et al., 2008). This has necessitated the building of automatic meth-

ods to sift through and extract valuable information from these unstructured resources.

2 Related Work

Natural Language Processing (NLP) models have been developed in attempts to resolve the above issues in unstructured text. Multiple papers have explored the utility of transformers for tasks in clinical NLP. BioBERT (Lee et al., 2020) and ClinicalBERT (Alsentzer et al., 2019), two models based on the original BERT model, both showed that models specifically pretrained on biomedical and clinical data performed better on a variety of tasks compared to their general-domain counter parts. This has been replicated across several different studies both within domain and for other domains (Gururangan et al., 2020; Si et al., 2019). Gu et al has since shown that given enough data, BERT models trained entirely on domain specific data outperform those pretrained on both general domain and domain specific data (Gu et al., 2021). Researchers have since focused on improving specific components of the model such as fine-tuning performance on a variety of tasks (Tinn et al., 2021). Most recently Li et al introduced ClinicalLongformer and ClinicalBigBird (Li et al., 2022) which pretrain the sparse transformer architectures with clinical data achieving state-of-the-art results.

An ontology encompasses a representation, formal naming, and definition of the categories properties, and relations between the concepts, data, and entities that substantiate one, many, or all domains of discourse. To put this simply, an ontology shows the properties of a subject area and how they are related, by defining a set of concepts and categories that represent the subject. In the medical field, an ontology describes the concepts of medical terminologies and the relation between them, thus, enabling the presentation of medical knowledge in a more structured manner. In the biomedical domain ontologies have been used to augment word and token representations. Additionally, they are used for breaking ties between words. These ontologies are helpful in natural language processing since they provide structure and understanding to the words which are often used in the medical domain. There has been much work done in using ontologies as embeddings and using those embeddings as input to the model. Often times they are paired with words from the sequence and then used as input to the model.

Biomedical ontologies have been used in several transformer-based architectures to improve the performance of deep neural networks for medical applications. When using ontologies in attention mechanisms there are two major methods of leveraging them. The first is using the ontology to augment the embeddings that are passed into the model. The second is augmenting the tokens within the model. One research group used the latter to learn robust representations of medical codes in the face of data scarcity (Choi et al., 2017). Instead of words this model uses diagnosis codes; when a term is rare they can use the parent of the term so that they can have a better representation. This is similar to our work in that tokens are weighted by presence in an ontology. However, ours differs in that it is directly applied to text and that the weight it uniform for those tokens that are selected and zero otherwise.

Another study uses clinical ontologies to improve transformer-based medical recommendation (Shang et al., 2019). Their model also produces an embedding of each medical concept with information on its ancestors in the ontology. However, in contrast to the model used by Choi et al. which used an attention unit to produce an ontology-aware embedding, the model designed by Shang et al. (2019) uses a graph neural network to produce the embedding, which is then used in a subsequent BERT-based architecture. This is the embedding method mentioned above, our work does not use this and instead it directly changes the tokens in the model. However, both of these are on a different domain, working directly on the entries in the form of diagnosis codes instead of words in a document. We give an example of ontologies being used in text in Sotudeh et al.

We will be discussing more on the two main works that we are interested in using to address the problem of time complexity in transformers and usage of ontology in medical domain.

2.1 Clinical Longformer and Clinical BigBird

Clinical BigBird (Li et al., 2022) demonstrated the importance of pre-training on clinical-corpora. (Li et al., 2022) basically pre-trained the Longformer and BigBird models on clinical corpora and evaluated the models on 10 baseline tasks including named entity recognition, question answering, and document classification tasks. The resultant models significantly outperformed ClinicalBERT as well

as other short-sequence transformers in all downstream tasks.

Our model is very similar to the Clinical BigBird model, as it is built from the same basic BigBird architecture and even borrows the Clinical BigBird's weights as a starting point for training. However, our model differs from theirs in the structure of the attention unit. While their attention unit uses global, window, and random attention from the original BigBird model, ours only uses the original window attention, while replacing random and global attention with our ontology-mediated attention. Since we largely based our model of the Clinical BigBird model, we used it as our primary baseline for experiments.

2.2 Attend to Medical Ontologies

In their work, Sotudeh et al. produce ontology-aware representations which are then fed into a transformer-based architecture (Sotudeh et al., 2020) for abstractive summarization of clinical records. Their approach involves creating a new vector representation containing information on which words from the source document are ontological terms. They then use this in a filtering gate function to filter the original embedded tokens. The ontology-aware representations produced through filtering are then directly fed into an attention unit.

Our work is similar to that of Sotudeh et al. in that we attempt to used ontological information to filter the representations that will be attended to. However, while Sotudeh et al. use ontological information to alter the word representations that are input into the attention unit, we use ontological information to alter the attention mask used in the attention unit. Additionally, we consider the tokens in a window around each ontological term, whereas they only considered ontological tokens in their filtering process.

3 Problem Statement

The goal of this study is the build an efficient language model for biomedical text. Biomedical documents can be difficult for language models to generalize to as they frequently have obscure words, abbreviations, and homonyms. The length of these document also poses an issue as modern language modeling has high memory and compute requirements. This makes language models that can process entire documents at once difficult to build. We intend to use existing knowledge bases to improve both the performance and time complexity for biomedical language modeling.

4 Proposed Solution

Our intended usage of ontological information most closely resembles that of Sotudeh et al. (Sotudeh et al., 2020). However, in our case we aim to use ontological information to filter keys in the attention layer itself, inducing a sparsely connected attention unit. This differs from Sotudeh et al. as there mechanism created a context embedding of the ontological terms which was concatenated to the attention vectors when passed into the decoder (Sotudeh et al., 2020). This limits the ontological information that flows through the model as they all this information has to be compressed down to a single vectors. This also does not take into account the direct context of the terms, only the terms that highlighted as ontological terms are added to the ontology embedding. Tokens that may augment the meaning of these terms are ignored which could have an impact on the model. We address this by having the terms directly attended to, instead of concatenating a vector presented them. We also introduce a window mechanism so that the context of the terms can be included in the attention.

Our model will also be built using a sparse attention mechanism. This means that every token will not be attending to every other token. We will be basing our architecture off of the Big Bird architecture which uses multiple sparse mechanism to improve time complexity (Zaheer et al., 2020). This will be expanded upon in the following sections.

4.1 QuickUMLS

While the model parses and extracts task-related information, it lacks prior domain knowledge. We aim to use an ontology to generate a structured heuristic to improve model learning speed and accuracy. The biomedical ontology that will be used is called the UMLS, or the Unified Medical Language System. It provides a graph-like structure compromised of medical vocabularies, along with connections between them. In itself, it can act as a knowledge base, and can be used to provide structure that helps relate any number of unstructured inputs or a subset thereof. We aim to modify the attention mechanism input by adding a preprocessing step that will aim to create a fast, efficient heuristic based on the ontology. To do that, we leverage

QuickUMLS (Soldaini and Goharian, 2016).

QuickUMLS is a fast, unsupervised algorithm that uses simstring matching to process documents and extract medical concepts based on a tuned modification of the UMLS database. More specifically, The original database is converted into a slimmed-down version that groups key-value pairs of UMLS terms and their id. For an arbitrary text input, an efficient matching algorithm extracts potential UMLS-related substrings, and looks for matching concept labels from the knowledge base. It then returns candidate concepts (represented as unique concept ids) as well as string similarity scores for each one. These similarity scores provide a fast heuristic for the model to relate input subsets that could inform one another.

With the generated heuristic, we can constrain the random attention as mentioned above to a subset of the input document based on the heuristic. In effect, this aims to boost the random attention that was used in the BigBird model by leveraging QuickUMLS to make fast educated guesses in regards to which input blocks should be attended to, instead of the random selection. The heuristic-based block attention thus allows for less vectorized attention operations, while aiming to minimize the loss of information that could propagate with the reduction from global attention. QuickUMLS is uniquely applicable for the task due its relatively minimal complexity, speed of processing, and reliable results. Additionally, we can leverage the similarity heuristic to control the broadness of concepts that the algorithm returns.

More concretely, by adjusting the simstring similarity score between the input and UMLS concepts, we can effectively control the labeling predictions. Higher similarity scores in QuickUMLS represent concept labels that match more closely with the string literal. Thus, a higher similarity threshold for acceptance will yield a shorter list of candidate concepts that will be used for attention, with each of the candidates being a closer match and, consequently, a more accurate possible-label prediction in probability. The trade-off is that a mislabeling can cause the model to skip terms that are actually related, which is undesirable. Conversely, a lower similarity threshold decreases the chances of mislabeling completely, as the output candidate labels will be a superset of the output with a higher threshold, but will also increase the amount of possibly unrelated concept labels, which is also undesirable.

The process of generating concept labels for the input is thus a varying heuristic, which we postulate can both help with the model results as well as accelerate the learning process due to more selective attention. By constraining the attention, the model will be able to deal with longer medical documents, effectively solving the issues associated with the infeasibily of performing full-attention on longer input sequences.

4.2 Ontology-Mediated Attention

The attention mechanism used in BigBird (Zaheer et al., 2020) can be described in terms of the generalized attention mechanism. That is, for an input sequence $\boldsymbol{X} \in \mathcal{R}^{n \times d}$ and a directed graph D comprised of n nodes, the ith output vector for the generalized attention can be formulated as

$$A_D(\boldsymbol{X})_i = \boldsymbol{x}_i + \sum_{h=1}^{H} \sigma\left(Q_h(\boldsymbol{x}_i) K_h(\boldsymbol{X}_{N(i)})^T\right) \cdot V_h(\boldsymbol{X}_{N(i)}) \qquad (1)$$

where Q_h, K_h, and V_h are the query, key, and value matrices respectively, σ is the softmax function, H is the number of heads, and $N(i)$ is the set of out-neighbors for node i in D. The graph of D is not fully-connected for sparse attention, in contrast to that of full attention, which is fully-connected. For our implementation of attention, we replaced the edges that were added via random attention with edges that are between nodes selected using ontological information.

For our transformer implementation, we have forked the BigBird code from Hugging Face transformers (Wolf et al., 2019). We chose to use this over the code written by the original authors of the BigBird paper since this code was written in PyTorch (Paszke et al., 2019). The defining feature of the BigBird model is its sparse attention unit, which utilizes three different types of attention: random, window, and global. In our implementation, we replaced random and global attention with our ontology-mediated attention, using the output of QuickUMLS to inform our attention.

The sparse attention unit for the BigBird model does not operate on individual input vectors, but rather on blocks of input vectors. This is because computations over random elements at sporadic locations of a matrix are not conducive to parallelization. Thus, to effectively leverage GPU computing, the attention mechanism performs operations on blocks of size 64. To accommodate for block size

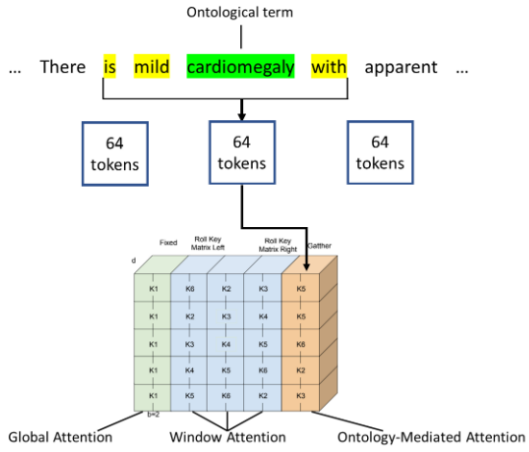

Figure 1: Diagram of ontology-mediated attention for random attention only.

in processes involving selecting groups or windows of vectors to attend to, we had to ensure that the sizes of these windows or groups were an integer divisor of the block size so that we can pack the collected vectors into complete blocks.

For our implementation of attention, we had modified BigBird's sparse attention function to take, as input, a list of indices corresponding to the positions of the ontological tokens in the input sequence (Figure 1). This list was obtained by checking whether each token in the input sequence was in the ontology via QuickUMLS. To implement our replacement for random attention, we randomly chose a fixed number of these indices corresponding to ontological tokens. For each index, we added it, and the indices in a window around it, to the set of indices that would be attended to by our ontology-mediated attention. We initially tried using a window size of 4, but we have also experimented with a window size of 16. Then, from the key and value matrices, we gathered the vectors associated with the selected indices into three blocks of size 64 for each of keys and values. For our replacement for global attention, we followed the same procedure as for random attention, but we selected vectors for two blocks instead of three.

For our model in which we replaced both random and global attention with our ontology-mediated attention, we aggregated the blocks from ontology-mediated attention, along with those selected from the unaltered window attention, into a large, blocked matrix containing vectors for all tokens that would be attended to in this attention unit. For model instantiations where we didn't replace global attention or random attention, we also included the blocks of vectors selected by the unaltered global attention or unaltered random attention mechanisms respectively in this blocked-attention matrix. Then, using these blocked-attention matrices for keys and values, we computed attention via the standard dot-product attention formula.

5 Experiments

5.1 Datasets

i2b2 is a benchmark dataset for clinical entity recognition. It has also been used to derive a question and answering dataset known as emrQA which we used for evaluation. This is done in a semi-automatic manner by using the location of entities in a document and building sets of questions around those entities. For example, if the medication Tylenol is identified as an entity, emrQA will generate the question "Is this patient taking Tylenol" and use the line with the entity as the answer (Pampari et al., 2018). This makes emrQA an extractive question answering dataset similar to SQUAD. We selected Medication, Relations, and Heart Disease as the topics for question answering due to their use in previous papers.

5.2 Data Preparation

We followed preprocessing and sampling techniques identified in (Yue et al., 2020). They showed under-sampling leads to no change in performance for Medication and Relation tasks, as there is a large amount of redundancy in each of the questions from the automatic generation mechanism. When documents were longer than the allowed input sequences, we broke them up into smaller documents for each question, and allowed for "no answer" to be possible in sections without the answer. We also add overlap into these sections to ensure that the full answer would appear in at least one document and not be cut off. Training, validation, and test sets were built in a ratio of 7:1:2 for each dataset.

5.3 Evaluation Metrics

For evaluation we select two common metrics used for question answering, Exact Match (EM) and f1-score. Exact Match operates just as it sounds: an exact match between the answer string and the predicted string. F1-score, on the other hand, is derived from token level performance, and measures

the overlap between the predicted sequence and the actual sequence.

5.4 Baselines

Using this dataset we evaluated several baseline models. BERT, BioBERT, and ClinicalBERT are all full attention models with the latter two being pretrained on biomedical documents. As they are full attention models with input size 512, they required the preprocessing steps that were outlined above. The last model we ran as a baseline is Clinical BigBird. This is a Big Bird model that was pretrained on clinical data and is the current state of the art for this task.

5.5 Attention Configuration

For the selective step of our model, when QuickUMLS is run over the text, we only include the actual document and not the question. We chose to do this as we did not want information in the question to be attended to when it was not next to the SEP token. If we did not do this, it is possible that information from the question could leak into the attention, without the proper context. All other steps outlined above were followed.

We evaluated three different attention formats: Random, Global, and Random + Global. You can find these described above. We also looked at different window sizes: 4, 16. These were chosen to fit inside the block size of 64 evenly. We did not use 1 or 2 as many entities identified by QuickUMLS spanned multiple tokens.

For trained each model started from the ClinicalBigBird Checkpoint. This is important as these models did not have pretraining for their specific mechanism.

5.6 Ablation

To evaluate the benefits of placing selected tokens in our proposed attention blocks we will perform an ablation study that removes instances of ontology-informed token selection. Instead of using the tokens that were selected from QuickUMLS we will randomly select a token uniformly along with its window. This is different from normal BigBird since we are filling the block with tokens from disparate sections of the document instead of the standard method which takes one continuous block We only did this study for the highest performing model due to time constraints. This experiment should indicate if our method is offering any benefit to the performance.

6 Results

Results for each task and model are shown in Table 1. The best performing models were ClincialBigBird and Rand-4 (Random Ontology mediated attention with window of 4) with ClinicalBigBird performing best on medications and Rand-4 performing best on relations and heart disease risk. More concretely, Rand-4 showed a 0.14% increase on the relation task and a 0.10% increase on the heart disease task for F1-score. Our other proposed models followed with lower performance compared to the ClinicalBigBird baseline. The ablation study performed worse than the ontology study of the same type. The full attention models all performed significantly worse than the all sparse models.

7 Discussion

Our models did not perform as hoped in this study. Almost universally our proposed method performed worse than the previous state-of-the-art. In the sections that it did outperform the state-of-the-art there was only a slight gain, which could have been due to differences in the training. One of the major trends that was noted from the results is that when the global attention was augmented using selected tokens, the performance of the model consistently degraded by a significant margin. The likely explanation for this phenomenon is due to the nature of the task, revolving around question answering. Global attention in this context results in the question always being attended to. By replacing these tokens in the attention our model would no longer attend to the question at every step, which likely led to this decrease in performance.

Another trend that was observed across the board was that a window of size 16 had worse performance than the window of size 4. This is the opposite of what was expected. Our initial hypothesis was that the context around the ontological tokens would be very important, and a larger window capturing greater context would result in increased performance. This result could indicate that the smaller window size provides all the attention that is needed, and that any benefits from accounting for more context tokens is offset by noise from irrelevant tokens in the vicinity of the ontological tokens. We see in the ablation study that selected tokens are bringing some benefit to the model, especially with the F1 score. The exact matches for the ablation study seemed to degrade less than the F1 score. This might imply that the selected tokens

Table 1: Performance of Baselines and Ontology Mediated Attention Models. Rand and Global were the two types of attention mentioned above, while the number if the size of the window that was used when selecting tokens.

	emrQA-Medication		emrQA-Relation		emrQA-Heart Disease	
Model	F1	EM	F1	EM	F1	EM
BERT	17.81	14.23	63.07	59.25	16.65	14.10
BioBERT	20.36	18.59	70.35	65.93	21.94	17.31
ClinicalBERT	23.42	17.62	70.98	67.45	18.81	14.38
ClinicalBigBird	**68.38**	**25.19**	87.72	78.36	68.46	24.94
Ablation Rand-4	67.18	24.86	86.97	77.45	67.60	24.71
Rand-4	67.84	25.03	**87.85**	**78.67**	**68.53**	**25.15**
Global-4	65.89	24.72	85.01	77.36	65.46	23.49
Rand + Global-4	65.29	23.22	84.40	77.28	65.87	24.17
Rand-16	66.58	24.02	87.51	78.37	67.63	24.66
Global-16	65.13	24.50	84.00	76.79	64.79	23.43
Rand + Global-16	64.74	23.44	84.28	76.88	65.08	23.75

are benefiting the model for the difficult questions, since the easier ones are the questions that would have an exact match. However these results are very limited, making inference difficult.

The most significant limitation with this study lies with the training of our proposed model. Due to time constraints we were not able to fully pretrain our models. Instead we used the checkpoint from ClinicalBigBird. Our attention mechanism was not exactly the same, but in essence it is the similar model with some of the weights permuted. We used the checkpoint to get around pretraining because the time and resources required to do so for so many models would have been beyond those allocated for this research project. This makes the ablation study important as it is the only one-to-one comparison that can be made in this study. This limitation also makes the fact that these ablated models performed better than our proposed model more surprising. While it is unfortunate that such a large limitation is present in this study, we think that this opens future doors to explore this methodology more carefully.

While this was difficult for us to validate, there is also a chance that the model learned to only rely on global and windowed attention. In previous studies it was shown that Longformer, a model that only uses window and global attention, had comparable results to BigBird in these tasks. This could mean that only the windowed and global attentions are needed for good results. If this is the case, it would explain why all of the models that are using global and windowed attention perform at a similar level. As seen in the results, the four best models all use this type of architecture.

As mentioned before, the dataset is limited due to the fact that it was automatically generated, which has given way to some artifacts that are undesirable. The first of these is that the sequences are not always correctly terminated. In the generation of the dataset, the text only has the line number for the medications and heart risk, and not the individual words. This constraint makes the ends of the lines a noisy label in the dataset. This is also the likely cause in the difference in performance between the tasks, as relation has much higher reported EM and F1. While we understand the difficulties of acquiring data in this domain, future work should evaluate these models on more concrete datasets. Evaluation of this model must also be done using different tasks, and due to limited resources for this study we could evaluate our model against a few datasets. However, to more comprehensively evaluate the language modeling ability of this model, other tasks should be explored. We believe that this structure would be especially well suited for classification tasks. However, we were unable to find any with documents long enough to warrant using sparse attention.

While we only used the UMLS database in this study, ideally we would leverage multiple ontologies as well as the structure within them. The structure of an ontology could be used in several ways to leverage the knowledge within. As an example, terms could be selected to best cover the graph, making the information as complete as possible. Adding in more ontologies could also allow for selectivity within each ontology, specified for the task.

8 Conclusion

In this study we explored the utility of ontology-mediated attention in language modeling. We evaluated these models using a common question answering task for biomedical documents. While one version of our model did outperform previous state-of-the-art on two of the evaluation datasets, most of our proposed model configurations degraded the performance. However, through our ablation study we did find that ontology-mediated attention provides some benefits over completely random attention. This question is therefore still open. Extensions of this work could take two major directions.

The first is improving the quality of evaluation from this paper. Due to a lack of resources there were some limitations to the training in this study. The most glaring is the lack of pretraining for our proposed model. We instead used the checkpoints from models with the same shape but not same functionality. While this worked for quick experiments, to fully evaluate this method we encourage future work to train all models from scratch using the same methods outlined in (Li et al., 2022).

The second direction is in exploring different token selection methods. In this study we explored a very naive approach by randomly selecting tokens identified by QuickUMLS. One of the main advantages of ontologies is the entity-level information that they contain. Future work could explore further selection mechanisms based on the type of token as described by the UMLS. They could also fine-tune approaches to each task. For example, further research could aim to select the ontological tokens most similar to the question that was answered. There are many exciting benefits that could be gained through leveraging the structure of knowledge bases.

This study did not prove the benefits of ontology mediated attention. However, we hope that this novel method will be explored further.

References

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

Beltagy, Iz, Peters, Matthew E, Cohan, and Arman. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Rewan Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*.

Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F. Stewart, and Jimeng Sun. 2017. Gram: Graph-based attention model for healthcare representation learning. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 787–795.

Choromanski, Krzysztof, Likhosherstov, Valerii, Dohan, David, Song, Xingyou, Gane, Andreea, Sarlos, Tamas, Hawkins, Peter, Davis, Jared, Mohiuddin, Afroz, Kaiser, Lukasz, et al. 2020. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*.

Zihang Dai, Zhilin Yang Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

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

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

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. *arXiv preprint arXiv:2004.10964*.

Kitaev, Nikita, Kaiser, Łukasz, Levskaya, and Anselm. 2020. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*.

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

Li, Yikuan, Wehbe, Ramsey M, Ahmad, Faraz S, Wang, Hanyin, Luo, and Yuan. 2022. Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences. *arXiv preprint arXiv:2201.11838*.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrqa: A large corpus for question answering on electronic medical records. *arXiv preprint arXiv:1809.00732*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style,

high-performance deep learning library. *Advances in neural information processing systems*, 32.

Hude Quan, Bing Li, Duncan Saunders, Gerry Parsons, and Carolyn Nilsson. 2008. Assessing validity of icd-9-cm and icd-10 administrative data in recording clinical conditions in a unique dually coded database. *Health Services Research*, 43(4):1424–1441.

Shang, Junyuan, Ma, Tengfei, Xiao, Cao, Sun, and Jimeng. 2019. Pre-training of graph augmented transformers for medication recommendation. *arXiv preprint arXiv:1906.00346*.

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.

Luca Soldaini and Nazli Goharian. 2016. Quickumls: a fast, unsupervised approach for medical concept extraction. *academia: 29618874*.

Sotudeh, Sajad, Goharian, Nazli, Filice, and Ross W. 2020. Attend to medical ontologies: Content selection for clinical abstractive summarization. *arXiv preprint arXiv:2005.00163*.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*.

Robert Tinn, Hao Cheng, Yu Gu, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Fine-tuning large neural language models for biomedical natural language processing. *arXiv preprint arXiv:2112.07869*.

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

Wang, Sinong, Li, Belinda Z, Khabsa, Madian, Fang, Han, Ma, and Hao. 2020. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing.

Angela Wu, Alexei Baevski, Yann N. Dauphin, and Michael Auli. 2019. Pay less attention with lightweight and dynamic convolutions. *arXiv preprint arXiv:1901.10430*.

Zihao Ye, Qipeng Guo, Quan Gan, Xipeng Qiu, and Zheng Zhang. 2019. Bp-transformer: Modelling long-range context via binary partitioning. *arXiv preprint arXiv:1911.04070*.

Xiang Yue, Bernal Jimenez Gutierrez, and Huan Sun. 2020. Clinical reading comprehension: A thorough analysis of the emrqa dataset. *arXiv preprint arXiv:2005.00574*.

Zaheer, Manzil, Guruganesh, Guru, Dubey, Kumar Avinava, Ainslie, Joshua, Alberti, Chris, Ontanon, Santiago, Pham, Philip, Ravula, Anirudh, Wang, Qifan, Yang, Li, et al. 2020. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297.