ANU-CSIRO at MEDIQA 2019:
Question Answering Using Deep Contextual Knowledge

Vincent Nguyen
Australian National University
CSIRO Data61
vincent.nguyen@anu.edu.au

Sarvnaz Karimi
CSIRO Data61
Sydney, Australia
sarvnaz.karimi@csiro.au

Zhenchang Xing
Australian National University
Canberra, Australia
zhenchang.xing@anu.edu.au

Abstract

We report on our system for textual inference and question entailment in the medical domain for the ACL BioNLP 2019 Shared Task, MEDIQA. Textual inference is the task of finding the semantic relationships between pairs of text. Question entailment involves identifying pairs of questions which have similar semantic content. To improve upon medical natural language inference and question entailment approaches to further medical question answering, we propose a system that incorporates open-domain and biomedical domain approaches to improve semantic understanding and ambiguity resolution. Our models achieve 80% accuracy on medical natural language inference (6.5% absolute improvement over the original baseline), 48.9% accuracy on recognising medical question entailment, 0.248 Spearman's rho for question answering ranking and 68.6% accuracy for question answering classification.

1 Introduction

Medical health search is the second most searched thematic query, representing 5% of all queries on Google (Cocco et al., 2018). However, many queries are semantically identical and are potentially already answered by experts (Abacha and Demner-Fushman, 2016). However, these questions may not be directly retrievable due to semantic ambiguity involving abbreviations (Wu et al., 2017), patient colloquialism (Graham and Brookey, 2008) or esoteric terminology (Lee et al., 2019). Furthermore, in regards to disease, temporality is a key factor in determining the relevance of retrieved answers (Lee et al., 2019). For example, it is more appropriate to retrieve answers relating to the *summer cold* in the summer.

As a means to retrieve these questions that are already answered by experts, question entailment has been proposed to discern relationships between pairs of questions. Recognising Question Entailment (RQE) is the task of determining the relationship between a question pair, $RQE(Q_1, Q_2)$, as either entailment or not entailment, where Abacha and Demner-Fushman (2016) define question entailment as the situation where "a question, Q_1, entails another question, Q_2, if every answer to Q_2 is also a complete or partial answer to Q_1."

Natural Language Inference (NLI) is determining the relationship between pairs of sentences, not just questions. NLI is the task of determining whether a *hypothesis*, H, is inferred (entailment), not inferred (contradiction) or neither (neutral), given a *premise*. In the context of question answering (QA), it can be used to validate if the answer can be inferred from the question.

Though RQE and NLI have thrived in the open-domain setting (Bowman et al., 2015; Rajpurkar et al., 2016), there are unique challenges in applying these tasks directly to the biomedical question answering field. Previous models in the medical domain that used NLI and RQE relied on models which were shallowly bidirectional (Romanov and Shivade, 2018) or rule-based approaches with shallow keyword matching techniques (Abacha and Demner-Fushman, 2016) which would not generalise well.

The MEDIQA (Ben Abacha et al., 2019) challenge, as part of the ACL BioNLP workshop, aims to further research efforts in NLI and RQE by introducing their applications to Biomedical QA.

In this paper, we detail our approach in MEDIQA which addresses some of the problems with biomedical text such as utilising deep contextual relationships between words within a sentence for semantic understanding and ambiguity associated with esoteric terminology, abbreviations, and patient colloquialism. We combine biomedical and open-domain approaches as a means to improve generalisation and bridge the gap between patient colloquialism and biomedical terminology.

Proceedings of the BioNLP 2019 workshop, pages 478–487
Florence, Italy, August 1, 2019. ©2019 Association for Computational Linguistics

2 Datasets

MEDIQA 2019 (Ben Abacha et al., 2019) provides datasets to be used for three different tasks.

Task 1: Natural Language Inference The MEDNLI dataset is used for this task (Romanov and Shivade, 2018). A collection of 11232 medical premise-hypothesis pairs are used for training, 2817 pairs for validation and 405 for testing. We preprocessed the text to remove punctuation, that were designed to ensure patient anonymity as a means to reduce noise while ensuring that sentence integrity was not broken.

For example, *cerebrovascular accident in [**2948**] → cerebrovascular accident in 2948.* Furthermore, we expand all medical abbreviations using the ADAM database (Wu et al., 2017). For example, *On arrival to the ED T97 BP 184/94 HR 92 → On arrival to the emergency department Temperature 97 Blood Pressure 184/94 Heart rate 92.*

Task 2: Recognizing Question Entailment For RQE, a collection of 8588 medical question pairs for training, 302 pairs for validation (Abacha and Demner-Fushman, 2016) and 230 pairs for testing is released. The RQE collection aims to match consumer health questions from the National Library of Medicine with Frequently Asked Questions (FAQs) from NIH websites.

Task 3: Question Answering Two separate training datasets were provided from the MEDIQA challenge (Ben Abacha et al., 2019):

LiveQAMed: 104 consumer health questions covering different types of questions about diseases and drugs alongside their associated answers.

Alexa: 104 simple questions about the most frequent diseases and associated answers.

No external data was used for any of the tasks as a conscious decision in order to assess the fine-tuning performance of our models. However, external data has shown to be useful in knowledge-based approaches (Romanov and Shivade, 2018) and we leave this as future work.

3 Our System

Due to the similarity of our approaches in the three tasks, we first describe a shared model that was utilised by all the tasks. Our approach extends upon the current state-of-the-art models (Lee et al., 2019; Devlin et al., 2019) in the open-domain and apply them to the MEDIQA biomedical tasks. As the state-of-the-art models currently employ transfer learning, we modelled an ensemble transfer learning approach used in the medical computer vision domain (Menegola et al., 2017; Kumar et al., 2017).

Algorithm 1: Ensemble Approach for NLI, RQE and QA

```
Input: Training Data, $x \in X$, Test Data,
       $z \in Z$, Hyperparameters $\Theta$,
       Pre-trained Models $M_{Brt}$ and $M_{Bio}$
Output: Label Predictions, $y \in Y$
$X \leftarrow PreprocessText(X)$;
$Z \leftarrow PreprocessText(Z)$;
while numEpochs < totalEpochs do
    for $b_x \in X$ do
        //$b_x$ is a minibatch of $X$
        $M_{BioFT} \leftarrow Train(M_{Bio}, b_x, \Theta)$;
        $M_{BrtFT} \leftarrow Train(M_{Brt}, b_x, \Theta)$;
        //$M_{FT}$ denotes the fine-tuned model
    end
    $numEpochs$++;
end
for $x \in X$ do
    $Pred^x_{Bio} \leftarrow Predict(M_{BioFT}, x)$;
    $Pred^x_{Brt} \leftarrow Predict(M_{BrtFT}, x)$;
    //$Pred$ is the softmax score outputs from
      each model
      $SVM \leftarrow Train(Pred^x_{Bio} \oplus Pred^x_{Brt})$
end
$Pred^Z_{Bio} \leftarrow Predict(M_{BioFT}, Z)$;
$Pred^Z_{Brt} \leftarrow Predict(M_{BrtFT}, Z)$;
$Y = Predict(SVM, Pred^Z_{Bio} \oplus Pred^Z_{Brt})$;
return $Y$
```

BERT As part of our strategy to combine open-domain approaches to a biomedical focused one, we elected to use a current state-of-the-art open-domain approach, *BERT* (Devlin et al., 2019), that is based on deeply bidirectional, unsupervised language representation that has been trained on Wikipedia.

BioBERT From the biomedical focused approach, we used *BioBERT* (Lee et al., 2019), a version of *BERT* that has been pre-trained using additional biomedical datasets, including *PubMED* and *PMC*.

Table 1: Hyperparamters used for each run for Tasks 1 & 2.

	Task 1				Task 2		
Run	**Model**	**Learning Rate**	**Batch Size**	**Epochs**	**Learning Rate**	**Batch Size**	**Epochs**
1	BioBERT	2e-5	64	1	2e-5	64	1
	BERT	8e-6	32	1	8e-6	32	1
2	BioBERT	2e-5	64	40	2e-5	64	40
	BERT	8e-6	32	40	8e-6	32	40
3	BioBERT x3	2e-5	64	40	2e-5	64	40
	BERT	8e-6	32	40	8e-6	32	40
4	BioBERT x3	-	-	-	2e-5	64	-
	BERT	-	-	-	2e-5	32	-
5	BioBERT x3	1e-6	32	100	1e-6	32	100
	BERT	1e-6	32	100	1e-6	32	100

Table 2: Tokenisation statistics for all Tasks.

Task	**Statistic**	**Training**	**Validation**	**Testing**
1	Average Sequence Length	386	190	64
2	Average Sequence Length	176	276	230
3	Average Sequence Length	605	632	582
	Portion of Docs >512 Sequence Length	0.32	0.37	0.32

Support Vector Machine We combined our predictions from our open-domain and biomedical domain approaches using a support vector machine (Cortes and Vapnik, 1995), which here, is akin to using a data-driven weighting function.

Learning-to-Rank We also used learning-to-rank models such as LambdaRank (Burges et al., 2007) and RankNet (Burges et al., 2005), which were implemented in Tensorflow Ranking[1] for the ranking portion of the challenge.

Sentence Embeddings When encoding our features into sentence embeddings, we used bert-as-service[2] in conjunction with BioBERT to create context-rich embeddings of text. In one of our post-challenge runs, we used a biomedical word2vec word embedding model (Chiu et al., 2016).

Hyperparameters For all three tasks, we experimented with batch sizes (2^N, $n \in \{3, 4, 5, 6, 7\}$) and learning rates ($A \times 10^B$, $A \in \{1, 2, 3...10\}$, $B \in \{2, 3, 4, 5, 6\}$) and selected the parameters that maximised performance on the validation set. We used the default sequence length of 64 for training, validation and testing of all three tasks.

Algorithm For the classification tasks in the challenge, we used an ensemble approach (see Algorithm 1). First, the text training data, X, and testing data, Z, is preprocessed. This preprocessing is done differently depending on the submission and task. Preprocessing includes punctuation removal and abbreviation expansion. This training data is used to train the BERT and BioBERT models using hyperparameters, Θ. The softmax scores for each training example, X, predicted by the final fine-tuned models are concatenated (denoted by $\oplus$ and used to train an SVM). The final predictions for the testing set, Z, are collected by first using the fine-tuned models to predict the softmax scores. These softmax scores are concatenated and fed as input into the SVM which outputs predictions, Y, for the test set.

Task 1: Natural Language Inference

The models were trained as follows: For the first and second run, BERT[3] is trained for a single epoch with a learning rate of 8e-6 with a batch size of 32, while the BioBERT[4] models were trained with a learning rate of 2e-5 with a batch size of 64. The models had their predictions combined

[1] Tensorflow Ranking
[2] Bert-As-Service Sentence Embeddings
[3] BERT Base Model
[4] BioBERT Pretrained Models

via an SVM (sklearn-pandas, version 1.8.0) with a penalty of 1.0, RBF kernel and gamma with the 'auto' parameter, which was then used as a data-driven weighting function. The code used for this portion was based on the following code from the BERT repository.[5]

For run 1, we established a baseline approach with no preprocessing and the models were trained only for one epoch. From run 2 onwards, preprocessing was done to the text to remove punctuation used for patient anonymity and expand medical abbreviations as mentioned previously. For runs 3 and 5, instead of using a single BioBERT model, the three variants of BioBERT were trained individually using the same parameters as in run 2. However, in the fourth run, early stop validation was used to select the best models that maximised validation accuracy. However, we excluded this run because it had the same predictions as run 3. In the final run, the learning rate was lowered and trained over a larger number of epochs.

Task 2: Recognizing Question Entailment

We use the same runs as Task 1. However, we did not do any preprocessing for any runs as it did not have any benefit on the validation set.

Task 3: Question Answering Task 3 was a 2-part challenge where answer snippets needed to be ranked and classified as relevant or irrelevant.

```
Algorithm 2: Ensemble Approach for Ranking QA
Input: Alexa Training Data, $T_A$, LiveQA
       Training Data, $T_L$, Test Data, $Z$
Output: Ranked List, $RL$
while $numEpochs < totalEpochs$ do
    for $b_a, b_l \in (T_A, T_L)$ do
        //$b_a$ is a minibatch of $T_A$
        $M_{Alexa} \leftarrow Train(FE(b_a), \Theta)$;
        $M_{LiveQA} \leftarrow Train(FE(b_l), \Theta)$;
        //FE is a feature extractor that
          vectorizes input
    end
    $numEpochs$++;
end
$RL_{Alexa} \leftarrow Predict(M_{Alexa}, Z)$;
$RL_{LiveQA} \leftarrow Predict(M_{LiveQA}, Z)$;
$RL \leftarrow RankScore(RL_{Alexa}, RL_{LiveQA})$
return $RL$
```

In this task, for the ranking task, we mainly used an ensemble of two separate learning-to-rank models that were trained on LiveQA and Alexa (see Algorithm 2). We used the following features as input to the model:

1. BioBERT sentence embedding of Question
2. BioBERT sentence embedding of Answer
3. BioBERT sentence embedding of Entailed Answer from MedQUAD
4. NLI predictions over all candidates summed
5. NLI predictions over all candidates averaged

The first two features were embeddings that were encoded using BioBERT, as mentioned previously. The third feature was found through the following steps:

1. Use BM25 (Stephen Robertson, 1994) to find the question candidates in MedQUAD, M, which are most related to a Question, Q.

2. Set a cut-off value, ρ to minimise the number of candidates for RQE/NLI. For the challenge, we set $rho = 4$.

3. Predict the question entailment between all questions, Q and candidates M using the RQE model, $pred_{rqe}(Q, m) = RQE(Q, m \in M)$.

4. Retain all candidate answers, R, that had questions predicted to be entailed to the Question.

5. Perform NLI on the answers in the original ranked list, L, and all candidate answers extracted from MedQUAD, $pred_{nli}(l, r) = NLI(l \in L, r \in R)$.

6. Use the answer with the highest BM25 score for the third feature.

The fourth and fifth features were performed by summing NLI predictions, $\sum pred_{nli}(l \in L, r \in R)$, and averaging, $\frac{1}{|R|} \sum pred_{nli}(l \in L, r \in R)$.

The features were fed into Tensorflow learning-to-rank models (RankNet for run 1 and LambdaRank for runs 3 and 4) with 2307 features using the Adam optimizer (Kingma and Ba, 2015), a group size of 2 and a learning rate of 0.001.

We ensembled predictions from the two models in two different ways. We used simple averaging for Run 1. However, for subsequent runs, we used *RankScore* (Li et al., 2013), which we define as:

[5] Sentence Classification Bert Code

Table 3: Results for all 3 tasks in the MEDIQA shared task, additional post challenge runs are included. **Note:** With the exception of Task 1, all post challenge runs were evaluated using the official evaluation script.

	Task 1	**Task 2**	**Task 3**		
Run	**Accuracy**	**Accuracy**	**Accuracy**	**Spearman's Rho**	**Precision@1**
1	0.751	0.481	0.581	0.093	0.580
2	0.800	0.485	0.584	0.122	0.640
3	0.796	0.481	0.584	-0.007	0.520
4	-	0.489	0.584	-0.043	0.533
5	0.768	0.485	0.577	0.162	0.593
Post Challenge Runs					
Task	**Description**		**Accuracy**	**Spearman's Rho**	**Precision**
1	Run 5 + Maximum Sequence Length (Validation Set)		0.827 (+0.016)	-	-
2	Run 5 + Maximum Seq. Length		0.489 (+0.004)	-	-
3	Run 5 (Corrected Submission)		0.686 (+0.109)	0.0513 (-0.111)	0.771 (+0.178)
3	Run 5 (Corrected Submission) + Max Seq. Length		0.663 (-0.023)	0.0971 (+0.046)	0.749 (-0.022)
3	Run 1 with word2vec embedding		-	0.284 (+0.189)	-
3	Run 5 (Corrected Submission) with UMLS concept expansion		0.659 (-0.027)	0.0200 (-0.0313)	0.749 (-0.022)

$R_s(d \in D) = 1/d_r$. We use RankScore to score each item in the ranked lists of Alexa and LiveQA models. We then combine the items by summing the documents RankScore from each model and sorting.

For classification, the same architecture from Tasks 1 and 2 for Runs 3 - 5 was used (4-ensemble with SVM layer). For runs 2 and 5, we use softmax scores output from the classification to rank documents.

4 Results and Discussion

Ensembles have been successfully utilised in other biomedical domains (Kumar et al., 2017; Brijesh and Zahid, 2011), with the main idea behind using these being to incorporate complementary strengths of the members of the ensemble. Thus, BERT is used in conjunction with BioBERT in order to correct the mistakes that the model makes by injecting non-domain specific knowledge. This idea was supported in our baseline experiments on task 1 where BioBERT scored 0.7913 on validation, while BERT scored 0.7715 on validation, but ensembling resulted in a higher final score of 0.7950.

NLI Baseline System Problems Our baseline system made characteristic mistakes on the validation set, which is shown in Table 4 for Task 1. We found that our system had trouble with *numerical interpretation* and, for instance, was not able to determine the difference between type 1 and type 2 diabetes. Furthermore, this problem is exacerbated when *abbreviations and numerical interpretation* are required in phrases such as *T97 BP 184/94*. Thus, to aid the system in disambiguating abbreviations, we expanded all abbreviations using the ADAM database of common clinical abbreviations and resulted in an 0.049 increase in accuracy. Furthermore, the system would struggle with medical forms of *negation*. However, due to the use of BERT/BioBERT, conventional techniques such as NegEx or removal would break sentence integrity and reduce comprehension, thereby affecting word context, and thus were not viable. Furthermore, punctuation, in terms of patient anonymisation, is also a problem as the punctuation does not carry meaningful semantic content and will confuse the classifiers.

RQE Baseline System Problems In task 2, we found that our baseline system made similar mistakes for different reasons (see Table 5). We found examples of what we consider *near miss* where the definition of partial entailment depends on interpretation. For example, in this question, the user

Table 4: Common mistakes made by the baseline system in Task 1.

Type	**Premise**	**Hypothesis**
Numerical Interpretation	PAST MEDICAL HISTORY: Type 2 diabetes mellitus.	the patient has type 1 diabetes
Abbreviation and Numerical Interpretation	On arrival to the ED T97 BP 184/94 HR 92 RR 24 88% on RA ->98% on NRB.	The patient was hypertensive in the ED
Negation	He denied headache or nausea or vomiting.	He has no head pain
Semantic Gap	HISTORY OF PRESENT ILLNESS:,The patient is a 54 year old male with endstage renal disease secondary to type 1 diabetes who presents for kidney transplant from wife.	patient is on insulin

wants information on hypertension (high blood pressure). However, according to the gold standard, this is not a form of entailment, partial or otherwise. We hypothesise that this lies on the borderline of the entailment definition or may be due to bias. Furthermore, our system struggles with *abbreviations*. However, the examples in the second task dataset are more related to problems with co-reference resolution where abbreviations appear in the original question but not in the FAQ question.

Furthermore, phrases like "come out of" should be aligned to terms such as "discharge", which is an example of a *semantic gap* and require common sense comprehension. This is problematic as BERT is known to struggle with this sort of reasoning (Talmor et al., 2018). Also, we did not adjust the *sequence length* parameter (set to 64), which may have been a source of error. However, a later investigation through a post-challenge run that shows that only Task 1 benefits from an increase in sequence length (see Table 3). Finally, *patient colloquialism* presents a unique challenge where "hole in lung" is to be interpreted as "pleurisy" (lung inflammation). Although we did not address this complex problem, it could be potentially solved through crowd-sourcing of medical forum data. This may be suitable as an area to investigate for future work.

We found that in all our submissions on the test set of the challenge, although our system was able to achieve high results on the validation set of 79%, the models were not well suited for the test set. Our model predicted entailment 92% of the time on the test set, suggesting that the model is overfitting, even though our baseline was trained for only one epoch. We found that the cases where the models make errors are cases where the question contains words such as *diagnosis* and the disease is mentioned, but the semantic content of the question might be about *treatment* rather than the diagnosis. This is very different from the training and validation datasets that were provided, which were much more straightforward and did not require as much comprehension. An example illustrating this difficulty is *Question A: Glaucoma: Can you mail me patient information about Glaucoma, I was recently diagnosed and want to learn all I can about the disease."* and *Question B: How is glaucoma diagnosed?*

Question Answering Submission Problems For the third Task, we incorrectly trained our models to recognise documents with a relevance score of one as irrelevant. In contrast, the task is defined to classify documents of relevance score one and two as irrelevant. By fixing this error, we found that we had over a 10% increase in accuracy (Table 3). However, interestingly, we found that the ranking quality (shown through Spearman's Rho) decreased. Upon investigation, we found two reasons why this problem occurred: (1) our system was able to differentiate the relevance of one from the other three labels much better than differentiating between labels of one/two against three/four. This was reflected in the validation accuracy of our initial incorrect model, which achieved an accuracy of 95% whereas the corrected model scores only 70% on the validation set, (2) we found that the longer the models were trained, the worse the ranking quality became. We hypothesise that the problem is due to how cross entropy loss and softmax functions work. Since the models are minimising KL-Divergence, the softmax scores become more extreme, falling close to 1 or very

Table 5: Common mistakes made by the baseline system in Task 2.

Type	**Question A**	**Question B**
Near Miss	I want more information on Hypertension and fibromyalgia, I seem to be getting only topics on diabetes and I do not have this. I enjoy reading the current info.	What is high blood pressure?
Abbreviation	Hi I have retinitis pigmentosa for 3years, Im suffering from this disease. Please intoduce me any way to treat mg eyes such as stem cell ... Thank you	Are there treatments for RP?
Semantic Gap	Which drug we I take to stop water come out of my nipple	How to Treat Nipple Discharge
Sequence Length	... The problem is my binocular vision is not good enough ... is there any operation that can fix this?	What is Vision Therapy When and why is it needed [for binocular vision]?
Patient Colloquialism	Cure for hole in lung. I certainly would like to request for medical for hole in the lung	How Are Pleurisy and Other Pleural Disorders Treated?

close to 0. This results in the differences between scores of the documents to be very low (forming dense clusters) which reduces ranking quality as the ranking becomes more sensitive to noise and uncertainty (Siddhant and Lipton, 2018).

Question Answering Baseline System Problems Due to the error of our submissions for Task 3, we will not discuss the mistakes that occurred within the challenge for the pointwise ranking runs. Instead, we will look at the mistakes that the post-challenge run encountered for those. However, for the pairwise runs within the challenge, we found that it performed much worse than expected. We attribute this ranking deficit to two important factors.

The first is that BERT sentence embeddings are not useful to represent sentences because the vector space is too condensed (vector representations are very close together). The second is that our vector representations were too large, with BERT sentence embeddings producing embeddings up to 800 dimensions. Using 3 of these embeddings results in a very large input which would take too long to train or hinder convergence. This effect was observed in a post-challenge run where we used Chiu et al. (2016)'s biomedical word2vec embeddings and achieved a much higher Spearman's Rho. The second factor was that the LambdaLoss (Burges et al., 2007) function was not a suitable objective function as the RankNet model performed better.

From Table 2, we find that Task 3 is more verbose than the other two tasks and presents unique challenges as almost a third of the documents will have information loss due to the limitation of maximum sequence length by BERT being 512 due to quadratic memory explosion (Liu et al., 2018). However, we did a post-challenge run where we increased the sequence length with no noticeable difference. This is because the majority of information in these long sequence can be safely discarded. Furthermore, the BERT truncation strategy is to truncate from the end of the sentence, implying that the important information is typically at the start of the answer.

We also find that there are unique challenges in Task 3 due to the use of real patient questions shown in Table 6. We found that problems such as *typos*, *grammar and spellings* mistakes were not directly fixed by the BERT/BioBERT ensemble as the collections were pretrained on academic or formal language (Pubmed, PMC and Wikipedia). However, problems such as *synonyms* (for example, abetalipoproteinemia and Bassen-Kornzweig syndrome) which should be addressed by the model were also not addressable due to a limitation in the vocabulary of the models, which is discussed below. Furthermore, we found cases of *near miss*, for example, the model identifies anemia and treatment options, but it is not the target disease of the question. To address these problems, we use a heuristic to expand UMLS terms in the question and answer, and add these to the start of the sentence to combat the mentioned problems.

Table 6: Common mistakes made by baseline system in Task 3

Type	Question	Answer
Typo	abetalipoproteimemia hi, I would like to know if there is any support for those suffering with abetalipoproteinemia ... keen to learn how to get it diagnosed...	abetalipoproteinemia: Abetalipoproteinemia is an inherited disorder that affects the absorption of dietary fats, cholesterol, and fat-soluble vitamins...
Synonyms	abetalipoproteimemia hi, I would like to know if there is any support for those suffering with abetalipoproteinemia...	Bassen-Kornzweig syndrome (Exams and Tests): There may be damage to the retina of the eye (retinitis pigmentosa). Tests that may be done to help diagnose this condition include...
Near miss	about thalassemia treatment sir,my friend is suffering from thalassemia ,in that majorly red blood anemia,white blood anemia and the blood is comming out from mouth when she got cough .her condition is very severe...	Anemia (Treatment): Anemia treatment depends on the cause. - Iron deficiency anemia. Treatment for this form of anemia...
Grammar and spelling mistakes	Absence seizures Does any damage occurre from these spells. Mental or physical	Seizures: A seizure is a sudden, uncontrolled electrical disturbance in the brain. It can cause changes in your behavior, movements or feelings, and in levels of consciousness. If you have two...
Semantic Gap	Bad Breath I have very bad breath and at times it can make myself and others sick. I need some advice as to what I need to do.	Breath odor (Home Care): Use proper dental hygiene, especially flossing. Remember that mouthwashes are not effective in treating the underlying problem...

We found that the model performs better on the validation set than any of the post-challenge runs (79% accuracy, a 5% absolute increase over the other runs), but did not perform substantially better on the test set (see Table 3).

Problems with Underlying Models One problem in using models such as *BERT* and *BioBERT* is the limitation in the maximum sequence length. This is demonstrated in the test portion of the challenge, where test set answers were much longer than those seen in the training and validation collection. These sequences were longer than the 512 sequence length limit allowed by the *BERT* architecture, which is constrained due to a problem known as the quadratic memory explosion (Liu et al., 2018) leading to exponentially longer training times and memory usage.

Though there are ways to overcome these restrictions such as striding the sentences pairs and labels, this results in contextual information being lost and label imbalance. This restriction also hinders the encoding of long-range dependencies between sequences as only contexts within a fixed length can be considered (Dai et al., 2019).

In addition, we use *BioBERT* as a means of contributing deep clinical contextual understanding of sentences. However, we find that during WordPiece Tokenisation (Devlin et al., 2019), medical terms are *always* split into their sub-word representations as they are out-of-vocabulary, e.g., *arthralgias* → *art hra al gia s*. Wordpiece tokenisation relies on the idea that morphemes carry meaning. However, due to the use of this non-medical vocabulary, specific medical related morphemes are not being learned. For instance, arthr- (where - denotes prefix), means joints and -algias means pain, so the correct tokenisation should be *arthralgias* → *arthr algia s* so that the model can currently learn the semantic meaning behind the morpheme. We find that these limitations hindered the use of these models and their application to the MEDIQA tasks.

We emphasise that there is a real-world application with the models and methods in this challenge. However, if we were to scale our approach to real-world application, we would require external data. Therefore for future work, given more time, we would like to use external datasets such as emrQA (Pampari et al., 2018) and explore multi-task learning due to the similarity of the three tasks and aim to incorporate other medical tasks for a better generalisation of the biomedical question answering. We would also want to train the BERT models on biomedical-focused vocabulary and additional data in the future as a baseline to compare against multi-task learning.

5 Conclusions

In this shared task, we use and improve upon NLI and RQE techniques for medical question answering. Our approach involves utilising deep contextual relationships between words emphasising semantic understanding and resolving ambiguity. We combine biomedical and open-domain strategies to improve generalisation and bridge the gap between the open-domain and biomedical domain question answering.

Acknowledgements

This research is supported by the Australian Research Training Program and the CSIRO Postgraduate Scholarship.

References

Ben Abacha and Demner-Fushman. 2016. Recognizing Question Entailment for Medical Question Answering. *American Medical Informatics Association Annual Symposium Proceedings*, 2016:310–318.

Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. 2019. Overview of the MEDIQA 2019 shared task on textual inference, question entailment and question answering. In *Proceedings of the BioNLP 2019 workshop*, Florence, Italy. Association for Computational Linguistics.

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Lisbon, Portugal. Association for Computational Linguistics.

Verma Brijesh and Hassan Syed Zahid. 2011. Hybrid ensemble approach for classification. *Applied Intelligence*, 34(2):258–278.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. 2005. Learning to rank using gradient descent. In *Proceedings of the 22Nd International Conference on Machine Learning*, ICML '05, pages 89–96, New York, NY. ACM.

Christopher Burges, Robert Ragno, and Quoc Le. 2007. Learning to rank with nonsmooth cost functions. In *Advances in Neural Information Processing Systems 19*, pages 193–200.

Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. 2016. How to train good word embeddings for biomedical NLP. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 166–174, Berlin, Germany.

Anthony Cocco, Rachel Zordan, David Taylor, Tracey Weiland, Stuart Dilley, Joyce Kant, Mahesha Dombagolla, Andreas Hendarto, Fiona Lai, and Jennie Hutton. 2018. Dr Google in the ED: searching for online health information by adult emergency department patients. *The Medical Journal of Australia*, 209:342–347.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *Computing Research Repository*, abs/1901.02860.

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 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, MN.

Suzanne Graham and John Brookey. 2008. Do patients understand? *The Permanente journal*, 12(3):67–69.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*, San Diego, CA.

Ashnil Kumar, Jinman Kim, David Lyndon, Michael Fulham, and Dagan Feng. 2017. An ensemble of fine-tuned convolutional neural networks for medical image classification. *IEEE Journal of Biomedical and Health Informatics*, 21(1):31–40.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *arXiv e-prints*, page arXiv:1901.08746.

Vincent Li, Paul Thomas, and David Hawking. 2013. Merging algorithms for enterprise search. *ACM International Conference Proceeding Series*, pages 42–49.

Peter Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. *Computing Research Repository*, abs/1801.10198.

Afonso Menegola, Julia Tavares, Michel Fornaciali, Lin Li, Sandra Fontes de Avila, and Eduardo Valle. 2017. Recod titans at isic challenge 2017. *Computing Research Repository*, abs/1703.04819.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrqa: A large corpus for question answering on electronic medical records. *Computing Research Repository*, abs/1809.00732.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. *Computing Research Repository*, abs/1606.05250.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. *Computing Research Repository*, abs/1808.06752.

Aditya Siddhant and Zachary Lipton. 2018. Deep bayesian active learning for natural language processing: Results of a large-scale empirical study. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2904–2909, Brussels, Belgium.

Susan Jones Micheline Hancock-Beaulieu Mike Gatford Stephen Robertson, Steve Walker. 1994. Okapi at trec-3. In *Proceedings of the Third Text REtrieval Conference (TREC 1994)*, Gaithersburg, MD.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *Computing Research Repository*, abs/1811.00937.

Yonghui Wu, Joshua Denny, Rosenbloom Trent, Randolph Miller, Dario Giuse, Lulu Wang, Carmelo Blanquicett, Ergin Soysal, Jun Xu, and Hua Xu. 2017. A long journey to short abbreviations: developing an open-source framework for clinical abbreviation recognition and disambiguation (CARD). *Journal of the American Medical Informatics Association*, 24(e1):e79–e86.