Recognizing Question Entailment in Consumer Health Using a Query Formulation Approach

Jooyeon Lee, Luan Pham and Ozlem Uzuner

March 24, 2023

Recognizing Question Entailment in Consumer Health Using a Query Formulation Approach

Jooyeon Lee and **Luan Huy Pham** and **Özlem Uzuner**
George Mason University, Fairfax, Virginia, USA
{jlee252,lpham6,ouzuner}@gmu.edu

Abstract

The need for online assistance regarding healthcare has grown significantly; a deficiency which has become readily apparent after the advent of the SARS-COV-2/COVID-19 pandemic. A widespread, trusted means of dispersing the latest medical knowledge could have provided tremendous benefit from a public health standpoint and curtailed the spread of a disease which has claimed lives of millions. Question Answering (QA) systems are well-suited to provide this assistance for both medical professionals and the public at-large, especially considering the increased adoption in recent years of virtual digital assistants such as Samsung's Bixby and Google Assistant. The overall performance of QA systems can be improved by a variety of methods, including entailment-based methods. In this paper, we propose a Query-Based Framework for Recognizing Question Entailment (QBF-RQE), which leverages a query formulation method to identify whether two questions are in an entailment relationship – with a specific emphasis on Consumer Health Questions (CHQs). Our approach also incorporates *type* and *focus* features of CHQs to determine the entailment relationship. We evaluate our approach with the MEDIQA 2019 shared task organized at the ACL-BioNLP workshop. Our method gives 83.48%, while the best-performing model for MEDIQA 2019 was 74.9%.

1 Introduction

The purpose of the entailment recognition (or recognition) task is to classify the entailment relationship between a text pair (usually two separate sentences), which are known as the *premise* and the *hypothesis*. The entailment relationship are classified as: *entailment* (the hypothesis having a similar meaning as the premise), *neutral* (hypothesis having similar lexical items but has a different meaning than the premise), and *contradiction* (hypothesis having contradicting meaning versus the premise) relation (Neeraj, 2020; Paramasivam and Nirmala, 2021). Within entailment recognition, there exists Recognizing Question Entailment (RQE) where both the premise and the hypothesis are question sentences. Harabagiu and Hickl (2006) showed improving performance for RQE also improves the QA system in the general domain. Furthermore, Demner-Fushman et al. (2019) showed applicability to the CHQ domain as well by augmenting their Consumer Health Information Question Answering (CHiQA) system with a specific module for RQE.

According to Abacha and Demner-Fushman (2019), the definition of entailment in QA is as follows: *"a question A entails a question B if every answer to B is also a complete or partial answer to A"*. The primary goal of RQE is to ensure the answers of the premise and the answers of the hypothesis align with the entailment relationship, per the definition of entailment. According to Ben Abacha and Demner-Fushman (2016), achieving this goal in QA requires multiple, different approaches.

We propose a new framework – QBF-RQE – which recognizes entailment relations based on the query formulation approach. Our framework leverages insights from CHiQA model presented by Abacha and Demner-Fushman (2019), which uses *type* and *focus* information to form a query to retrieve answers using multiple AI models. Thus, if the hypothesis has the same *focus* and/or *type* as the premise, the retrieved answers to a premise can be a partial or full answer to the answers of the hypothesis. We can state the hypothesis is an entailment of the premise.

While AI models would be ideally trained with premise and hypothesis pairs to achieve high performance, there is a relative lack of suitable, generally-available datasets for CHQA. To directly address this aforementioned lack of available datasets, we attempted several different approaches to augment the official MEDIQA 2019 training set, via several merging-based methodologies: 1) a module

trained with a premise and hypothesis pairs 2) a module trained with question (Entailment Recognition Module: ER Module) and *type* pairs (Type Recognition Module: TR Module) 3) a module trained with question and *focus* pairs (Focus Recognition Module: FR Module), as shown in Figure 1.

2 Datasets

In this section, we describe datasets used to train and test our pipeline modules. The overall performance of RQE in CHQ is measured with the MEDIQA 2019 RQE Challenge test set.

2.1 Entailment Datasets

This section describes the dataset used for the ER module. We use different combinations of MeQSum, the MEDIQA 2019 training set, and the MEDIQA 2019 NLI dataset for training.

1. **MEDIQA2019 RQE Datasets**[1]: This dataset consists of sets of text-hypothesis pairs (clinical question-question pairs) provided by Abacha et al. (2017); Ben Abacha and Demner-Fushman (2016) at NLM. The pairs are labeled either Entailment or Not-Entailment. In the 8,890-pair training set, 4,680 pairs were labeled Entailment and 3,963 pairs were Not-Entailment. In the 302-pair validation set, 129 pairs were labeled Entailment and 173 pairs were Not-Entailment. In the 230-pair test set, the pairs are evenly divided with 115 each.

2. **MeQSum**[2]: We leveraged the fact that answers from summarized CHQ should result in the same answers as the original CHQ to include MeQSum to our RQE task training set. The dataset (Ben Abacha and Demner-Fushman, 2019), also provided by the NLM group, includes 1,000 pairs of CHQ and summarized CHQ.

3. **MEDIQA2019 NLI Datasets**[3]: While not consisting of pairs in question form, a few teams incorporated MEDIQA-NLI (MedNLI) (Romanov and Shivade, 2018) in the MEDIQA 2019 RQE task (Zhu et al., 2019; Pugaliya et al., 2019). The dataset includes clinical sentence pairs: Entailment (3,744 pairs), Neutral (3,744 pairs) and Contradiction (3,744 pairs). Each label has 465 pairs in the validation set, and 474 pairs in the test set.

2.2 Type and Focus Datasets

This section describes the dataset used to train and test the TR module and FR module. For the TR module for RQE task, we use LiveQA, MedInfo and MedQuAD to train the model. For the TR task itself, we use the LiveQA training set, MedInfo and MedQuAD to train, and the LiveQA test set to measure the performance of each model to compare the performance with the baseline (Demner-Fushman et al., 2019). The *type* names and their frequencies are shown in Table 6 in Appendix A.2. For the FR, we consider disease names as a focus of the CHQs to be consistent with the answer retrieval method of CHiQA. LiveQA, MedQuAD, and MedInfo all are in the CHQ domain and have focus entities labeled. However, the Named Entity Recognition (NER) task to identify disease names are already widely available, and for the purposes of this paper, we do not perform re-training for the NER task with CHQ datasets.

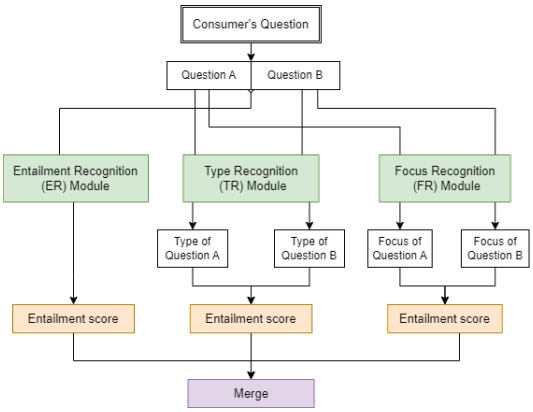

Figure 1: Architecture of the proposed QBF-RQE.

1. **TREC-2017 LiveQA**[4]: The TREC-2017 LiveQA: Medical Question Answering Task (Abacha et al., 2017) organizer provides a dataset (LiveQA) that has 446 pairs in the training set and 104 pairs in the test set.

2. **MedInfo**[5]: The MedInfo (Ben Abacha et al., 2019) dataset is about medication CHQs. The dataset has CHQs, answers, focus, type, section title and URL of the information source.

3. **MedQuAD**[6]: MedQuAD (Abacha and Demner-Fushman, 2019) has 47,457 pairs of medical questions/answers created from NIH websites.

3 Methodology

We describe our model in this section. Our model has 3 modules for different tasks: ER, TR and FR. The detailed architecture of our model is shown in Figure 1. All the models are transformer-based models, and we use pretrained models publicly available in the Hugging Face repository. The parameters we used are listed in Appendix B.

[1] https://github.com/abachaa/MEDIQA2019/tree/master/MEDIQA_Task2_RQE
[2] https://github.com/abachaa/MeQSum
[3] https://physionet.org/content/mednli-bionlp19/1.0.1/
[4] https://github.com/abachaa/LiveQA_MedicalTask_TREC2017
[5] https://github.com/abachaa/Medication_QA_MedInfo2019
[6] https://github.com/abachaa/MedQuAD

3.1 ER Module

For ER, we experiment with 4 different dataset combinations with 6 different models.

Data With MedNLI, MedQuAD and MEDIQA2019, we create 4 combinations of sets: 1) MEDIQA 2019 training set, 2) MEDIQA 2019 training set + MedNLI, 3) MEDIQA 2019 training set + MeQSUM, 4) MEDIQA 2019 training set + MeQSUM + MedNLI.

Model

1. **Bio-Clinical-BERT**: Bio-Clinical-BERT (Alsentzer et al., 2019) is domain-specific contextual word embedding model, which is initialized with BIOBERT model and trained on all MIMIC notes (Johnson et al., 2016).

2. **BiomedNLP-PubMedBERT-base-uncased-abstract** Gu et al. (2020) provide a BERT-based neural language model pretrained on the biomedical NLP benchmark. BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext and BiomedNLP-PubMedBERT-base-uncased -abstract are pretrained models which are available in the Hugging Face repository[7][8].

3. **BioELECTRA-base-discriminator-PubMed** BioELECTRA-base-discriminator-PubMed (Kanakarajan et al., 2021) is a pretrained ELECTRA model-based, biomedical domain-specific language model using discriminators, showing great performance in MedNLI (Romanov and Shivade, 2018) (Language inference task), i2b2-2010 (Uzuner et al., 2011) (NER and relation extraction task), ShARe/CLEFE (Kelly et al., 2013) (NER task) and ClinicalSTS (Wang et al., 2020) (Sentence Similarity task).

4. **BioMed-RoBERTa-base** BioMed-RoBERTa-base (Gururangan et al., 2020) is a language model based on the RoBERTa-base (Liu et al., 2019) model, fine-tuned with 2.68 million scientific papers from the Semantic Scholar corpus. Both full-text of papers and abstracts were used to train.

3.2 TR Module

Data We union labels of the LiveQA (26 types), MedInfo (17 types) and MedQuAD (16 types in Disease-related questions, 20 types in the drug category), resulting in 39 labels. For similar labels, we prioritized matching with LiveQA labels. Specific details regarding the label union methodology/procedure as well as the labels after union are shown in the bottom row of Table 6 in Appendix A.1.

Model We use the same models in the TR module as those in the ER Module.

[7]https://huggingface.co/microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract

[8]https://huggingface.co/microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext

Entailment Score We measure the score as either 1 or 0 (Consistent or Not-Consistent), based on overlapping type labels. If there is any overlap between the *type* of text and the *type* of hypothesis, then we consider it Consistent. If there are no overlaps, then it is Not-Consistent.

3.3 FR Module

NER tasks to identify disease names are popular research area and publicly-available datasets and models are easily accessible. For this paper, we used 2 of the state-of-the-art models for the task and selected the one that gives higher accuracy based on the validation set. One is biobert-diseases-ner (Casero, 2021), which is a BERT-based model trained on NCBI-disease. NER-disease-ncbi-bionlp-bc5cdr-PubMed (Zhang, 2021) is a RoBERTa-based model (Liu et al., 2019), trained on NCBI-disease and BC5CDR datasets.

Entailment Score We test 2 different methods to measure the score: 1) Exact-Match, and 2) Similarity-Based Match. Exact-Match occurs when there is overlap in disease names; which we then classify as entailment. If there is no overlap, then it is not entailment. Similarity-Based Match is utilized to address minor differences/typos in the disease names. We measure the similarity score between each focus in the premise and the hypothesis. If the similarity scores of focus in the hypothesis and premise score is above a threshold, then we consider the pairs to be in an entailment relationship. The similarity score is measured with the S-BioBert-snli-multinli-stsb sentence similarity model (Deka and Jurek-Loughrey, 2021) and the spaCy sentence similarity model (Honnibal and Montani, 2017). S-BioBert-snli-multinli-stsb model is BioBERT (Lee et al., 2019) finetuned with several language inference datasets: SNLI (Bowman et al., 2015), MultiNLI (Williams et al., 2017) and STS-b (Cer et al., 2017). The spaCy model measures similarity by measuring the distance between word vectors trained on a large English general domain.

3.4 Merge

To merge the results, we test a majority-voting (m-voting) and a weighted-voting (w-voting) system. We evaluate three methods for the w-voting, giving a weight on the ER module result, the TR module result or the FR module result.

4 Evaluation

In this section, we discuss the overall performance of the QBF-RQE, along with the performance of each individual module.

4.1 QBF-RQE Results

We measure the performance of the QBF-RQE by calculating the number of correctly-predicted labels over the total number of premise and hypothesis pairs. We then compare our results to the results of the MEDIQA 2019 challenge participants. In Table 1, we list the results of each module for the RQE task, along with the pipeline result with the m-voting method, and the pipeline result with the w-voting methods. The table also includes the list of accuracies from the top 3 best-performing teams at MEDIQA 2019. Accuracy scores of Team Sieg for the Validation set were not reported. Team IIT-KGP reported multiple experiment results. Their model with Sci-BERT with the Hinge-Loss method gave the best performance with the testing set, model with QSpider gave the best result with the validation set.

With the test set, the QBF-RQE model with the w-voting method on the TR module gave 8.58% higher accuracy than the best-performing team. It also showed the best performance amongst all the experiments we performed: ER module, TR module, FR module, QBF-RQE with the m-voting method, QBF-RQE with the w-voting method and QBF-RQE with the union merging method.

Model	**test acc**	**val acc**
ER module	57.39%	**83.04%**
TR module	82.17%	74.17%
FR module	51.3%	70.76%
QBF-RQE m-voting	60.0%	82.46%
QBF-RQE w-voting (ER module)	62.17%	81.29%
QBF-RQE w-voting (TR module)	**83.48%**	**83.04** %
QBF-RQE w-voting (FR module)	56.52%	76.32%
QBF-RQE ER module ∪ (TR module&FR module)	60.43%	81.87%
PANLP (Zhu et al., 2019): Ensemble, transfer learning, re-ranking with BERT and MT-DNN	**74.9%**	**84.77%**
Sieg (Bhaskar et al., 2019): MT-DNN	70.6%	-
IIT-KGP (Sharma and Roychowdhury, 2019): The best model result for Test set - Sci-BERT with Hinge Loss	68.4%	62.0%
IIT-KGP (Sharma and Roychowdhury, 2019): The best model result for Val Set - QSpider	51.3%	80.5%
Baseline - SVM (Ben Abacha et al., 2019)	54.1%	-

Table 1: Evaluation of the QBF-RQE for RQE task.

4.2 Module Performance

We also investigate the performance of each module of the pipeline separately.

4.2.1 ER Module

We notice that the testing set is challenging to classify entailment or not-entailment compared to the validation set or training set. This may be caused by using different methods to create the training, validation, and testing set. The training set includes a repetition of hypothesis questions for not-entailment relationships, while we see fewer of these cases on the validation and none in the testing set. We believe these are the cause of a lot of challenge scores lying below 80% accuracy, while validation accuracies from challenge participants are usually above 80%. To overcome this problem, the majority of teams, including this paper, incorporated augmentation of the training sets.

Hence, for the ER module, we investigate the different combinations of available datasets, along with the performance of 5 different models, and results are shown in Table 2.

Augmenting training set with MedNLI, or MeQSum gives higher performance for all models. Both the MEDIQA training set+MedNLI combination and the MEDIQA training set+MeQSum combination demonstrated a greater than 10% increase vs just the MEDIQA 2019 training set. This shows that augmenting the MEDIQA 2019 helps to improve RQE models. However, merging the MedNLI, MeQSUM and MEDIQA training sets together did not necessarily improve the performance. Combining all datasets gave the best score of 80.99% and the average score of 79.13%, which is higher than the MEDIQA + MedNLI combination, but lower than the MEDIQA + MeQSUM combination. We can therefore conclude that MedNLI may increase performance with a training set which is relatively small/limited, but if there is a training set that has closer characteristics to the test set, merging with MedNLI may not be advantageous.

For the test accuracy on Table 2, we selected the dataset combination which gave the best accuracy to the validation set (MEDIQA 2019 + MeQSum) and added the MEDIQA 2019 validation set to the training set to train and tested on the MEDIQA 2019 test set.

Model with Train-set, MedNLI, MeQSum	Test Accuracy	Validation Accuracy
Bio-Clinical-BERT	51.52%	70.20%
BiomedNLP-PubMedBERT-base-uncased-abstract-full text	51.52%	78.65%
BiomedNLP-PubMedBERT-base-uncased-abstract	52.38%	79.24%
BioELECTRA-base-discriminator-PubMed	54.98%	80.99%
BioELECTRA-base-discriminator-PubMed-PMC-lt	53.68%	80.12%
Biomed-RoBERTa-base	51.08%	74.56%
Model with Train-set, MedNLI		
Bio-Clinical-BERT	49.35%	72.81%
BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext	53.25%	79.82%
BiomedNLP-PubMedBERT-base-uncased-abstract	51.95%	81.29%
BioELECTRA-base-discriminator-PubMed	**56.28%**	80.12%
BioELECTRA-base-discriminator-PubMed-PMC-lt	54.98%	78.95%
Biomed-RoBERTa-base	52.38%	73.98 %
Model with Train-set, MeQSum		
Bio-Clinical-BERT	51.08%	69.0%
BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext	51.52%	77.78%
BiomedNLP-PubMedBERT-base-uncased-abstract	52.38%	80.12%
BioELECTRA-base-discriminator-PubMed	55.41%	**83.04%**
BioELECTRA-base-discriminator-PubMed-PMC-lt	53.25%	80.99%
Biomed-RoBERTa-base	51.95%	76.61%
Model with Train-set		
Bio-Clinical-BERT	54.55%	58.19%
BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext	52.38%	56.43%
BiomedNLP-PubMedBERT-base-uncased-abstract	54.55%	64.33%
BioELECTRA-base-discriminator-PubMed	**56.28%**	79.53%
BioELECTRA-base-discriminator-PubMed-PMC-lt	55.84%	78.65%
Biomed-RoBERTa-base	53.25%	75.44%
Model with Train-set, MeQSum, Validation set		
Bio-Clinical-BERT	50.43%	-
BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext	**57.83%**	-
BiomedNLP-PubMedBERT-base-uncased-abstract	55.22%	-
BioELECTRA-base-discriminator-PubMed	57.39%	-
BioELECTRA-base-discriminator-PubMed-PMC-lt	56.09%	-
Biomed-RoBERTa-base	51.73%	-

Table 2: Evaluation of the ER module for RQE task.

4.2.2 TR Module

Demner-Fushman et al. (2019) thoroughly investigated the individual TR and FR models using Recall, Precision and F1 score with LiveQA test set. They used combinations of SVM and rule-based methods (regular expressions) and deep learning methods to extract the Type from CHQs. We consider this method as a baseline and compare it with transformer-based models. As shown in Table 3, simply merging LiveQA, MedInfo and MedQuAD showed improved performance.

We use the same models to test the ER purpose, to pick the best performing model and plug it into the pipeline, BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext showed the best performance on both the test and validation set. Results are shown in Table 4.

4.2.3 FR Module

ner-disease-ncbi-bionlp-bc5cdr-PubMed shows slightly higher performance than biobert-disease-ner model for the RQE task. Therefore, QBF-RQE results listed in the Section 1, ner-disease-ncbi-bionlp-bc5cdr-PubMed model was used for the FR module, with the entailment score calculated with the Similarity-Based Match method. Exact accuracy is listed on the Table 5.

4.3 Error Cases

In this section, we show 2 cases of error types to show the effect of the TR module and the FR module of our framework.

Case 1. Incorrect prediction from a module. In this case, the QBF-RQE error was caused by the error in the individual module.

- **Not-Entailment (PREMISE)** (Type: information, Focus: fibromyalgia): 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. thanks **(HYPOTHESIS)** (Type:

Data	Precision	Recall	F1
Bio-Clinical-BERT	62.77%	44.70%	52.21%
BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext	**67.39%**	46.97%	55.36%
BiomedNLP-PubMedBERT-base-uncased-abstract	64.95%	47.73%	55.02%
BioELECTRA-base-discriminator-PubMed	65.93%	45.45%	53.81%
BioELECTRA-base-discriminator-PubMed-PMC-lt	61.7%	43.94%	51.33%
Biomed-RoBERTa-base	66.33%	**49.24%**	**56.52%**
SVM+Rule-Based+BiLSTM (Demner-Fushman et al., 2019)	55.5%	42.5%	48.1%

Table 3: Evaluation of a TR module with LiveQA Test Set.

Model	Test Accuracy	Validation Accuracy
Bio-Clinical-BERT	80.43%	71.19%
BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext	**82.17%**	**74.17%**
BiomedNLP-PubMedBERT-base-uncased-abstract	81.3%	73.18%
BioELECTRA-base-discriminator-PubMed	80.0%	73.51%
BioELECTRA-base-discriminator-PubMed-PMC-lt	80.43%	70.20%
Biomed-RoBERTa-base	81.74%	70.20%

Table 4: Evaluation of TR module on RQE task.

Information, Focus: Fibromyalgia): How Is Fibromyalgia Treated?

The premise and hypothesis in the above example are in a not-entailment relationship. While our TR model incorrectly predicted it as an entailment relationship by classifying both premise and hypothesis *Types* as “information", and Focus as “fibromalgia" for premise and “Fibromyalgia", while the *Type* of the premise should be “treatment".

Case 2. Correct prediction from each module This case is caused by either *Type* or *Focus* was not enough information to determine two questions are in entailment relationship.

- **Not-Entailment (PREMISE)** (Type: Information, complication, Focus: Arrhythmia) Arrhythmia. can arrhythmia occurs after ablation? What is the success rate of Ablation? During my Holter test it was found that my Heart rate fluctuates from 254 to 21. How do you rate the situation? **(HYPOTHESIS)** (Type: Information, Focus: Arrhythmia) What is an Arrhythmia?

The above example with a question that has “What is [disease name]" format. This is one of the most commonly occurring errors with QBF-RQE. The answers to this question could include a definition of a potential treatment used for the disease, medicine types, symptoms, and list continues. There is a possibility that answers to “Can arrhythmia occurs after ablation?" could be a part of an answer to “What is [disease name]". Thus, with *Type* information can not be the primary factor of determining the entailment or not-entailment relationship.

- **Entailment (PREMISE)** (Type: information, Focus: Itching): babygirl vagina itching. My newborn is 9 weeks and I noticed when I went to clean inside her vagina there was a little bit of spotting. Should I be concerned? **(HYPOTHESIS)** (Type: Consideration): When to Worry About Your Newborn’s Genitals?

The above example shows an example of *Focus* problem. The *Focus* of premise is “Itching" and there is no disease name we can detect from the hypothesis. Thus no *Focus* was found in the hypothesis. From both the similarity match method and the exact match method, the entailment score is 0 (Not-Entailment). This is caused by limiting the *Focus* to disease names only. Expanding the boundary of *focus* could be one of our future works.

4.4 Limitations and Future work

While the performance of the QBF-RQE is generally improved by combining multiple modules, it is important to note that if the accuracy of a single module is significantly lower for a particular use case, the net effect can decrease overall performance. This characteristic is prominent on the test set. Individual module accuracy of the ER module, TR module and FR modules are 57.39%, 82.17% and 51.3% respectively. The TR Module has the highest recognition and difference of

Data	Test Accuracy	Validation Accuracy
biobert-diseases-ner (Exact Match)	52.17%	66.67%
biobert-diseases-ner (Similarity-Based Match)	51.3%	70.76%
ner-disease-ncbi-bionlp-bc5cdr-PubMed (Exact Match)	48.70%	60.82%
ner-disease-ncbi-bionlp-bc5cdr-PubMed (Similarity-Based Match)	**52.61%**	**71.93%**

Table 5: Evaluation of FR module on RQE task.

accuracy between the TR module vs FR and ER modules is more than 20%. With the majority-voting system, we can see the accuracy reduced to 60% from 82.17%. With weighted-voting based on type, the accuracy is increased by 1.31%. Therefore, when using this approach, it is advantageous primarily when the individual modules have a balanced performance profile. Otherwise, simply employ the module with the best performance, particularly when the individual module is has an overwhelmingly superior performance profile. Second, when there is an bias in performance in one module (though not to an overwhelming degree), the weighted merge imparts improved performance. In the future, we hope to explore methods to further improve the performance of each module and hopefully investigate the different methods to merge the ER, FR and TR modules.

5 Conclusion

Ideally, the best scenario for RQE would be having one AI model and suitably training the dataset with appropriate premise and hypothesis pairs. But, the CHQ domain lacks such a dataset, which therefore limits the performance of the AI models. However, we showed significant improvement in performance in RQE in the CHQ domain by using the query formulation method inspired by the definition of Entailment in QA. In the future, we hope to investigate different ways to incorporate queries and study different methods of extracting queries (not limited to question focus and type characteristics) to build a more versatile RQE pipeline.

References

Asma Ben Abacha, Eugene Agichtein, Yuval Pinter, and Dina Demner-Fushman. 2017. Overview of the medical question answering task at trec 2017 liveqa. In *TREC*.

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *CoRR*, abs/1901.08079.

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. Publicly available clinical BERT embeddings. *CoRR*, abs/1904.03323.

Asma Ben Abacha and Dina Demner-Fushman. 2016. Recognizing question entailment for medical question answering. In *AMIA 2016, American Medical Informatics Association Annual Symposium, Chicago, IL, USA, November 12-16, 2016*.

Asma Ben Abacha and Dina Demner-Fushman. 2019. On the summarization of consumer health questions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2228–2234, Florence, Italy. Association for Computational Linguistics.

Asma Ben Abacha, Yassine Mrabet, Mark Sharp, Travis Goodwin, Sonya E. Shooshan, and Dina Demner-Fushman. 2019. Bridging the gap between consumers' medication questions and trusted answers. In *MEDINFO 2019*.

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 18th BioNLP Workshop and Shared Task*, pages 370–379, Florence, Italy. Association for Computational Linguistics.

Sai Abishek Bhaskar, Rashi Rungta, James Route, Eric Nyberg, and Teruko Mitamura. 2019. Sieg at MEDIQA 2019: Multi-task neural ensemble for biomedical inference and entailment. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 462–470, Florence, Italy. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Álvaro Alonso Casero. 2021. Named entity recognition and normalization in biomedical literature: a practical case in sars-cov-2 literature.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and

crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Pritam Deka and Anna Jurek-Loughrey. 2021. Unsupervised keyword combination query generation from online health related content for evidence-based fact checking. In *The 23rd International Conference on Information Integration and Web Intelligence*, pages 267–277.

Dina Demner-Fushman, Yassine Mrabet, and Asma Ben Abacha. 2019. Consumer health information and question answering: helping consumers find answers to their health-related information needs. *Journal of the American Medical Informatics Association*, 27(2):194–201.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing.

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 ACL*.

Sanda Harabagiu and Andrew Hickl. 2006. Methods for using textual entailment in open-domain question answering. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 905–912, Sydney, Australia. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Alistair Johnson, Tom Pollard, Lu Shen, Li-wei Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Celi, and Roger Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035.

Kamal raj Kanakarajan, Bhuvana Kundumani, and Malaikannan Sankarasubbu. 2021. BioELECTRA:pretrained biomedical text encoder using discriminators. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 143–154, Online. Association for Computational Linguistics.

Liadh Kelly, Lorraine Goeuriot, Hanna Suominen, Tobias Schreck, Gondy Leroy, Danielle L. Mowery, Sumithra Velupillai, Wendy W. Chapman, David Martínez, G. Zuccon, and João Palotti. 2013. Overview of the share/clef ehealth evaluation lab 2013. In *CLEF*.

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. *CoRR*, abs/1901.08746.

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. *CoRR*, abs/1907.11692.

Trishala Neeraj. 2020. Semantic entailment. *trishalaneeraj.github.io*.

Aarthi Paramasivam and S. Jaya Nirmala. 2021. A survey on textual entailment based question answering. *Journal of King Saud University - Computer and Information Sciences*.

Hemant Pugaliya, Karan Saxena, Shefali Garg, Sheetal Shalini, Prashant Gupta, Eric Nyberg, and Teruko Mitamura. 2019. Pentagon at MEDIQA 2019: Multitask learning for filtering and re-ranking answers using language inference and question entailment. *CoRR*, abs/1907.01643.

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

Prakhar Sharma and Sumegh Roychowdhury. 2019. IITKGP at MEDIQA 2019: Recognizing question entailment using sci-BERT stacked with a gradient boosting classifier. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 471–477, Florence, Italy. Association for Computational Linguistics.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Yanshan Wang, Sunyang Fu, Feichen Shen, Sam Henry, Ozlem Uzuner, and Hongfang Liu. 2020. The 2019 n2c2/ohnlp track on clinical semantic textual similarity: Overview. *JMIR Med Inform*, 8(11):e23375.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *CoRR*, abs/1704.05426.

Xiaochen Zhang. 2021. raynardj/ner-disease-ncbi-bionlp-bc5cdr-pubmed.

Wei Zhu, Xiaofeng Zhou, Keqiang Wang, Xun Luo, Xiepeng Li, Yuan Ni, and Guotong Xie. 2019. PANLP at MEDIQA 2019: Pre-trained language models, transfer learning and knowledge distillation. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 380–388, Florence, Italy. Association for Computational Linguistics.

A Drug Label Merge Details

A.1 Union Labels

To avoid redundant labels, we merged labels originating in different datasets which are functionally identical but possess minor variations in spelling. The bottom row in Table 6 contains the labels resulting from the union of LiveQA, MedInfo and MedQuAD. The labels in black originated in LiveQA. Blue labels are labels that originated in MedInfo that do not exist in LiveQA. Red labels only exist in MedQuAD.

We union the labels manually, setting the priority of labels as: #1 LiveQA, #2 MedInfo and #3 MedQuAD. We prioritized LiveQA due to its previous use in CHiQA research, better facilitating comparisons. Thus, we rename the labels to match the spelling as it exists with LiveQA, if possible. If the label does not exist in the LiveQA but only in MedInfo and MedQuAD, we arbitrarily modified the MedQuAD label to match with MedInfo. For example, side effects in MedInfo is modified to side-effect to match with LiveQA, while side effects, severe reaction in MedQuAD are modified to *side-effect* to match with LiveQA. In Table 6, the example is represented as "side effects (side effects, side effects, severe reaction)".

Another manual task is to identify entailment relationships. For the purposes of the paper, if the two types are in an obvious entailment relationship, we unified the labels. The label special instructions, important warning, precautions, are renamed to *considerations*. In Table 6, the example is represented as "considerations (special instructions, important warning, precautions)".

A.2 Union Datasets

After merging the 3 datasets with the method mentioned in Appendix A.1, the total number of questions and the type pairs are 48,577 and the total number of labels is 39. Due to the resource limitations and to prevent the dataset from overfitting on MedQuAD characteristics, we only select a max of 500 question and *type* pairs for each *type*. Among the 48,577, more than 97% of the dataset is from MedQuAD. With this limit, we have a total of 12,620 pairs. The detailed distribution is listed in Table 7.

B Parameters

We use default parameters of hugging face for 6 models for ER and TR except warmup_steps, save_steps, batch size, epochs, weight_decay and learning_rate. We perform a grid search method to find an optimal parameter for each model: warmup_steps=100, save_steps = 500, batch size = 16, epochs = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}, weight_decay={0.01, 0.1}, learning_rate = {5e-6, 1e-5, 5e-5, 1e-4}. weight_decay of 0.01 gave the best result for all tasks and models. learing_rate of 5e-5 gave the best results for the TR module. For the ER module, the best results were given when learning_rates lies between 1e-5 and 5e-5. Bio-Clinical-BERT, BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext, Biomed-RoBERTa-base results are model results with learning_rate of 1e-5. BiomedNLP-PubMedBERT-base-uncased-abstract, BioELECTRA-base-discriminator-PubMed and BioELECTRA-base-discriminator-PubMed-PMC-lt results are models trained with learning_rate of 5e-5.

Dataset Name	**Labels**	**# of Pairs / # of Labels**
LiveQA	treatment, information, cause, diagnosis, susceptibility, interaction, person-organization, side-effect, effect, ingredient, prevention, symptom, tapering, usage, complication, contraindication, dosage, indication, prognosis, storage-disposal, comparison, inheritance, action, alternative, lifestyle-diet, other-question, genetic changes, resources	Train: (p) 446 / (L) 23 Test: (P) 104 / (L) 26
MedInfo	information, dose, usage, side effects, indication, interaction, action, appearance, usage/time, stopping/tapering, ingredient, action/time, storage and disposal, comparison, contraindication, overdose, alternatives, usage/duration, time, brand names, combination, pronunciation, manufacturer, availability, long term consequences	(P) 674 / (L) 25
MedQuAD	Diseases: information, research (or clinical trial), causes, treatment, prevention, diagnosis (exams and tests), prognosis, complications, symptoms, inheritance, susceptibility, genetic changes, frequency, considerations, contact a medical professional, support groups Drugs: information, interaction with medications, interaction with food, interaction with herbs and supplements, important warning, special instructions, brand names, how does it work, how effective is it, indication, contraindication, learn more, side effects, emergency or overdose, severe reaction, forget a dose, dietary, why get vaccinated, storage and disposal, usage, dose, precaution Medical Entities (ME): information	(P) 47,457 / (L-disease) 16, (L-drug) 20, (L-ME) 1
Union	treatment, information (other-question, learn more), cause (causes), diagnosis, susceptibility, interaction (interaction with food, interaction with herbs and supplements, interaction with medications), person-organization (contact a medical professional, support groups), side-effect (side effects, side effects, severe reaction), effect (how effective is it), ingredient, prevention, symptom (symptoms), tapering (stopping/tapering), usage, complication (complications), contraindication, dosage (dose, overdose, dose, forget a dose, emergency or overdose), indication, prognosis(long term consequences), storage-disposal (storage and disposal, storage and disposal), comparison, inheritance, action (how does it work), alternative, lifestyle-diet (dietary), genetic changes, resources (research), appearance, time (duration), comparison, alternatives, brand names, combination, pronunciation, manufacturer, availability, frequency, considerations (special instructions, important warning,precautions), why get vaccinated	Train: (P) 48,577 / (L) 39 Test: (P) 104 / (L) 26

Table 6: Consumer Health Question *Type* Dataset.

***Type* Name**	**# of pairs with 500 max limit**	**# of pairs without max limit on MedQuAD**
information	705	10724
symptom	515	4353
treatment	725	4131
consideration	500	2653
cause	537	2473
dosage	583	2422
prognosis	524	2256
diagnosis	523	2081
organization	517	1976
brand names	503	1471
inheritance	508	1454
side_effect	568	1393
usage	609	1353
indication	559	1317
prevention	505	1244
storage_disposal	515	1132
complication	508	1128
frequency	500	1120
lifestyle_diet	500	1092
genetic changes	501	1088
susceptibility	489	489
resources	401	401
interaction	359	359
action	161	161
effect	103	103
stages	80	80
time	80	80
tapering	62	62
appearance	38	38
contraindication	34	34
ingredient	28	28
why get vaccinated	16	16
comparison	12	12
alternative	8	8
pronounce name	3	3
combination	3	3
manufacturer	2	2
availability	1	1

Table 7: *Type* distribution after Union of LiveQA, MedInfo and MedQuAD with max number of pairs limit to 500.