Overview of the MEDIQA 2019 Shared Task on Textual Inference, Question Entailment and Question Answering

Asma Ben Abacha[1] asma.benabacha@nih.gov
Chaitanya Shivade[2] cshivade@us.ibm.com
Dina Demner-Fushman[1] ddemner@mail.nih.gov
[1]LHC, NLM, Bethesda, MD [2]IBM, Almaden Research Center, San Jose, CA

Abstract

This paper presents the MEDIQA 2019 shared task organized at the ACL-BioNLP workshop. The shared task is motivated by a need to develop relevant methods, techniques and gold standards for inference and entailment in the medical domain, and their application to improve domain specific information retrieval and question answering systems. MEDIQA 2019 includes three tasks: Natural Language Inference (NLI), Recognizing Question Entailment (RQE), and Question Answering (QA) in the medical domain. 72 teams participated in the challenge, achieving an accuracy of 98% in the NLI task, 74.9% in the RQE task, and 78.3% in the QA task. In this paper, we describe the tasks, the datasets, and the participants' approaches and results. We hope that this shared task will attract further research efforts in textual inference, question entailment, and question answering in the medical domain.

1 Introduction

The first open-domain challenge in Recognizing Textual Entailment (RTE) was launched in 2005 (Dagan et al., 2005) and has prompted the development of a wide range of approaches (Bar-Haim et al., 2014). Recently, large-scale datasets such as SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018) were introduced for the task of Natural Language Inference (NLI) targeting three relations between sentences: Entailment, Neutral, and Contradiction. Few efforts have studied the benefits of RTE and NLI in other NLP tasks such as text exploration (Adler et al., 2012), identifying evidence for eligibility criteria satisfaction in clinical trials (Shivade et al., 2015), and the summarization of PMC articles (Chachra et al., 2016).

NLI can also be beneficial for Question Answering (QA). Harabagiu and Hickl (2006) presented entailment-based methods to filter and rank answers and showed that RTE can enhance the performance of open-domain QA systems and provide the inferential information needed to validate the answers. Çelikyilmaz et al. (2009) presented a graph-based semi-supervised method for QA exploiting entailment relations between questions and candidate answers and demonstrated that the use of unlabeled entailment data can improve answer ranking. Ben Abacha and Demner-Fushman (2016) noted that the requirements of question entailment in QA are different from general question similarity, and introduced the task of Recognizing Question Entailment (RQE) in order to answer new questions by retrieving entailed questions with pre-existing answers. Ben Abacha and Demner-Fushman (2019) proposed a novel QA approach based on RQE, with the introduction of the MedQuAD medical question-answer collection, and showed empirical evidence supporting question entailment for QA.

Although the idea of using entailment in QA has been introduced, research investigating methods to incorporate textual inference and question entailment into QA systems is still limited in the literature. Moreover, despite a few recent efforts to design RTE methods and datasets from MEDLINE abstracts (Ben Abacha et al., 2015) and to create the MedNLI dataset from clinical data (Romanov and Shivade, 2018), the entailment and inference tasks remain less studied in the medical domain.

MEDIQA 2019[1] aims to highlight further the NLI and RQE tasks in the medical domain, and their applications in QA and NLP. Figure 2 presents the MEDIQA tasks in the AIcrowd platform[2]. For the QA task, participants were tasked to filter and re-rank the provided answers. Reuse of the systems developed in the first and second tasks was highly encouraged.

[1] https://sites.google.com/view/mediqa2019
[2] https://www.aicrowd.com/organizers/mediqa-acl-bionlp

Proceedings of the BioNLP 2019 workshop, pages 370–379
Florence, Italy, August 1, 2019.

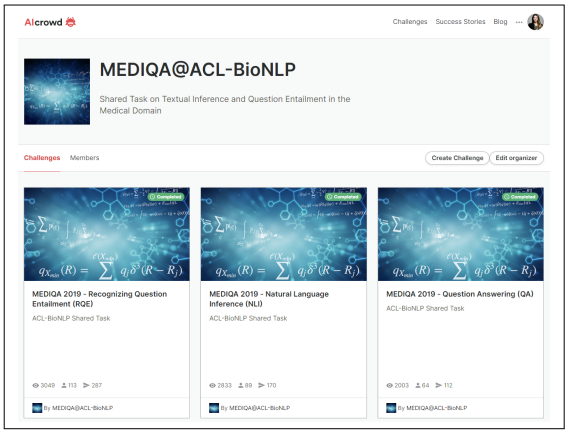


Figure 1: MEDIQA tasks on AIcrowd

2 Tasks

2.1 Natural Language Inference (NLI)

The first task focuses on Natural Language Inference (NLI) in the medical domain. We use three labels for the relation between two sentences: Entailment, Neutral and Contradiction.

2.2 Recognizing Question entailment (RQE)

The second task tackles Recognizing Question entailment (RQE) in the medical domain. We use the following definition tailored to QA: "a question A entails a question B if every answer to B is also a complete or partial answer to A" (Ben Abacha and Demner-Fushman, 2016).

2.3 Question Answering (QA)

The objective of this task is to filter and improve the ranking of automatically retrieved answers. The input ranks are generated by the medical QA system CHiQA[3]. We highly recommended the reuse of the RQE and NLI systems (first tasks). For instance (i) the RQE system could be used to retrieve answered questions (e.g. from the MedQuAD dataset[4]) that are entailed from the original questions and use their answers to validate the system's answers and re-rank them; and (ii) the NLI system could be used to identify the relations (i.e. entailment, contradiction, neutral) between the answers of the same question, as well as the answers of the questions related by the entailment relation. We encouraged all other ideas and approaches for using textual inference and question entailment to filter and re-rank the retrieved answers.

[3] https://chiqa.nlm.nih.gov/

[4] github.com/abachaa/MedQuAD

3 Data Description

3.1 NLI Datasets

The MEDIQA-NLI test set consists of 405 text-hypothesis pairs. The training set is the MedNLI dataset, which includes 14,049 clinical sentence pairs derived from the MIMIC-III database (Romanov and Shivade, 2018). Both datasets are publicly available[5].

3.2 RQE Datasets

The MEDIQA-RQE test set consists of 230 pairs of Consumer Health Questions (CHQs) received by the U.S. National Library of Medicine (NLM) and Frequently Asked Questions (FAQs) from NIH institutes. The collection was created automatically and double validated manually by medical experts. Table 1 presents positive and negative examples from the test set. The RQE training and validation sets contain respectively 8,890 and 302 medical question pairs created by (Ben Abacha and Demner-Fushman, 2016) using a collection of clinical questions (Ely et al., 2000) for the training set and pairs of CHQs and FAQs pairs for the validation set. All the RQE training, validation and test sets are publicly available[6].

3.3 QA Datasets

The MEDIQA-QA training, validation and test sets were created by submitting medical questions to the consumer health QA system CHiQA (Demner-Fushman et al., 2019), and then rating and re-ranking the retrieved answers manually by medical experts to provide reference ranks (1 to 11) and scores (4: Excellent Answer, 3: Correct but Incomplete, 2: Related, 1: Incorrect).

We provided two training sets for the QA task:

- 104 consumer health questions from the TREC-2017-LiveQA medical data (Ben Abacha et al., 2017) covering different topics such as diseases and drugs, and 839 associated answers retrieved by CHiQA and manually rated and re-ranked.

- 104 simple questions about the most frequent diseases (dataset named Alexa), and 862 associated answers.

[5] https://alpha.physionet.org/content/mednli-bionlp19/1.0.0/

[6] https://github.com/abachaa/MEDIQA2019/tree/master/MEDIQA_Task2_RQE

ID (Label)	Type	Question
Pair#1 (True)	Premise	I have a list of questions about Tay sachs disease and clubfoot 1. what is TSD/Clubfoot, and how does it effect a baby 2. what causes both? can it be prevented, treated, or cured 3. How common is TSD? how common is Clubfoot 4. How can your agency help a women/couple who are concerned about this congenital condition, and is there a cost? If you can answer these few questions I would be thankful, please get back as soon as you can.
	Hypothesis	How does congenital talipes equinovarus affect a child?
Pair#2 (True)	Premise	When and how do you know when you have congenital night blindness?
	Hypothesis	What are the symptoms of X-linked congenital stationary night blindness ?
Pair#3 (True)	Premise	Polycystic ovarian syndrome Is it possible for parents to pass this on in the genes to their children - is there any other way this can be acquired?
	Hypothesis	Can polycystic ovary syndrome be inherited ?
Pair#4 (True)	Premise	polymicrogyria. My 16 month old son has this. Does not sit up our crawl yet but still trying and is improving in grabbing things etc etc. Have read about other cases that seem 10000 time worse. It's it possible for this post of his brain to grown to normal and he grow out of it?
	Hypothesis	What is the outlook for Polymicrogyria ?
Pair#5 (False)	Premise	spina bifida; vertbral fusion;syrinx tethered cord. can u help for treatment of these problem
	Hypothesis	Does Spina Bifida cause vertebral fusion?
Pair#6 (False)	Premise	varicella shingles How can I determine whether or not I've had chicken pox. If there is a test for it, what are the results of the tests I need to know that will tell me whether or not I have had chicken pox? I want to know this to determine if I should have shingles vaccine (Zostavax) Thank you.
	Hypothesis	Who can catch shingles ?
Pair#7 (False)	Premise	Would appreciate any good info on Lewy Body Dementia, we need to get people aware of this dreadful disease, all they talk about is alzheimers. Thank you
	Hypothesis	What is alzheimer's ?
Pair#8 (False)	Premise	Can you please send me as much information as possible on hypothyroidism. I was recently diagnosed with the disease and I am struggling to figure out what it is and how I got it (...)
	Hypothesis	How is Hypothyroidism diagnosed?

Table 1: Positive and negative examples from the MEDIQA-RQE test set.

The MEDIQA-QA validation set consists of 25 consumer health questions and 234 associated answers returned by CHiQA and judged manually.

The MEDIQA-QA test set consists of 150 consumer health questions and 1,107 associated answers.

All the QA training, validation and test sets are publicly available[7].

In addition, the MedQuAD dataset of 47K medical question-answer pairs (Ben Abacha and Demner-Fushman, 2019) can be used to retrieve answered questions that are entailed from the original questions.

The validation sets of the RQE and QA tasks were used for the first (validation) round on AIcrowd. The test sets were used for the official and final challenge evaluation.

4 Evaluation

4.1 Evaluation Metrics

The evaluation of the NLI and RQE tasks was based on accuracy. In the QA task, participants were tasked to filter and re-rank the provided answers. The QA evaluation was based on accuracy, Mean Reciprocal Rank (MRR), Precision, and Spearmans Rank Correlation Coefficient (Spearman's rho).

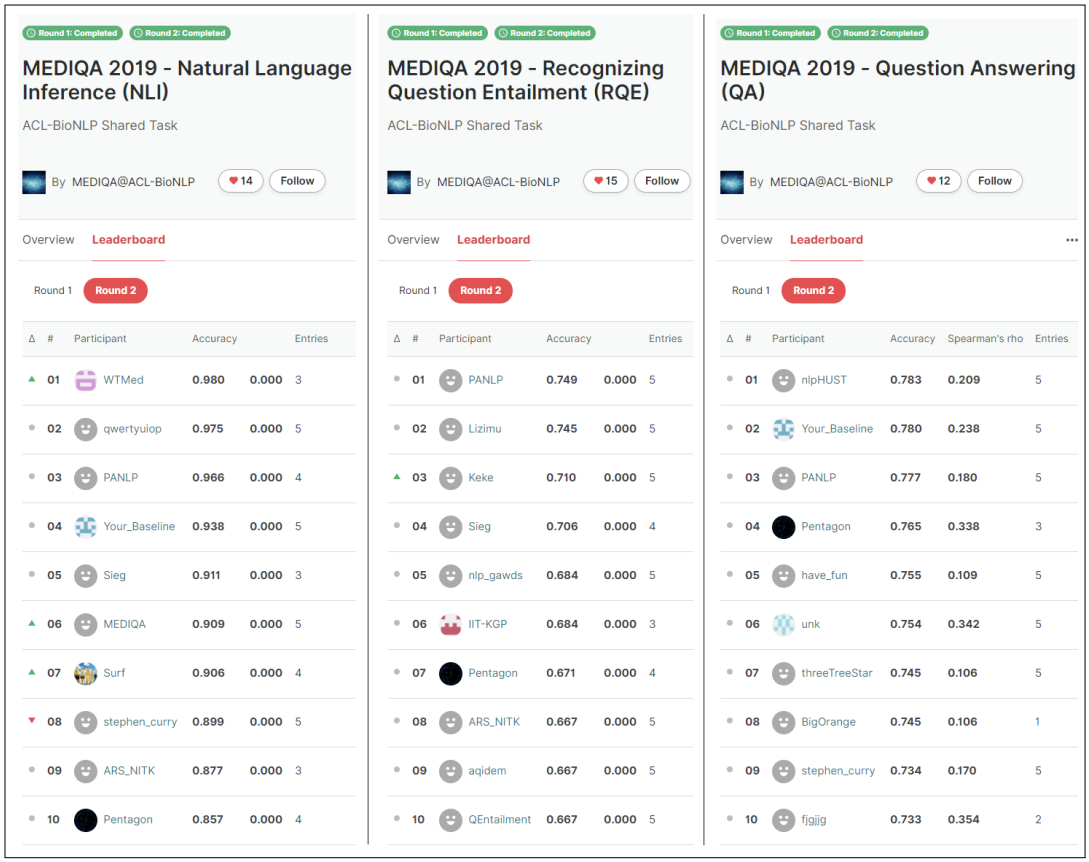
4.2 Baseline Systems

- The NLI baseline is the InferSent system (Conneau et al., 2017) based on fasttext (Bojanowski et al., 2017) word embeddings trained on the MIMIC-III data Romanov and Shivade (2018).

- The RQE baseline is a feature-based SVM classifier relying on similarity measures and semantic features (Ben Abacha and Demner-Fushman, 2016).

- The QA baseline is the CHiQA question-answering system (Demner-Fushman et al., 2019). The system was used to provide the answers for the QA task.

[7] https://github.com/abachaa/MEDIQA2019/tree/master/MEDIQA_Task3_QA



Round 1: Completed · Round 2: Completed

MEDIQA 2019 - Natural Language Inference (NLI)

ACL-BioNLP Shared Task

By MEDIQA@ACL-BioNLP · 14 · Follow

Overview · Leaderboard

Round 1 · Round 2

Δ	#	Participant	Accuracy		Entries
▲	01	WTMed	0.980	0.000	3
•	02	qwertyuiop	0.975	0.000	5
•	03	PANLP	0.966	0.000	4
•	04	Your_Baseline	0.938	0.000	5
•	05	Sieg	0.911	0.000	3
▲	06	MEDIQA	0.909	0.000	5
▲	07	Surf	0.906	0.000	4
▼	08	stephen_curry	0.899	0.000	5
•	09	ARS_NITK	0.877	0.000	3
•	10	Pentagon	0.857	0.000	4

Round 1: Completed · Round 2: Completed

MEDIQA 2019 - Recognizing Question Entailment (RQE)

ACL-BioNLP Shared Task

By MEDIQA@ACL-BioNLP · 15 · Follow

Overview · Leaderboard

Round 1 · Round 2

Δ	#	Participant	Accuracy		Entries
•	01	PANLP	0.749	0.000	5
•	02	Lizimu	0.745	0.000	5
▲	03	Keke	0.710	0.000	5
•	04	Sieg	0.706	0.000	4
•	05	nlp_gawds	0.684	0.000	5
•	06	IIT-KGP	0.684	0.000	3
•	07	Pentagon	0.671	0.000	4
•	08	ARS_NITK	0.667	0.000	5
•	09	aqidem	0.667	0.000	5
•	10	QEntailment	0.667	0.000	5

Round 1: Completed · Round 2: Completed

MEDIQA 2019 - Question Answering (QA)

ACL-BioNLP Shared Task

By MEDIQA@ACL-BioNLP · 12 · Follow

Overview · Leaderboard

Round 1 · Round 2

Δ	#	Participant	Accuracy	Spearman's rho	Entries
•	01	nlpHUST	0.783	0.209	5
•	02	Your_Baseline	0.780	0.238	5
•	03	PANLP	0.777	0.180	5
•	04	Pentagon	0.765	0.338	3
•	05	have_fun	0.755	0.109	5
•	06	unk	0.754	0.342	5
•	07	threeTreeStar	0.745	0.106	5
•	08	BigOrange	0.745	0.106	1
•	09	stephen_curry	0.734	0.170	5
•	10	fjgjjg	0.733	0.354	2

Figure 2: Top-10 results of the three tasks in MEDIQA 2019 among 72 participating teams on AIcrowd

5 Official Results

Seventy two teams participated in the challenge on the AIcrowd platform. Figure 2 presents the original top-10 scores for each task.

The official scores include only the teams who sent a working notes paper describing their approach. The accepted teams are presented in table 2. The official scores for the MEDIQA NLI, RQE, and QA tasks are presented respectively in tables 3, 4, and 5.

5.1 NLI Approaches & Results

Seventeen official teams submitted runs along with a paper describing their approaches among 43 participating teams on NLI@AIcrowd[8]. Most systems build up on the BERT model (Devlin et al., 2019). This model is pretrained on a large open-domain corpus. However, since MedNLI is from the clinical domain following variations of BERT were used.

SciBERT (Beltagy et al., 2019) is a set of variants of the original BERT trained with full text scientific articles, primarily from PubMed. Variants of the model either use the vocabulary of the original BERT model or a new vocabulary learnt specifically for this corpus.

BioBERT (Lee et al., 2019a) is initialized with the original BERT model and then pretrained on biomedical articles from PMC full text articles and PubMed abstracts. BioBERT can be fine-tuned for specific tasks like named entity recognition, relation extraction, and question answering. The data used for pretraining BioBERT is much larger (4.5B words from abstracts and 13.5B words from full text articles) than that used for SciBERT (3.1B words).

ClinicalBERT (Huang et al., 2019) is initialized with the original BERT model and then pretrained on clinical notes from the MIMIC-III dataset. Alsentzer et al. (2019) also released another resource with the same name. These are BERT and BioBERT models further pretrained on the full set of MIMIC-III notes and a subset of discharge summaries.

[8] www.aicrowd.com/challenges/mediqa-2019-natural-language-inference-nli/leaderboards

Table 2: Official teams in MEDIQA 2019 among 72 participating teams on AIcrowd

Team	Task(s)
ANU-CSIRO (Nguyen et al., 2019)	NLI, RQE, QA
ARS_NITK (Agrawal et al., 2019)	NLI, RQE, QA
DoubleTransfer (Xu et al., 2019)	NLI, RQE, QA
Dr.Quad (Bannihatti Kumar et al., 2019)	NLI, RQE, QA
DUT-BIM (Zhou et al., 2019a)	QA
DUT-NLP (Zhou et al., 2019b)	RQE, QA
IITP (Bandyopadhyay et al., 2019)	NLI, RQE, QA
IIT-KGP (Sharma and Roychowdhury, 2019)	RQE
KU_ai (Cengiz et al., 2019)	NLI
lasigeBioTM (Lamurias and Couto, 2019)	NLI, RQE, QA
MSIT_SRIB (Chopra et al., 2019)	NLI
NCUEE (Lee et al., 2019b)	NLI
PANLP (Zhu et al., 2019)	NLI, RQE, QA
Pentagon (Pugaliya et al., 2019)	NLI, RQE, QA
Saama Research (Kanakarajan, 2019)	NLI
Sieg (Bhaskar et al., 2019)	NLI, RQE
Surf (Nam et al., 2019)	NLI
UU_TAILS (Tawfik and Spruit, 2019)	NLI, RQE
UW-BHI (Kearns et al., 2019)	NLI
WTMED (Wu et al., 2019)	NLI

Table 3: Official Results of the MEDIQA-NLI Task

Rank	Team	Accuracy
1	WTMED	0.980
2	PANLP	0.966
3	DoubleTransfer	0.938
4	Sieg	0.911
5	Surf	0.906
6	ARS_NITK	0.877
7	Pentagon	0.857
8	Dr.Quad	0.855
9	UU_TAILS	0.852
10	KU_ai	0.847
11	NCUEE	0.840
12	IITP	0.818
13	MSIT_SRIB	0.813
14	uw-bhi	0.813
15	ANU-CSIRO	0.800
16	Saama Research	0.783
17	lasigeBioTM	0.724
-	*NLI-Baseline*	*0.714*

Table 4: Official Results of the MEDIQA-RQE Task

Rank	Team	Accuracy
1	PANLP	0.749
2	Sieg	0.706
3	IIT-KGP	0.684
4	Pentagon	0.671
5	ARS_NITK	0.667
5	Dr.Quad	0.667
7	DoubleTransfer	0.662
8	DUT-NLP	0.636
9	UU_TAILS	0.584
10	IITP	0.532
11	ANU-CSIRO	0.489
12	lasigeBioTM	0.485
-	*RQE-Baseline*	*0.541*

Table 5: Official Results of the MEDIQA-QA Task

Rank	Team	Accuracy	Precision	MRR	Spearman's rho
1	DoubleTransfer	0.780	0.8191	0.9367	0.238
2	PANLP	0.777	0.7806	0.9378	0.180
3	Pentagon	0.765	0.7766	0.9622	0.338
4	DUT-BIM	0.745	0.7466	0.9061	0.106
4	DUT-NLP	0.745	0.7466	0.9061	0.106
6	IITP	0.717	0.7936	0.8611	0.024
7	lasigeBioTM	0.637	0.5975	0.91	0.211
8	ANU-CSIRO	0.584	0.5568	0.7843	0.122
9	Dr.Quad	0.565	0.6679	0.6069	0.009
10	ARS_NITK	0.536	0.5596	0.6293	0.196
-	*Provided Answers*	*0.517*	*0.5167*	*0.895*	*0.315*

Another common model used by participating systems was the Multi-Task Deep Neural Network MT-DNN (Liu et al., 2019) which builds up on BERT to perform multi-task learning and is evaluated on the GLUE benchmark (Wang et al., 2018). A common theme across all the papers was training of multiple models and then using an ensemble as the final system which performed better than the individual models. Tawfik and Spruit (2019) trained 30 different models as candidates to the ensemble and experimented with various aggregation techniques. Some teams also leveraged dataset-specific properties to enhance the performance. The WTMED team (Wu et al., 2019) modeled parameters specific to the index of the text-hypothesis pair in the dataset which shows a significant boost in performance.

5.2 RQE Approaches & Results

Twelve official teams participated in MEDIQA-RQE among 53 participating teams in the second round on RQE@AIcrowd[9]. The results of the RQE task were surprisingly good knowing the challenges of the test set. For instance, positive question pairs can use different synonyms of the same medical entities (e.g. Pair#1 in table 1) and/or express differently the same information needs (e.g. Pair#4), while negative pairs can use similar language (e.g. Pair#8). Also, the test set is a realistic dataset consisting of actual consumer health questions including one or multiple sub-questions, when the training set consisted of automatically generated question pairs created from doctors' questions. This highlights the fact that several of the proposed deep networks reached relevant generalizations and abstractions of the questions.

The best results on the RQE task were obtained by the PANLP team (Zhu et al., 2019) with an approach based on multi-task learning. More specifically, their approach relied on a language model learned by the recent MT-DNN (Liu et al., 2019). In a post-processing step, they applied re-ranking heuristics based on grouping observations from the NLI and RQE datasets. E.g., for NLI the text pairs came in groups of three, where a given premise text had three counter-parts for the three relation types: entailment, neutral, and contradiction. Their heuristic re-ranking approach eliminated potential conflicts in the resuls accorcing to the group observation, and led to an increase of 5.1% in accuracy.

More generally, approaches combining ensemble methods and transfer learning of multi-task language models were the clear winners of the competition for RQE with the first and second scores (Zhu et al., 2019; Bhaskar et al., 2019). Approaches that used ensemble methods without multi-task language models (Sharma and Roychowdhury, 2019) or multi-task learning without ensemble methods (Pugaliya et al., 2019) performed worse than the first category but made it to the top 4.

Domain knowledge was also used in several participating approaches with a clear positive impact. For instance, several systems used the UMLS (Bodenreider, 2004) to expand acronyms or to replace mentions of medical entities (Bhaskar et al., 2019; Bannihatti Kumar et al., 2019). Data augmentation also played a key role for several

[9] www.aicrowd.com/challenges/mediqa-2019-recognizing-question-entailment-rqe/leaderboards

systems that used external data to extend batches of in-domain data (Xu et al., 2019), created synthetic data (Bannihatti Kumar et al., 2019), or used models trained on external datasets (e.g. MultiNLI) in ensemble methods (Bhaskar et al., 2019; Sharma and Roychowdhury, 2019).

5.3 QA Approaches & Results

Ten official teams participated in the QA task among 23 participating teams in the second round on QA@AIcrowd[10]. The relevant answer classification problem was relatively challenging with a best accuracy of 78%, however most systems did well on the first answer ranking with a best MRR of 96.22%. Precision also ranged from 79.3% to 81.9% for the six first systems. Many teams used their RQE and/or NLI models in the QA task (Bannihatti Kumar et al., 2019; Pugaliya et al., 2019; Zhu et al., 2019; Nguyen et al., 2019). The DUT-NLP team (Zhou et al., 2019b) used an adversarial multi-task network to jointly model RQE and QA.

The approach that had the best accuracy and precision in the QA task (Xu et al., 2019) relied on multi-task language models (MT-DNN) and ensemble methods. To avoid overfitting, the Double-Transfer team proposed a method, called Multi-Source, that enriches the data batches during training from external datasets by a 50% ratio and random selection. The final ensemble method further combines the Multi-Source method with pre-trained MT-DNN and SciBERT models by taking the majority vote from their predictions and resolving ties by summing the prediction probabilities for each label. The PANLP team's best run (Zhu et al., 2019) ranked second in the QA task despite the fact that the QA data do not have a group structure that could be used in re-ranking heuristics. This shows that their core model is a strong approach, and highlights further the outstanding performance of ensemble methods and multi-task language models for transfer learning for natural language understanding tasks.

Interestingly, the runs that did best on accuracy and precision did not have the best performance in terms of MRR and Spearman's rank correlation coefficient. The best team on these two metrics, Pentagon (Pugaliya et al., 2019), used the MedQuAD and the iCliniq datasets to retrieve entailed answers and used them to build more general embeddings of the considered answer. They also integrated the top-3 RQE candidates from these datasets for the considered question to build joint embeddings. The final answer embeddings were enriched with metadata such as the candidate answer source, answer length, and the original system rank. The same joint embeddings are then used in a filtering classifier for answer relevance and in a binary answer-to-answer classifier that decides if an answer is better than another. These generalized joint answer embeddings and the focus on the answer-to-answer relationship are likely to be the key elements that led to the best performance in MRR and Spearman's rho, despite the fact that the approach did not rely on the state-of-the-art ensemble models from the NLI and RQE tasks.

5.4 Multi-Tasking & External Resources

One of the aims of the MEDIQA 2019 shared task was to investigate ideas that can be reused across the three tasks. Of the twenty working notes papers, ten papers describe systems attempting more than one task. Eight papers describe systems attempting all three tasks. The multi-task nature of MEDIQA 2019 was leveraged by teams to train models such as MT-DNN (e.g. (Bannihatti Kumar et al., 2019; Xu et al., 2019; Zhu et al., 2019)). The Sieg team (Bhaskar et al., 2019) trained a model with shared layers being trained for the NLI and RQE tasks. Some teams also reused models across the three tasks. Pugaliya et al. (2019) used models developed for NLI and RQE as feature extractors in the QA task, which led to the best performance in MRR and Spearman's rho.

The shared task also encouraged the use of external resources other than the training data provided for the three tasks. Below is a non-exhaustive list resources used by various teams.

- **Abbreviation expansion** Many teams pre-processed the training data with UMLS for abbreviation expansion. While Nguyen et al. (2019) used the ADAM database (Zhou et al., 2006) for this task, Bannihatti Kumar et al. (2019) used a CAMC[11] gazzeteer.

- **External datasets** Bannihatti Kumar et al. (2019) used the Quora question pairs dataset (Shankar Iyer and Csernai, 2017) to boost the training for the RQE task, applied

[10] www.aicrowd.com/challenges/mediqa-2019-question-answering-qa/leaderboards

[11] http://www.camc.org

MetaMap[12] to recognize medical entities, and synthetically created new questions and paraphrases. Bhaskar et al. (2019) and Pugaliya et al. (2019) used the online iCliniq forum to augment training data for the RQE task. Pugaliya et al. (2019), Xu et al. (2019), Lamurias and Couto (2019), and Nguyen et al. (2019) used the MedQuAD[13] dataset of medical questions and answers (Ben Abacha and Demner-Fushman, 2019).

- **Word Embeddings** While many teams used BERT (Lamurias and Couto, 2019; Zhou et al., 2019a; Bandyopadhyay et al., 2019; Nguyen et al., 2019; Sharma and Roychowdhury, 2019)[14], some teams also used word embeddings as the input to their models. Bhaskar et al. (2019) used biomedical word embeddings from Chen et al. (2018) while Kearns et al. (2019) used cui2vec (Beam et al., 2018).

6 Conclusions

We presented the MEDIQA 2019 shared task on Natural Language Inference (NLI), Recognizing Question Entailment (RQE), and Question answering (QA) in the medical domain. The runs submitted to the challenge by 20 official teams among 72 participating teams achieved promising results and highlighted the strength of multi-task language models, transfer learning, and ensemble methods. Integrating domain knowledge and targeted data augmentation were also key factors for best performing systems. We hope that further research works and insights will be developed in the future from the MEDIQA tasks and their publicly available datasets.

Acknowledgments

This work was supported by the intramural research program at the U.S. National Library of Medicine, National Institutes of Health.
We would like to thank Sharada Mohanty, CEO and co-founder of AIcrowd, and Yassine Mrabet from the NLM for his support with the CHiQA system. We are also thankful to Vandana Mukherjee from IBM Research for supporting the project.

[12] metamap.nlm.nih.gov
[13] github.com/abachaa/MedQuAD
[14] github.com/Team-IIT-KGP/Qspider

References

Meni Adler, Jonathan Berant, and Ido Dagan. 2012. Entailment-based text exploration with application to the health-care domain. In *The 50th Annual Meeting of the Association for Computational Linguistics, System Demonstrations, 2012, Korea*.

Anumeha Agrawal, Rosa Anil George, Selvan Sunitha Ravi, Sowmya Kamath, and Anand Kumar. 2019. Ars_nitk at mediqa 2019:analysing various methods for natural language inference, recognising question entailment and medical question answering system. In *Proceedings of the BioNLP 2019 workshop, Florence, Italy, August 1, 2019*. Association for Computational Linguistics.

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

Dibyanayan Bandyopadhyay, Baban Gain, Tanik Saikh, and Asif Ekbal. 2019. Iitp at mediqa 2019: Systems report for natural language inference, question entailment and question answering. In *Proceedings of the BioNLP 2019 workshop, Florence, Italy, August 1, 2019*. Association for Computational Linguistics.

Vinayshekhar Bannihatti Kumar, Ashwin Srinivasan, Aditi Chaudhary, James Route, Teruko Mitamura, and Eric Nyberg. 2019. Dr.quad at mediqa 2019: Towards textual inference and question entailment using contextualized representations. In *Proceedings of the BioNLP 2019 workshop, Florence, Italy, August 1, 2019*. Association for Computational Linguistics.

Roy Bar-Haim, Ido Dagan, and Idan Szpektor. 2014. Benchmarking applied semantic inference: The PASCAL recognising textual entailment challenges. In *Language, Culture, Computation. Computing - Theory and Technology - Essays Dedicated to Yaacov Choueka on the Occasion of His 75th Birthday*.

Andrew L Beam, Benjamin Kompa, Inbar Fried, Nathan P Palmer, Xu Shi, Tianxi Cai, and Isaac S Kohane. 2018. Clinical concept embeddings learned from massive sources of multimodal medical data. *arXiv preprint arXiv:1804.01486*.

Iz Beltagy, Arman Cohan, and Kyle Lo. 2019. Scibert: Pretrained contextualized embeddings for scientific text. *arXiv preprint arXiv:1903.10676*.

Asma Ben Abacha, Eugene Agichtein, Yuval Pinter, and Dina Demner-Fushman. 2017. Overview of the medical question answering task at TREC 2017 LiveQA. In *Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA, November 15-17, 2017*.

Asma Ben Abacha and Dina Demner-Fushman. 2016. Recognizing question entailment for medical question answering. In *AMIA 2016, American Med-*

ical Informatics Association Annual Symposium, Chicago, IL, USA, November 12-16, 2016.

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

Asma Ben Abacha, Duy Dinh, and Yassine Mrabet. 2015. Semantic analysis and automatic corpus construction for entailment recognition in medical texts. In *Artificial Intelligence in Medicine - 15th Conference on Artificial Intelligence in Medicine, AIME 2015, Pavia, Italy, June 17-20, 2015*.

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 BioNLP 2019 workshop, Florence, Italy, August 1, 2019*. Association for Computational Linguistics.

Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database-Issue):267–270.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

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, EMNLP 2015, Lisbon, Portugal*.

Asli Çelikyilmaz, Marcus Thint, and Zhiheng Huang. 2009. A graph-based semi-supervised learning for question-answering. In *ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2-7 August 2009, Singapore*.

Cemil Cengiz, Ula Sert, and Deniz Yuret. 2019. Ku_ai at mediqa 2019: Domain-specific pre-training and transfer learning for medical nli. In *Proceedings of the BioNLP 2019 workshop, Florence, Italy, August 1, 2019*. Association for Computational Linguistics.

Suchet K. Chachra, Asma Ben Abacha, Sonya E. Shooshan, Laritza Rodriguez, and Dina Demner-Fushman. 2016. A hybrid approach to generation of missing abstracts in biomedical literature. In *COLING 2016, 26th International Conference on Computational Linguistics, 2016, Osaka, Japan*.

Qingyu Chen, Yifan Peng, and Zhiyong Lu. 2018. Biosentvec: creating sentence embeddings for biomedical texts. *arXiv preprint arXiv:1810.09302*.

Sahil Chopra, Ankita Gupta, and Anupama Kaushik. 2019. Msit_srib at mediqa 2019: Knowledge directed multi-task framework for natural language inference in clinical domain. In *Proceedings of the BioNLP 2019 workshop, Florence, Italy, August 1, 2019*. Association for Computational Linguistics.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges, Evaluating Predictive Uncertainty, Visual Object Classification and Recognizing Textual Entailment, First PASCAL Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK*.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*.

John W. Ely, Jerome A. Osheroff, Paul N. Gorman, Mark H. Ebell, M. Lee Chambliss, Eric A. Pifer, and P. Zoe Stavri. 2000. A taxonomy of generic clinical questions: classification study. *British Medical Journal*, 321:429–432.

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

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv:1904.05342*.

Kamal raj Kanakarajan. 2019. Saama research at mediqa 2019: Pre-trained biobert with attention visualisation for medical natural language inference. In *Proceedings of the BioNLP 2019 workshop, Florence, Italy, August 1, 2019*. Association for Computational Linguistics.

William Kearns, Wilson Lau, and Jason Thomas. 2019. Uw-bhi at mediqa 2019: An analysis of representation methods for medical natural language inference. In *Proceedings of the BioNLP 2019 workshop, Florence, Italy, August 1, 2019*. Association for Computational Linguistics.

Andre Lamurias and Francisco Couto. 2019. Lasigebiotm at mediqa 2019: Biomedical question answering using bidirectional transformers and named entity recognition. In *Proceedings of the BioNLP 2019 workshop, Florence, Italy, August 1, 2019*. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019a. Biobert: pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*.

Lung-Hao Lee, Yi Lu, Po-Han Chen, Po-Lei Lee, and Kuo-Kai Shyu. 2019b. Ncuee at mediqa 2019: Medical text inference using ensemble bert-bilstm-attention model. In *Proceedings of the BioNLP 2019 workshop, Florence, Italy, August 1, 2019*. Association for Computational Linguistics.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*.

Jiin Nam, Seunghyun Yoon, and Kyomin Jung. 2019. Surf at mediqa 2019: Improving performance of natural language inference in the clinical domain by adopting pre-trained language model. In *Proceedings of the BioNLP 2019 workshop, Florence, Italy, August 1, 2019*. Association for Computational Linguistics.

Vincent Nguyen, Sarvnaz Karimi, and Zhenchang Xing. 2019. Anu-csiro at mediqa 2019: Question answering using deep contextual knowledge. In *Proceedings of the BioNLP 2019 workshop, Florence, Italy, August 1, 2019*. Association for Computational Linguistics.

Hemant Pugaliya, Karan Saxena, Shefali Garg, Sheetal Shalini, Prashant Gupta, Eric Nyberg, and Teruko Mitamura. 2019. Pentagon at mediqa 2019: Multi-task learning for filtering and re-ranking answers using language inference and question entailment. In *Proceedings of the BioNLP 2019 workshop, Florence, Italy, August 1, 2019*. Association for Computational Linguistics.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 2018*.

Nikhil Dandekar Shankar Iyer and Kornl Csernai. 2017. First quora dataset release: Question pairs.

Prakhar Sharma and Sumegh Roychowdhury. 2019. Iit-kgp at mediqa 2019: Recognizing question entailment using sci-bert stacked with a gradient boosting classifier. In *Proceedings of the BioNLP 2019 workshop, Florence, Italy, August 1, 2019*. Association for Computational Linguistics.

Chaitanya Shivade, Courtney Hebert, Marcelo A. Lopetegui, Marie-Catherine de Marneffe, Eric Fosler-Lussier, and Albert M. Lai. 2015. Textual inference for eligibility criteria resolution in clinical trials. *Journal of Biomedical Informatics*, 58.

Noha Tawfik and Marco Spruit. 2019. Uu_tails at mediqa 2019: Learning textual entailment in the medical domain. In *Proceedings of the BioNLP 2019 workshop, Florence, Italy, August 1, 2019*. Association for Computational Linguistics.

Alex Wang, Amapreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA*.

Zhaofeng Wu, Yan Song, Sicong Huang, Yuanhe Tian, and Fei Xia. 2019. Wtmed at mediqa 2019: A hybrid approach to biomedical natural language inference. In *Proceedings of the BioNLP 2019 workshop, Florence, Italy, August 1, 2019*. Association for Computational Linguistics.

Yichong Xu, Xiaodong Liu, Chunyuan Li, Hoifung Poon, and Jianfeng Gao. 2019. Doubletransfer at mediqa 2019: Multi-source transfer learning for natural language understanding in the medical domain. In *Proceedings of the BioNLP 2019 workshop, Florence, Italy, August 1, 2019*. Association for Computational Linguistics.

Huiwei Zhou, Bizun Lei, Zhe Liu, and Zhuang Liu. 2019a. Dut-bim at mediqa 2019: Utilizing transformer network and medical domain-specific contextualized representations for question answering. In *Proceedings of the BioNLP 2019 workshop, Florence, Italy, August 1, 2019*. Association for Computational Linguistics.

Huiwei Zhou, Weihong Yao Xuefei Li, Chengkun Lang, and Shixian Ning. 2019b. Dut-nlp at mediqa 2019:an adversarial multi-task network to jointly model recognizing question entailment and question answering. In *Proceedings of the BioNLP 2019 workshop, Florence, Italy, August 1, 2019*. Association for Computational Linguistics.

Wei Zhou, Vetle I Torvik, and Neil R Smalheiser. 2006. Adam: another database of abbreviations in medline. *Bioinformatics*, 22(22):2813–2818.

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 BioNLP 2019 workshop, Florence, Italy, August 1, 2019*.