Sieg at MEDIQA 2019: Multi-task Neural Ensemble for Biomedical Inference and Entailment

Sai Abishek Bhaskar*, **Rashi Rungta***, **James Route, Eric Nyberg, Teruko Mitamura**
Language Technologies Institute, Carnegie Mellon University
{sabhaska, rashir, jroute, ehn, teruko}@cs.cmu.edu

Abstract

This paper presents a multi-task learning approach to natural language inference (NLI) and question entailment (RQE) in the biomedical domain. Recognizing textual inference relations and question similarity can address the issue of answering new consumer health questions by mapping them to Frequently Asked Questions on reputed websites like the NIH[1]. We show that leveraging information from parallel tasks across domains along with medical knowledge integration allows our model to learn better biomedical feature representations. Our final models for the NLI and RQE tasks achieve the 4^{th} and 2^{nd} rank on the shared-task leaderboard respectively.

1 Introduction

The MEDIQA challenge (Abacha et al., 2019) aims to improve textual inference and entailment in the medical domain to build better domain-specific Information Retrieval and Question Answering systems. There are three subtasks (NLI, RQE, QA), out of which we focus on -

1. **Natural Language Inference (NLI):** Identifying the three types of inference relations (Entailment, Neutral and Contradiction) between two sentences.
2. **Recognizing Question Entailment (RQE):** Predicting entailment between two questions (if every answer for question 1 is at least a partial answer for question 2) in the context of QA.

The task is motivated by the need to explore and develop better question answering systems in the medical domain. Identifying the type of correlation between questions as well as medical sentences will help the biomedical community cope with the increasing number of consumer health questions posted on community question answering websites, many of which have already been asked before and can easily be answered by linking them with a previously answered question by an expert.

In this paper, we start with a discussion of the previous work done on multi-task learning and textual inference and entailment in the biomedical domain in Section 2, followed by the dataset description in Section 3. The baselines and our proposed approach are detailed in Section 4 and 5 respectively. We conclude with the discussion of our results in Section 6 and a detailed error analysis in Section 7.

2 Related Work

2.1 Multi-Task Learning

Multi-task Learning (MTL) is inspired by the idea that it is useful to jointly learn multiple related tasks so that the knowledge gained in one task can benefit other tasks. Recently, there is growing interest in using deep neural networks (DNNs) to apply MTL to representation learning (Collobert et al. 2011, Liu et al. 2017). MTL provides an effective way to use supervised data from a number of related tasks and also provides for a regularization effect by not overfitting to a specific task, thus making the learned representations more robust.

2.2 Biomedical Textual Inference

The initial approaches for predicting inference relations between two sentences in the medical domain involved several neural architectures. (Ro-

*denotes equal contribution
[1] https://www.nih.gov/

Proceedings of the BioNLP 2019 workshop, pages 462–470
Florence, Italy, August 1, 2019.

manov and Shivade, 2018) details the curation of the MedNLI dataset, and describes multiple baseline approaches. A Feature-based, Bag-of-Words (BOW), the ESIM model (Chen et al., 2016) and the InferSent model (Conneau et al., 2017) being among them.

2.3 Biomedical Question Entailment

The initial work (Ben Abacha and Demner-Fushman, 2017), in addition to creating the working dataset for RQE, uses handcrafted lexical and semantic features as an input to traditional machine learning models like SVM, Logistic Regression, and Naive Bayes for question entailment in the clinical domain. The lexical features include word overlap, bigram similarity and best similarity from a set of 5 similarity measures (Levenshtein, Jaccard, Cosine, Bigram, Word Overlap) while semantic features include the number of overlapping medical entities and problems based on a CRF classifier trained across different corpora. Ben Abacha and Demner-Fushman 2019 use question analysis based features such as question type and focus recognition which helps identify the different focus points of consumer health questions such as information, symptoms, or treatments based on specific trigger words.

3 Datasets

3.0.1 NLI

MedNLI (Romanov and Shivade) is a dataset annotated by doctors for NLI in the clinical domain. It is available through the MIMIC-III derived data repository.

- Train: 11232 sentence pairs
- Validation: 1395 sentence pairs
- Test: 1422 sentence pairs
- Test (Leaderboard): 230 sentence pairs

Labels: {contradiction, entailment, neutral}
Evaluation Metric: Accuracy

Since the train, validation and test sets are from the same distribution, we combined them and took a subset of 90% to be the new training set and the rest 10% to be the held-out validation set.

3.0.2 RQE

The RQE dataset comprises of consumer health questions (CHQs) received by the National Library of Medicine and frequently asked questions (FAQs) collected from the National Institutes of Health (NIH) websites (Ben Abacha and Demner-Fushman, 2017).

- Training Set: 8,588 medical question pairs
- Test: 302 medical question pairs
- Test Set (Leaderboard): 230 medical question pairs

Labels: {true, false}
Evaluation Metric: Accuracy

On further analysis of the RQE train and test data, we found that the two datasets come from different distributions. The CHQs in the training set follow a more formal third person based language structure while CHQs in the test set are verbose with more colloquial language phrases. For example, a CHQ from the training set is - "How should I treat polymenorrhea in a 14-year-old girl?" while a CHQ from the test set is - "lupus. Hi, I want to know about Lupus and its treatment. Best, Mehrnaz".
In light of this, we modify our training set to contain 302 examples from the original training set, all the 302 examples in the test set and 930 questions from *icliniq* as explained in section 5.1.3. As with NLI, we took a subset of 90% to be the new training set and the rest 10% to be the held-out validation set.

4 Baselines

4.1 NLI

InferSent (Romanov and Shivade, 2018) is a sentence encoder model that has given near state-of-the-art results across the NLP (including NLI) and computer vision domains. For the MedNLI dataset, the model uses a Bi-directional LSTM with domain knowledge incorporated through retrofitting and attention. We use this InferSent model as our baseline for the NLI task. A re-implementation using data preprocessed with UMLS (5.2.3) and abbreviation expansion (5.2.5), along with different word embeddings (5.2.2) gives a slight bump in the accuracy value.

InferSent	Accuracy	Embeddings
Reported	78.3	MIMIC FastText
Re-implementation	79.3	PubMed MIMIC FastText

Table 1: Baseline accuracy values for NLI dev set

4.2 RQE

The SVM model described in Ben Abacha and Demner-Fushman 2017 is our RQE baseline. The input features are detailed in 2.3 and the corresponding metrics are shown in Table 2.

	P	R	F
SVM	75.0	75.2	75.0

Table 2: Baseline precision, recall and F1 values for RQE

5 Proposed Approach

5.1 Additional Datasets

Our hypothesis is that these parallel datasets will help our multi-task neural model capture salient biomedical features to help our main NLI and RQE tasks.

5.1.1 PubMed RCT

The Pubmed RCT dataset contains 2.3m sentences from 200k PubMed abstracts of randomized controlled trial (RCT) articles. We use the smaller subset of the sentences from 20k abstracts. The sentences are labeled based on their role in the abstract which belongs to one of the following five classes: background, objective, method, result, or conclusion. This single sentence classification is a parallel dataset for the NLI task.

5.1.2 MultiNLI

The MultiNLI dataset (Williams et al., 2017) contains 433k sentences which have been annotated with textual entailment information. This textual inference classification corpus forms one of the parallel datasets for the NLI task.

5.1.3 *icliniq.com* Questions

Given the limited size of the RQE dataset, we looked for ways to augment our data with additional examples from the same distribution.
We use data scraped from *icliniq.com*, which is an online doctor consultation platform. The website has a format where each question has a summary question, followed by the entire text entered by the user. We take the summary question to be the FAQ and the question text as the CHQ corresponding to the RQE task. 465 question pairs were scraped (Regin, 2017) and an equal number of negative examples is generated through negative sampling. This gives us a total of 930 additional question pairs. An example from icliniq is:

Q1 (CHQ): Hello doctor, I do not have a white half moon on my nails. Is there any thyroid issue? If yes, please suggest some treatment."
Q2 (FAQ): Does the absence of the white half moon on nails indicate a thyroid problem?
Gold Label: True

5.1.4 GARD Question Type

The dataset released by the Genetic and Rare diseases information center (Roberts et al., 2014) allows our model to learn question type information necessary for the RQE task. It contains 3137 questions each of which has one of 13 unique labels. Since the question type is an important handcrafted feature while considering traditional ML approaches for the RQE task, we use this dataset so that our multi-task model can leverage this information. The merit of this approach is shown in Table 3.

5.1.5 Quora Question Pairs

The Quora Question Pairs dataset (Quora, 2017) contains more than 400k duplicate question pairs released by Quora, a popular community QA website. We hypothesize that using this as a parallel dataset for the RQE task will help us generalize better since Quora users adopt an informal and colloquial form of language which is similar to the language of CHQs.

5.2 Domain Knowledge Integration and Preprocessing

5.2.1 ScispaCy

We use ScispaCy (Neumann et al., 2019), a tool for practical biomedical/scientific text processing, based on the spaCy library to preprocess and incorporate domain knowledge in the NLI and RQE datasets. Its use is detailed in the subsequent sections.

5.2.2 Biomedical Word Vectors

We use the biomedical word vectors released by the NCBI BioNLP Research Group (Chen et al., 2018) as the word embeddings for the InferSent model for the NLI task. Fasttext (Bojanowski et al., 2017) was used to train 200-dimensional word vectors on PubMed abstracts and MIMIC III clinical notes.

5.2.3 UMLS Metamap

We use a python wrapper for UMLS Metamap (Aronson and Lang, 2010), called pyMetamap[2] to extract preferred names and CUIs (Concept Unique Identifiers) for medical entities from the UMLS Metathesaurus (Bodenreider, 2004). As a pre-processing step, we identify medical terms in the data using ScispaCy, and replace them with their *preferred_name* occurring with the highest score in UMLS.
Using ScispaCy helps us by acting as a filter against common terms like *patient* and *lab*, which would otherwise get identified to be medical entities.

In cases where the preferred name for a medical entity was exactly the term itself, we used the additional dataset MRCON (Rogers et al., 2012) to extract all entity names with the same CUI as the one for the entity identified initially. We created a set of these synonymous entities and picked the one which had the highest semantic similarity to the medical entities identified in the parallel sentence/question. We then append this identified synonymous entity's name to where the originally identified entity was found in the first sentence/question.

[2] https://github.com/AnthonyMRios/pymetamap/

5.2.4 DrugBank

DrugBank (Wishart et al., 2017) is a bioinformatics and cheminformatics dataset containing detailed drug data for more than 12k drugs along with their synonyms, parent medical categories (i.e. what kind of drug it is) and pharmacological information.
Our use of DrugBank to augment the RQE and NLI datasets with domain knowledge is as follows:

- We load SciSpacy with two pretrained Spacy models. The first is a NER model trained on the BC5CDR corpus to identify drug names and the second is a general pipeline for biomedical data.
- From the first sentence, we extract drug names using the first SciSpacy model.
- From the second sentence in the particular sentence-pair, we extract biomedical terms and search for a string overlap with the relevant drug information from the Drugbank dataset.
- If a particular phrase exists in the drug information, we append this phrase after the drug name in the first sentence.

5.2.5 Abbreviation expansion

We use the Recognizing Abbreviation Definitions dataset (S Schwartz and Hearst, 2003) to construct an initial dictionary. To further augment it, we use the CAMC (Charleston Area Medical Center) medical word list[3]. In order to get an extended dictionary which took into account the several newly created acronyms, or those which are more colloquial than formal, we scraped the medical abbreviation Wikipedia pages and appended this to our dictionary. If more than one medical phrase was found for an abbreviation, we gave preference to the first one. On manual combing of the thus created dictionary, we edited/deleted entries which felt incorrect. For example, *FS* which was being mapped to *Flow Sheet* was changed to *Fingerstick*.
As one of the preprocessing steps, ScispaCy is used to identify abbreviations in the text which are then appended with their corresponding expanded medical term.

[3] https://www.camc.org/documents/patientlink/Abbreviations-List.pdf

5.2.6 Bio-BERT

BioBERT (Lee et al., 2019) uses the pretrained BERT base model and finetunes it for the biomedical domain by further training on PubMed abstracts and PMC full-text articles. We converted the Tensorflow version of the saved model weights to PyTorch using the PyTorch pretrained BERT library. The three variants of the BioBERT model based on the data used to finetune it are-

- PubMed abstracts (4.5B words)
- PMC full-text articles (13.5B words)
- Both PubMed abstracts and PMC full-text articles

The latter variant outperforms single dataset trained BioBERT with respect to most of the biomedical named entity recognition datasets but has mixed results for the relation extraction and question answering datasets as mentioned in (Lee et al., 2019).
We use the PubMed+PMC BioBERT v1.0 model (cased vocabulary) to initialize our MT-DNN architecture.

5.2.7 SciBERT

SciBERT (Beltagy et al., 2019), is another BERT based model for the scientific and biomedical domain which outperforms BioBERT by an average of 0.51 F1 score at biomedical named entity recognition, text classification and relation classification. It was trained on 1.14M papers from Semantic Scholar (Ammar et al., 2018) of 18% is from the computer science domain and 82% is from the biomedical domain. The full text of the papers are used, not just the abstracts.

There are four variants of SciBERT -

- Cased or Uncased
- BERT-Base vocab or scivocab (30k words, having a 42% overlap with BERT-Base vocab)

We use the recommended uncased scivocab SciBERT model to initialize our MT-DNN architecture. Our final model ensemble consists of SciBERT in addition to BioBERT as the two models were trained on different datasets and hence they will be able to capture different salient features of biomedical knowledge.

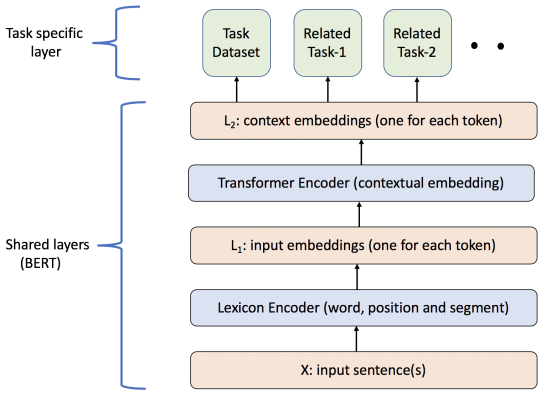

Figure 1: Architecture of the multi-task MT-DNN model

Datasets	Test Accuracy
RQE	58.2
RQE + GARD Question Type (GARD)	62.6
RQE + Quora Question Pairs (QQP)	66.0
RQE + QQP + GARD	66.0

Table 3: Parallel dataset results (values obtained post the shared task completion) for the RQE task using the MT-DNN base model.

5.3 Model

We are interested in leveraging multi-task learning across different datasets to improve the learning of the biomedical text representations. For the current work, we use the Multi-Task Deep Neural Networks for Natural Language Understanding (MT-DNN) introduced in Liu et al. 2019, which demonstrates the effectiveness of multi-task learning by beating the state-of-art on eight out of nine GLUE benchmark tasks (Wang et al., 2019). The architecture of our MT-DNN model is shown in Figure 1. Both the NLI and RQE tasks share the lower layers, while the top layers represent task-specific outputs. The input X, which is a word sequence (biomedical question for RQE and sentence text for NLI) is first represented as a sequence of embedding vectors, one for each word, in L_1. Then the transformer encoder captures the contextual information for each word via self-attention and generates a sequence of contextual embeddings in L_2. This is the shared semantic representation that is trained by the multiple task objectives. The lexicon encoder (L_1) and transformer encoder (L_2) pre-training involves the approach introduced in the BERT model (Devlin et al., 2018).

Model	Datasets	Domain Knowledge	Test Accuracy
InferSent Baseline (Romanov and Shivade 2018)	NLI (train set only)	UMLS	71.4
MT-DNN + MT-DNN(BioBERT)	NLI	UMLS	83.5
MT-DNN + MT-DNN(BioBERT) + MT-DNN(SciBERT)	NLI + MultiNLI + PubMed 20k RCT	UMLS	87.2
MT-DNN + MT-DNN(BioBERT) + MT-DNN(SciBERT) + InferSent	NLI + MultiNLI + PubMed 20k RCT	UMLS + DrugBank + Abbreviation Expansion	**91.1**

Table 4: Results for the NLI Task

Model	Datasets	Domain Knowledge	Test Accuracy
SVM Baseline (Ben Abacha and Demner-Fushman 2017)	RQE (train set only)	biomedical NER	54.1
MT-DNN + MT-DNN(SciBERT)	RQE	UMLS + DrugBank + Abbreviation Expansion	65.8
MT-DNN + MT-DNN(SciBERT)	RQE + GARD Question Type	UMLS + DrugBank + Abbreviation Expansion	66.7
MT-DNN + MT-DNN(BioBERT) + MT-DNN(SciBERT)	RQE + Quora Question Pairs + GARD Question Type	UMLS + DrugBank + Abbreviation Expansion	**70.6**

Table 5: Results for the RQE Task

5.3.1 Implementation details

The BERTAdam optimizer with a learning rate of 5e-5, batch size of 32, linear learning rate decay schedule with warm-up over 0.1 and gradient clipping is used. These hyperparameters are in accordance with those proposed in the MT-DNN work (Liu et al. 2019). In each epoch, a mini-batch from all the parallel datasets is taken and the model is updated.

The training procedure of the model consists of two stages: pretrained BERT model loading and multi-task fine-tuning. We use BioBERT (5.2.6), SciBERT (5.2.7) and the MT-DNN base model (pretrained on the GLUE benchmark tasks) to initialize our MT-DNN model variants.

6 Experiments and Results

The accuracy values obtained on the shared task's leaderboard are listed in Table 4 and Table 5 for the NLI and the RQE task respectively.

For the NLI task, Table 4, we see that an ensemble of the MT-DNN base model along with MT-DNN initialized with SciBERT and BioBERT keeping PubMed RCT and MultiNLI as the parallel datasets achieved a better accuracy than using only the NLI dataset with an MT-DNN base model, BioBERT ensemble.
To account for missing drug information and the lack of biomedical context around abbreviations in the input data, we preprocess our dataset by expanding medical abbreviations (5.2.5) and including DrugBank (5.2.4) information.

We see that taking a four-way ensemble of the MT-DNN base model, MT-DNN initialized with BioBERT, SciBERT and InferSent along with a three-pronged domain knowledge inclusion with MultiNLI and PubMed RCT as the parallel datasets gave us the best result of **91.1%** on the leaderboard. Our hypothesis behind this model ensemble was that since BioBERT and SciBERT are trained on different datasets, they will capture different features and hence taking an ensemble of these two models along with InferSent based on majority confidence scores will help us achieve a better accuracy than a single model. Our Infersent re-implementation results are shown in Table 1.

To demonstrate the usefulness of parallel datasets for the RQE task and for easy comparison with the results on the leaderboard (Table 5), we measure the test accuracy for different dataset combinations using the test dataset labels released by the task organizers post completion of the shared task. These results are shown in Table

Category	Premise	Hypothesis	Predicted	Gold Label
Numeric Reasoning	On transfer, patient VS were 102, 87/33, 100% on 60% 450 x 18 PEEP 5.	The patient's vitals were normal on transfer	neutral	contradiction
	Was given a 500cc bolus and responded to 89/50.	The patient was hypotensive.	neutral	entailment
	His initial BP at OSH 130/75, down to 93/63 after nitro.	The patient was initially normotensive.	contradiction	entailment
Inconclusive cases	The pt was discharged home [**2188-5-3**].	the patient was discharged with home medications	entailment	neutral
	On the floor, he is doing relatively well.	The patient is stable.	entailment	neutral
	His symptoms occur about every day to every other day and have been stable over the past year.	His symptoms are severe.	contradiction	neutral

Table 6: Error types observed during the qualitative analysis for the NLI Task

Category	Q1 (CHQ)	Q2 (FAQ)	Predicted	Gold Label
Understanding	milroy disease hello , my daughter has lymph edema her both legs and left hand is swelling , this problem started when she was of 3 months now she is 16 months , her swelling is growing day by day , im clue less what to do and what kind of treatment i should do plz help and suggest us	Is walking good for lymphedema?	true	false
	If oleandor was ingested by touching the plant stems inner part and then directly eating without washing hands, how long would u exspect symptoms would start? And how severe would you say symptoms may get.	What are the symptoms of Oleander poisoning?	false	true
Multiple Questions	more information in relation to Ellis van creveld syndrome Specifically in later life can they have children has it ever been reported any researchcarried out and just as much information as possible to help my understanding of what I have Many thanks	What is Ellis-van Creveld syndrome?	true	false
	Achondroplasia research. Hello, We are students from [LOCATION] and we are doing a biology project of genetic diseases. We chose Achondroplasia as our disease to research. We have a few question and we are hoping you could answer them. Our questions are, can you tell if your child will have Achondroplasia when you are pregnant? When do people usually come in when they think something isn't right with their child? what are the worse cases of Achondroplasia you've ever seen? Thank you in advance. sincerely, [NAME]	How to diagnose Achondroplasia?	false	true

Table 7: Error types observed during the qualitative analysis for the RQE Task

3. We see that using only the RQE dataset got us an accuracy of 58.2% while using the GARD question type decomposition and Quora Question Pairs increased our accuracy by 4.3% and 7.8% respectively.

Building on the observation of variation in performance of the different parallel datasets, we see that having GARD question types as the parallel dataset gives us a slight boost in accuracy from 65.8% to 66.7% as shown in Table 5. Our best result of **70.6%** is obtained when we take an ensemble of the MT-DNN base model along with MT-DNN initialized with BioBERT and SciBERT, keeping Quora Question Pairs and GARD Question Type as the parallel datasets.

7 Error Analysis

7.1 NLI

Equivalent to the error analysis in Romanov and Shivade 2018, we present some of the representative examples from the Test set (using the gold labels released by the task organizers) in Table 6. We broadly classify them into categories we felt they were closest to.

7.1.1 Numeric values

Example pairs where the premise is solely based on *numeric values* describing the patient's vitals are often classified incorrectly due to the several variations in the values used across examples. This can be seen in Example 1, 2 and 3 from the table. Most of such examples are often incorrectly predicted to be neutral by our model.

7.1.2 Inconclusive cases

We also come across examples where the sentences are not entirely conclusive, but the model assumes them to be, hence making an incorrect prediction. These examples are clubbed under the *Inconclusive cases* category.

Consider the case of Example 5 from Table 6, the hypothesis claims the patient to be stable, while the premise does not state this explicitly,

thus leaving a margin for a less definite hypothesis. Our model predicts *entailment* for this pair, when the expected label is *neutral*.

7.2 RQE

Table 7 shows a few examples representative of the two broad categories of errors observed in the Test set (using the gold labels released by the task organizers) for the RQE task.

7.2.1 Understanding

The *CHQ* from Example 1 in the table is asking for treatment suggestions for the condition *lymphedema*, and the *FAQ* is a question verifying if walking is good for *lymphedema*. The expected label is *false*, while the model predicts it to be *true*. The two questions are semantically different because of which one does not entail the other, but the model might be confusing a suggestive question (*FAQ* in this example) to be a part of the broader question (*CHQ*) thus failing to understand the subtle difference between the two.

In Example 2, the *CHQ* asks about two questions related to the symptoms - how long they will take to occur, and how severe they would get. The *FAQ* inquires about what the symptoms are. These questions have the same focus, but could be understood as being different when compared semantically. However, since the answer to the *FAQ* might partially answer the *CHQ*, the expected label is true, while our model predicted this as false.

7.2.2 Multiple Questions

The other kind of errors we observed were when the *CHQs* had multiple questions within them. For instance in Example 3, in the *CHQ* the user seems to have decent knowledge about the said syndrome and wants more in-depth knowledge on the subject. The repeated questions about more *information* might have misled the model into predicting this as true, when the expected label was false.

In Example 4, we see that the several questions contained in the *CHQ* confuse our classifier to predict false when the *FAQ* is actually entailed.

8 Future Work

Going forward, this work could be improved by more intensive domain knowledge incorporation. To start with, using medical side effects relations from SIDER (Kuhn et al., 2015) and leveraging the ontology relations in UMLS (Bodenreider, 2004) would be appropriate steps to strengthen the proposed system. We would like to thank our anonymous reviewers for these inputs.

A large part of the success of this work can be attributed to preprocessing the input data to incorporate biomedical knowledge which, at the same time makes it harder to generalize this pipeline to other domains. Therefore, investigating the performance of our proposed approach in non-biomedical domains by training with different parallel datasets to enforce generalization is an interesting avenue for future research.

9 Conclusion

In this paper, we investigate various preprocessing pipelines along with parallel dataset combinations in a multi-task learning setup for efficient language processing in the biomedical domain. We demonstrate the effectiveness of using transformer based neural models for predicting natural language inference and recognizing question entailment in the medical domain which beat the baselines (as shown in Table 4 and Table 5) by a margin of 19.7% and 16.5% for the NLI and RQE tasks respectively.

References

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. *ACL-BioNLP 2019*.

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew E. Peters, Joanna Power, Sam Skjonsberg, Lucy Lu Wang, Chris Wilhelm, Zheng Yuan, Madeleine van Zuylen, and Oren Etzioni. 2018. Construction of the literature graph in semantic scholar. *CoRR*, abs/1805.02262.

Alan Aronson and Franois-Michel Lang. 2010. An overview of metamap: Historical perspective and re-

cent advances. *Journal of the American Medical Informatics Association : JAMIA*, 17:229–36.

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

Asma Ben Abacha and Dina Demner-Fushman. 2017. Recognizing question entailment for medical question answering. *American Medical Informatics Association*.

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

Olivier Bodenreider. 2004. The unified medical language system (umls): Integrating biomedical terminology. *Nucleic acids research*, 32.

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.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, and Hui Jiang. 2016. Enhancing and combining sequential and tree LSTM for natural language inference. *CoRR*, abs/1609.06038.

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

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel P. Kuksa. 2011. Natural language processing (almost) from scratch. *CoRR*, abs/1103.0398.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *CoRR*, abs/1705.02364.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Michael Kuhn, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. 2015. The SIDER database of drugs and side effects. *Nucleic Acids Research*, 44(D1):D1075–D1079.

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.

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

Xiaodong Liu, Yelong Shen, Kevin Duh, and Jianfeng Gao. 2017. Stochastic answer networks for machine reading comprehension. *CoRR*, abs/1712.03556.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. Scispacy: Fast and robust models for biomedical natural language processing.

Quora. 2017. Quora question pairs - kaggle. *Kaggle.com*.

Lasse Regin. 2017. Medical question answer data. *https://github.com/LasseRegin/medical-question-answer-data*.

Kirk Roberts, Halil Kilicoglu, Marcelo Fiszman, and Dina Demner-Fushman. 2014. Decomposing consumer health questions. *Proceedings of BioNLP 2014*, pages 29–37.

Willie Rogers, Francois-Michel Lang, and Cliff Gay. 2012. Metamap data file builder. *metamap.nlm.nih.gov*, page 6.

Alexey Romanov and Chaitanya Shivade. Lessons from natural language inference in the clinical domain.

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

Ariel S Schwartz and Marti Hearst. 2003. A simple algorithm for identifying abbreviation definitions in biomedical text. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 4:451–62.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In the Proceedings of ICLR.

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

David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, Nazanin Assempour, Ithayavani Iynkkaran, Yifeng Liu, Adam Maciejewski, Nicola Gale, Alex Wilson, Lucy Chin, Ryan Cummings, Diana Le, Allison Pon, Craig Knox, and Michael Wilson. 2017. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research*, 46(D1):D1074–D1082.