ARS_NITK at MEDIQA 2019:Analysing Various Methods for Natural Language Inference, Recognising Question Entailment and Medical Question Answering System

Anumeha Agrawal, Rosa Anil George, Selvan Sunitha Ravi, Sowmya Kamath S and Anand Kumar M
Department of Information Technology
National Institute of Technology Karnataka, Surathkal, India 575025
{anumehaagrawal29, rosageorge97@gmail.com, sunitha98selvan}@gmail.com,
{sowmyakamath, m_anandkumar}@nitk.edu.in

Abstract

This paper includes approaches we have taken for Natural Language Inference, Question Entailment Recognition and Question-Answering tasks to improve domain-specific Information Retrieval. Natural Language Inference (NLI) is a task that aims to determine if a given *hypothesis* is an entailment, contradiction or is neutral to the given *premise*. Recognizing Question Entailment (RQE) focuses on identifying entailment between two questions while the objective of Question-Answering (QA) is to filter and improve the ranking of automatically retrieved answers. For addressing the NLI task, the UMLS Metathesaurus was used to find the synonyms of medical terms in given sentences, on which the InferSent model was trained to predict if the given sentence is an entailment, contradictory or neutral. We also introduce a new Extreme gradient boosting model built on PubMed embeddings to perform RQE. Further, a closed-domain Question Answering technique that uses Bi-directional LSTMs trained on the SquAD dataset to determine relevant ranks of answers for a given question is also discussed. Experimental validation showed that the proposed models achieved promising results.

1 Introduction

Recent studies have shown that patient-specific data can be utilized for the development of intelligent Healthcare Information Management Systems (HIMS), that support a wide range of supporting applications that enhance healthcare delivery platforms. The application of natural language processing, sophisticated data modeling, and predictive algorithms make it a highly interesting area of research. Patient data is continuously generated in large volume and variety, given the multiple modalities, it is available in (e.g., discharge summaries, physician's notes, clinical reports, lab reports etc). With an abundance of such diverse information sources available in the medical domain, sophisticated solutions that can adapt to the heterogeneity and specific manifold nature of health-related information are a critical requirement for HIMS development.

In clinical text, a commonly occurring problem would be to understand the correlation and association between various factors like disease, symptoms, diagnoses and treatment. Clinical text is inherently unstructured and written in natural language, and hence is prone to significant issues in effective interpretability and utilization. Challenges like paraphrase detection, anaphora resolution, natural language inference etc must be effectively dealt with in order to extract useful knowledge that can be used to build intelligent decision support applications. Such support systems require extensive evidence-based analysis, and context-sensitive processing, in order to enable higher-level functionalities like clinical question-answering. Thus, dealing with such issues is paramount importance.

Natural Language Inference is used to determine whether a given *hypothesis* can be inferred from a given *premise* (Ben Abacha et al. (2019)). The three inference relations to be identified between the statements are Entailment, Neutrality and Contradiction. If a statement is a true description of the other then it is labelled *Entailment*. If it is a false description then it is labelled *Contradiction*, otherwise, it is considered to be *Neutral*. The goal of Recognizing Question Entailment(RQE) is to retrieve answers to a premise question by retrieving inferred or entailed questions, called hypothesis questions that already have associated answers. Therefore, we define the entailment relation between two questions as: a question A en-

Proceedings of the BioNLP 2019 workshop, pages 533–540
Florence, Italy, August 1, 2019.

tails a question B if every answer to B also correctly answers A (Abacha and Demner-Fushman, 2016). RQE is particularly relevant due to the increasing numbers of similar questions posted online (Luo et al., 2015). For Question Answering, the input ranks are generated by the medical QA system CHiQA. Extracting certain elements of a question like the question type and focus is the main approach in question answering. If the question happens to contain multiple subquestions then an answer will be considered complete only if all sub-questions are answered. The rest of this paper is organized as follows: Section 2 presents a summarization on relevant existing research done in the area of interest. We discuss the Proposed Architecture for NLI, RQE and QA in Section 3. Section 4 presents the results and performance of the various models for each task, followed by error analysis, conclusion and references.

2 Related Work

There has been considerable research in the field of Medical Question Answering Systems. Incorporating QA systems with NLI and RQE give a machine the ability to better understand a query and fetch precise answers.

Modeling natural language inference is a complicated task but with the introduction of MedNLI (Romanov and Shivade, 2018; Goldberger et al., 2000), a new publicly available expert annotated dataset for NLI it has become possible to train models in order to achieve state-of-the-art performance. Chen et al. (2017) experimented with the SNLI corpus (Bowman et al., 2015) and MultiNLI corpus (Williams et al., 2017) to train complex models, to increase the performance of the neural network based NLI models with external knowledge. Most previous works on NLI worked on relatively small datasets, Chen et al. (2016) designed an approach to merge the modeling ability of neural networks with extra external inference knowledge. The advantage of using external knowledge is more significant when the training data is of limited size and is beneficial as more information is obtained. They obtained good results with this approach.

Romanov and Shivade (2018) presented a systematic comparison of various open domain models for NLI on MedNLI and studied the applicability of transfer learning techniques from the open domain to the clinical domain. They discussed their experimentation with a feature-based system in order to establish a baseline performance on MedNLI. Models that were explored include the Bag of Words model, InferSent (Conneau et al., 2017), ESIM (Enhanced Sequential Inference Model)(Chen et al., 2016). Other techniques included those that employed transfer learning, use of word embeddings and knowledge integration. In our work, we built on the work of these authors, by adapting their models and benchmarking them on different features, various word embeddings and clinical domain-oriented knowledge base to predict the relationship between the hypothesis and premise. Abacha and Demner-Fushman (2016) developed a method where RQE is applied to find a frequently asked question similar to consumer health questions, in order to answer consumer health questions with the answers given to similar FAQs. Groenendijk and Stokhof (1984) define an entailment relation between two questions Q_1, Q_2 if every proposition giving an answer to Q_1 is also giving an answer to Q_2. In our case, we used a supervised machine learning approach to determine whether or not a question Q_2 can be inferred from a question Q_1 by modeling the medical context's syntactic and semantic features, including complex relationships like negation, medical entities like disease, symptom, diagnoses and treatment etc. Abacha and Demner-Fushman (2016) used the NLM (National Library of Medicine) collection of 4,655 clinical questions asked by family doctors to construct the training corpus for RQE. For test pairs, two types of test data were collected - pairs of manually validated questions from the NLM collections and pairs of questions including FAQs retrieved online with a manual search of NIH websites. Four different statistical learning algorithms, SVM, Logistic Regression, Naive Bayes and J48, were used for RQE on the feature vector created. They reported the best results using the SVM classifier in the form of 75% F-measure values.

Abacha and Demner-Fushman (2019) studied question entailment in the medical domain and the effectiveness of the end-to-end RQE-based QA approach is calculated by evaluating the relevance of the retrieved answers. They benchmarked machine learning and deep learning approaches to RQE using different kinds of datasets, including textual inference, question similarity and entailment in both the open and clinical domains. The RQE methods (i.e. deep learning model and logistic regression classifier) are evaluated using two datasets of sentence pairs (SNLI and multiNLI), and three datasets of question pairs (Quora, Clinical-QE, and SemEval-cQA). They analyzed two methods for RQE: a deep learning model and Logistic Regression Classifier. Deep learning models achieve good results on open-domain and clinical datasets but delivered a lower performance on consumer health questions. When trained and tested on the same corpus, the Deep learning model with GloVe embeddings (Pennington et al. (2014)) gave the best results. Logistic Regression gave the best Accuracy on the Clinical-RQE dataset. When tested on our test set (850 medical CHQs-FAQs pairs), Logistic Regression trained on Clinical-QE gave the best performance.

Question answering (QA) is a crucial task that requires both natural language processing and domain related knowledge. Many Question Answering systems have been developed around the Question Answering dataset from Stanford (SQuAD) (Rajpurkar et al., 2016). The public leaderboard on the SQuAD website displays many deep learning models built for the task. Since the seminal work by Rajpurkar et al. (2016), many researchers have proposed different architectures for the task. The main feature of the dataset is that the answers are present as a span in the reverence document. The present state-of-the-art model is an AoA neural network by Cui et al. (2016), with an F1 score of 89.281, EM score of 82.482 and also outperforms the performance of humans.

3 Proposed Approaches

In this section, a detailed discussion on the various models designed for addressing the NLI, RQE and QA tasks, are presented.

3.1 Natural Language Inference

The first model proposed for NLI, a recurrent neural network (RNN) method is designed. We use the content of the two sentences to determine the two new rows $sentence_{id_1}$ and $sentence_{id_2}$ respectively which are formed using 300 dimensional glove embeddings. This feature vector created has been passed through a RNN with 300 nodes. Once this model was trained, we were able to get an accuracy of 67.1% with the test dataset.

The second model is the InferSent Model, which is a sentence embedding method that provides semantic representations for English sentences. As shown in Figure 1, the architecture centralizes on the idea that two sentences (premise input and hypothesis input) will be transformed by sentence encoder (same weights). After that, it leverages three matching methods to recognize relations between premise input and hypothesis input. The three matching methods are: concatenation of two vectors, product two vectors element-wise and absolute difference of two vectors. Conneau et al. (2017) proposed the model which is trained using GloVe word embeddings(Pennington et al., 2014). In our work, we used the MedNLI (Romanov and Shivade, 2018; Goldberger et al., 2000) along with different word embeddings such as 300D GloVe embeddings, MIMIC clinic data embeddings (Johnson et al., 2016), Wikipedia (english) embeddings, the combination of Wikipedia english and MIMIC clinical data embeddings and even with the combination of 300D GloVe with BioASQ (Tsatsaronis et al., 2015) and MIMIC embeddings. All the techniques were set with number of training epochs as 100 and were trained on GPUs.

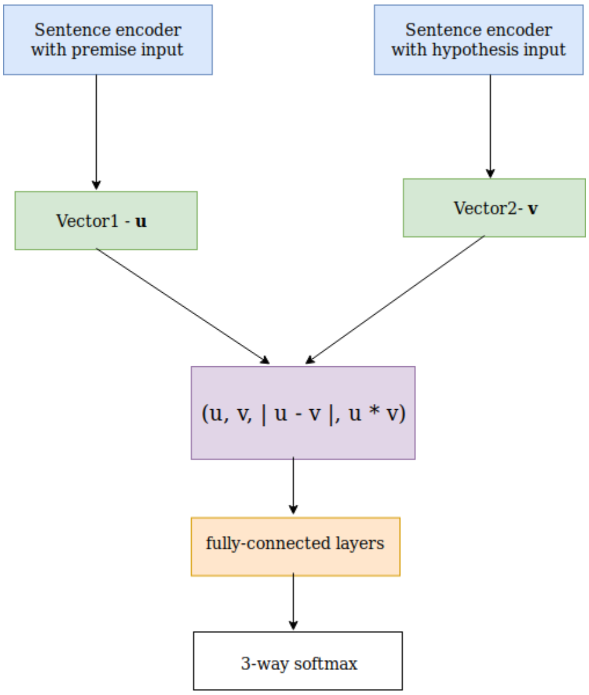

Figure 1: The Infersent Architecture adopted for NLI task

We also designed a novel technique to extract the semantic aspect of the clinical terms, for which we used the UMLS Metathesaurus (Aronson, 2001). The Metathesaurus is the largest component of UMLS that is organized by concept, or meaning, and links similar names for the term from over two hundred different vocabularies. The Metathesaurus is able to identify useful and relevant relationships between the various medical and non-medical concepts while preserving basic meaning and relationships from each vocabulary. We made use of the MetaMap tool for recognizing UMLS concepts in text. It can map medical texts to the UMLS Metathesaurus, using which we generated the synonyms for the terms that are not stop words and all synonyms have been generated with the use of UMLS Metathesaurus and NLTK corpus wordnet synsets. With this technique we were able to generate the highest accuracy yet of 87.7% on the test dataset given for the MediQA shared task.

3.2 Question Answering Task

The objective is to filter and improve the ranking of automatically retrieved answers, and the workflow employed is shown in Figure 2. Each question consists of several possible answers - relevant or irrelevant and are ranked based on the medical QA system CHiQA. We propose a system based on Question Answering model called Deeppavlov.ai (Burtsev et al., 2018). The context based question answering model uses SQuAD dataset to predict the answer. For every possible question, there are multiple answers and answer URLs associated with it. We scrape the content from the URL links and use that as context to the Deeppavlov model. The model takes in a question and a context to predict the answer. The answer provided in the AnswerText is a subset of the URL context. In case an answer does not have a URL associated with it, we use the AnswerText as the context.

Next, the model provides a score for every answer. This helps us determine how relevant the answer is for a question. We pass all the answers pertaining to a question to the model and obtain a score. We rank the answers based on the score obtained. We also need to determine if an answer is relevant or irrelevant. Based on the training dataset we set the threshold for relevance. If the answer has a score above the threshold then it is relevant, otherwise, it is irrelevant. This threshold is taken as the average of the scores of all answers belonging to a question. The threshold can be further improved based on the mode and median of the scores.

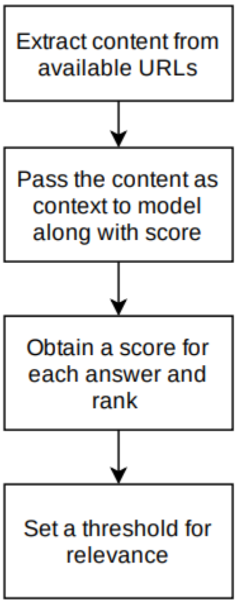
Extract content from available URLs
Pass the content as context to model along with score
Obtain a score for each answer and rank
Set a threshold for relevance

Figure 2: The proposed Question Answering workflow

3.3 Recognizing Question Entailment

For the next task, the stop words were removed and word stemming using the Porter algorithm (Porter, 1980) was performed for all (Q_1, Q_2) training pairs, to extract relevant features. We create the feature vector using lexical features and semantic features. The semantic features used are Negativity and Positivity, and Named entities count. The lexical features used are Jaccard similarity, Mover's distance and Bigram overlap (Abacha and Demner-Fushman, 2016). We used scikit-learn library (Pedregosa et al., 2011) for all the machine learning models. We also used NLTK (Natural Language Toolkit) (Loper and Bird, 2002) to find ngrams, Wordnet (pri, 2010) and StanfordNERTagger. Wordnet is a lexical database which can be used to find synonyms. StanfordNERTagger is used to find named entities in the text.

Four different models were benchmarked for the RQE task on the test dataset - Support Vector Machine (SVM), Logistic Regression Classifier (LRC), AdaBoost Classifier and XGBoost with PubMed Embeddings.

Support Vector Machine for RQE: We use word overlap, common bigrams, Jaccard similarity, cosine similarity and Levenshtein distance as the features(Abacha and Demner-Fushman

(2016)). We also calculate the Word Mover's distance and this is included in the feature vector. We pass this feature vector through a SVM model. In order to enhance the performance and alter it for the medical domain, we use PubMed (Pyysalo et al. (2013)) 200D embeddings to find the word vectors.

Logistic Regression Classifier for RQE: - The same feature vector that was used for the Support Vector Machine task is being used here. In addition to that, the feature list also includes the maximum and average values obtained with these measures and the question length ratio (length(P Q)/length(HQ)).The morphosyntactic feature indicating the number of common nouns and verbs between P Q and HQ is also used (Abacha and Demner-Fushman (2019)).

K-Nearest Neighbors Classifier for RQE: Using the same feature vectors that was used in the Logistic Regression Classifier, we have used the K-nearest neighbors classifier for RQE. In order to obtain the value of K resulting with the highest accuracy, we have ran the algorithm for K ranging from 5 to 70. With the varying values of K, the accuracy is measured and the highest accuracy measure was with K=47.

Ada Boost Classifier for RQE: We use a new approach which uses the Ada Boost Classifier. Adaptive Boosting uses results from weak learner algorithms and combines it into a weighted sum which represents the final output of the boosted classifier. Using the same feature set as above we pass the feature vector through the ensemble based model. AdaBoost produces better results as it is adaptive. This algorithm works better than the single classifiers as it pools the prediction of multiple classifiers and reduces model bias and variance.

XGBoost with PubMed embeddings for RQE: We present a new approach for RQE in the medical domain using XGBoost with PubMed embeddings. XGBoost (Chen and Guestrin (2016)) is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework and is an ensemble model. Ensemble learning offers a systematic solution to combine the predictive power of multiple learners. The result is a single model which gives the aggregated output from several models.

In XGboost, the ensemble trees are constructed much faster than any other ensemble classifier as it makes use of distributed computing. The feature vector created above is passed through the Extreme Gradient boosting algorithm and a new feature based on the PubMed 200D embedding is added to the feature vector for calculating the similarity between the two medical questions. This is one unique feature which can capture relations between medical terms and thus gives high accuracy.

4 Experimental Results and Discussion

We performed several experiments to benchmark the relative performance of the various proposed models for the three different tasks - NLI, RQE and QA. We analyzed the accuracy obtained using the standard metrics defined for the three tasks. The datasets provided from the ACL-BioNLP'19 Shared Task (Ben Abacha et al., 2019) were used for the experimental studies.

For the NLI task, we used the MedNLI dataset (Romanov and Shivade, 2018; Goldberger et al., 2000) built on different word embeddings such as 300D GloVe embeddings (Pennington et al. (2014)) , MIMIC clinic data embeddings (Johnson et al. (2016)) , Wikipedia English embeddings, combination of Wikipedia English and MIMIC clinical data embeddings and even with the combination of 300D GloVe with BioASQ (Tsatsaronis et al. (2015)) and MIMIC embeddings. All the techniques were set with number of training epochs as 100 and were trained on GPUs. The observed performance for the MediQA released test dataset is tabulated in Table 1.

Table 1: Performance of the InferSent model for NLI task when different embeddings are used

Embeddings used	Accuracy
Wiki English	71.4%
MIMIC	71.7%
GloVe with BioASQ and MIMIC	72.4%
Wiki English with MIMIC	74.4%

The accuracy obtained with different methods is listed in Table 2. The RNN based model for NLI was trained for 30 epochs and a validation accuracy of 67.1% was achieved. Further accuracy can be improved by using a Bidirectional LSTM or Bidirectional GRU. The InferSent Model for NLI with MIMIC (Johnson et al., 2016) embeddings that were trained for 100 epochs gave an accu-

Table 2: Comparative performance of the proposed approaches for the NLI, RQE and QA tasks

Methology Used	Task	Accuracy
RNN	NLI	67.1%
Infersent+MIMIC	NLI	71.7%
Infersent+MIMIC+Wiki	NLI	74.4%
UMLS Metathesaurus	NLI	87.7%
SVM	RQE	62%
Logistic Regression	RQE	64.5%
KNN	RQE	62.4%
Naive Bayes	RQE	65%
Ada Boosting	RQE	66%
XgBoost	RQE	66.7%
Closed domain QA	QA	53.6%

racy of 71.7% This is because the medical context of the sentences was taken care of by the MIMIC word embeddings. The InferSent Model for NLI with MIMIC and Wikipedia english words embeddings gave an accuracy of 74.4% when trained for 100 epochs. This is because the medical and grammatical concepts were given special emphasis during the modeling phase. The model built on UMLS Metathesaurus and NLTK wordnet synsets model achieved an accuracy of 87.7% on the test data and 93.2% on validation data.

In the case of the RQE task, the SVM model was trained using a few features like semantic features, bigram overlap, word movers distance and cosine similarity. An accuracy of 62% achieved. The Logistic Regression model was trained using several handcrafted features and an accuracy of 64.5% was achieved. The KNN algorithm was also used for this classification task and an accuracy of 62.4% was obtained with K=47. The Naive Bayes model was fine-tuned and trained using the constructed feature vector. This gave an accuracy of 64%. The Naive Bayes model feature vector was modified again to include a feature which will consider the content of both the questions, which improved the accuracy by 1%. The AdaBoost classifier was used and this ensemble based method performed better than the naive methods and gave an accuracy of 66%. The XGBoost method performed the best and gave an accuracy of 66.7% on the test set.

As can be seen from Table 2, the closed domain question answering model gave an accuracy of 53.6% which is much above the baseline fixed at 51%. This accuracy was achieved because this method focuses on finding the specific answer in the given context which is more relevant to a given question. Based on the scores obtained from the closed domain model, the answers have been ranked accordingly. The model achieved an accuracy of 53.6%, precision of 55.9% and Mean Reciprocal Rank of 62.93%.

4.1 Discussion

Based on the experimental results, we hereby present several observations and insights into the proposed models. In the case of the NLI task, we performed error analysis for the InferSent model by varying the embeddings and incorporating UMLS Metathesaurus and found that the error ratio also varies. The ratio of the error rate between neutral, entailment and contradiction was observed to be $5 : 4 : 3$, when the MIMIC word embeddings (Johnson et al. (2016)) were used. However, it changed to $2 : 1 : 1$ when the UMLS Metathesaurus with WordNet (pri, 2010) synsets are used. Thus, it can be concluded that the neutral label was the hardest to predict and differentiating between entailment and neutrality is also challenging. In our current implementation, if similar terms are present in the hypothesis and premise, then the label of entailment is still predicted, whereas the statements could actually be neutral. We also noticed that, by using clinical domain-specific embeddings, the predictions become more accurate.

Table 4 shows the Precision and Recall values for RQE using XgBoost. XgBoost provides a parallel tree boosting which improves the accuracy. Also, it uses continued training so it can further boost an already fitted model on new data, thus a significant improvement in accuracy is observed.

Table 3: Confusion Matrix for NLI

Label	True	False
Entailment	121	14
Neutral	114	21
Contradiction	128	15

Table 4: Confusion Matrix for RQE

Parameters	True	False
Precision	0.68	0.66
Recall	0.65	0.69

5 Concluding Remarks

In this paper, several techniques for the NLI, RQE and QA tasks were discussed. For addressing the NLI task, the UMLS Metathesaurus was used to find the synonyms of medical terms in given sentences, on which the InferSent model was trained to predict if the given sentence is an entailment, contradictory and neutral. We also designed a new Extreme gradient boosting model built on PubMed embeddings to perform RQE. Further, a closed-domain Question Answering technique that uses Bi-directional LSTMs trained on the SquAD dataset to determine relevant ranks of answers for a given question was also presented. Among the proposed models, the UMLS Metathesaurus and NLTK wordnet synsets model achieved the highest accuracy of 87.7% on the test dataset provided by the MediQA Challenge (Ben Abacha et al., 2019). For RQE, the highest accuracy of 66.7% was achieved using the XGBoost method. For the QA task, we achieved an accuracy of 53.6%, precision of 55.9% and Mean Reciprocal Rank of 62.93%.

As future work, we intend to extend the textual inference model for the clinical domain to develop decision support applications so that treatment methods can be simplified by grouping similar diseases and problems together. This can be achieved by using RQE which can aid in analyzing if two different health conditions are similar enough to have the same treatment. The model can also be trained on MedQuAD dataset (Abacha and Demner-Fushman (2019)) to improve the accuracy so that the model can perform more accurately in real-world hospital scenarios.

References

Princeton university ”about wordnet.” wordnet. princeton university. [online]. 2010.

Asma Ben Abacha and Dina Demner-Fushman. 2016. Recognizing question entailment for medical question answering. In *AMIA Annual Symposium Proceedings*, volume 2016, page 310. American Medical Informatics Association.

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *arXiv preprint arXiv:1901.08079*.

Alan R Aronson. 2001. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association.

Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. 2019. Overview of the mediqa 2019 shared task on textual inference, question entailment and question answering. In *Proceedings of the BioNLP 2019 workshop, Florence, Italy, August 1, 2019*. 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. *arXiv preprint arXiv:1508.05326*.

Mikhail Burtsev, Alexander Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nickolay Bushkov, Olga Gureenkova, Taras Khakhulin, Yuri Kuratov, Denis Kuznetsov, et al. 2018. Deeppavlov: Open-source library for dialogue systems. In *Proceedings of ACL 2018, System Demonstrations*, pages 122–127.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Diana Inkpen, and Si Wei. 2017. Neural natural language inference models enhanced with external knowledge. *arXiv preprint arXiv:1711.04289*.

Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2016. Enhanced lstm for natural language inference. *arXiv preprint arXiv:1609.06038*.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. *CoRR*, abs/1603.02754.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

Yiming Cui, Zhipeng Chen, Si Wei, Shijin Wang, Ting Liu, and Guoping Hu. 2016. Attention-over-attention neural networks for reading comprehension. *arXiv preprint arXiv:1607.04423*.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. 2000. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220.

Jeroen Antonius Gerardus Groenendijk and Martin Johan Bastiaan Stokhof. 1984. *Studies on the Semantics of Questions and the Pragmatics of Answers*. Ph.D. thesis, Univ. Amsterdam.

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

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics*.

Jake Luo, Guo-Qiang Zhang, Susan Wentz, Licong Cui, and Rong Xu. 2015. Simq: Real-time retrieval of similar consumer health questions. *J Med Internet Res*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

M.F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

Sampo Pyysalo, Filip Ginter, Hans Moen, Tapio Salakoski, and Sophia Ananiadou. 2013. Distributional semantics resources for biomedical text processing.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. *arXiv preprint arXiv:1808.06752*.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artiéres, Axel-Cyrille Ngonga Ngomo, Norman Heino, Eric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*.

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