NLP Term Project

Task : Retrieval of text spans from relevant articles for biomedical question answering

Team : SUV

Team Members : Suyash Namdeo (17CE10059) | Vivek Pal (17MA20048) | Vinay Agarwal (17MA20051)

Problem Statement / Task Overview

Given a question and a set of relevant articles independent of the question, we are supposed to extract the relevant article along with the word span of the text relevant for answering that particular question. In this task, we have to explore answering the factoid /yes-no/ list questions. The corpus used is the BioASQ shared task-7b phase b corpus. We were supposed to explore various state-of-the-art deep learning QA techniques to solve the problem.

Introduction

In this report, we investigate how the recently introduced pre-trained language model can be adapted for biomedical corpora. This report provides an application of NLP in the constantly expanding field of automated answering of questions. For this, we took BioASQ 7b corpus that has training and testing corpus of biomedical questions and answering dataset and trained it in our model and analysed the important statistical parameters to get the optimum performance of our model. BIOASQ assesses the ability of systems to semantically index very large numbers of biomedical scientific articles, and to return concise and user-understandable answers to given natural language questions by combining information from biomedical articles and ontologies. In this task we explored the task of answering the factoid/Yes or No/list questions. In this report we'd give the details of the models and their performances that we explored in the state-of-the-art deep learning QA techniques. Our analysis results show that pre-training on biomedical corpora helps it to understand complex biomedical texts.

Motivation

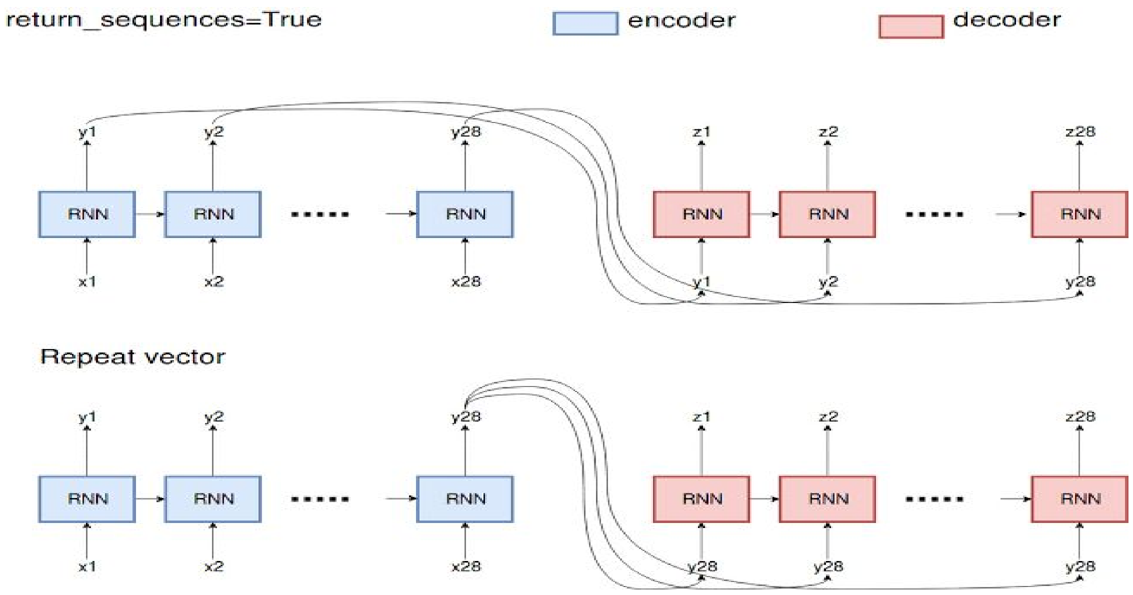
In the modern era of constantly evolving medical sciences and the age of the internet, a remarkable amount of medical literature is continuously being posted and is available online. This has led to a need for an effective retrieval and indexing system which can allow us to extract sensible meaningful insights from those available information resources. Biomedical text mining is becoming increasingly important as the number of biomedical documents rapidly grows. With the progress in natural language processing (NLP), extracting valuable information from biomedical literature has gained popularity among researchers, and deep learning has boosted the development of effective biomedical text mining models. However, directly applying the advancements in NLP to biomedical text mining often yields unsatisfactory results due to a word distribution shift from general domain corpora to biomedical corpora. The most effective way is to take the amount of this huge dataset that is available globally and build a Question-Answering (QA) system that allows us to directly query this data in the corpus and extract sensible, meaningful and structured or semi structured information in a human readable format.

Model Architecture

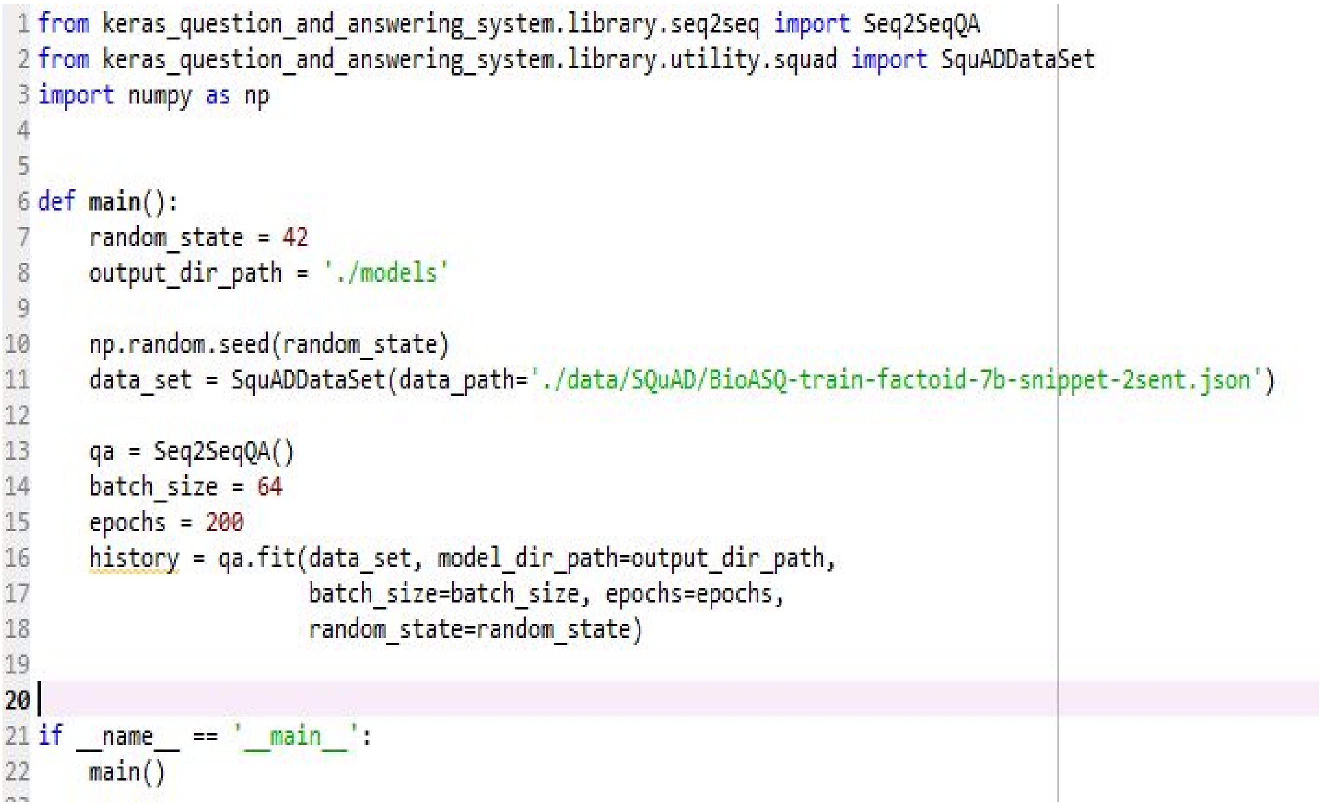
Question answering system developed using seq2seq and memory network model in Keras. In this project we worked with two different encoding for the input of the encoder (one hot encoding and GloVe word2vec encoding) while merging the paragraph context and the question as the input of the encoder.

The implemented models include:

- **Seq2seq.py** : One-hot encoding input that is paragraph_context + ' Q ' + question
- **Seq2seq_v2.py** : One-hot encoding input that is add(paragraph_context, RepeatVector(LSTM(question)))
- **Seq2seq_glove.py** : GloVe encoding input that is paragraph_context + ' Q ' + question
- **Seq2seq_v2_glove.py** : GloVe encoding input that is add(paragraph_context, RepeatVector(LSTM(question)))



The trained models are included in the demo folder in the project . The training was done using the Full-Abstract/ BioASQ-train-factoid-7b-snippet data set on with 200 epochs and batch size of 64.



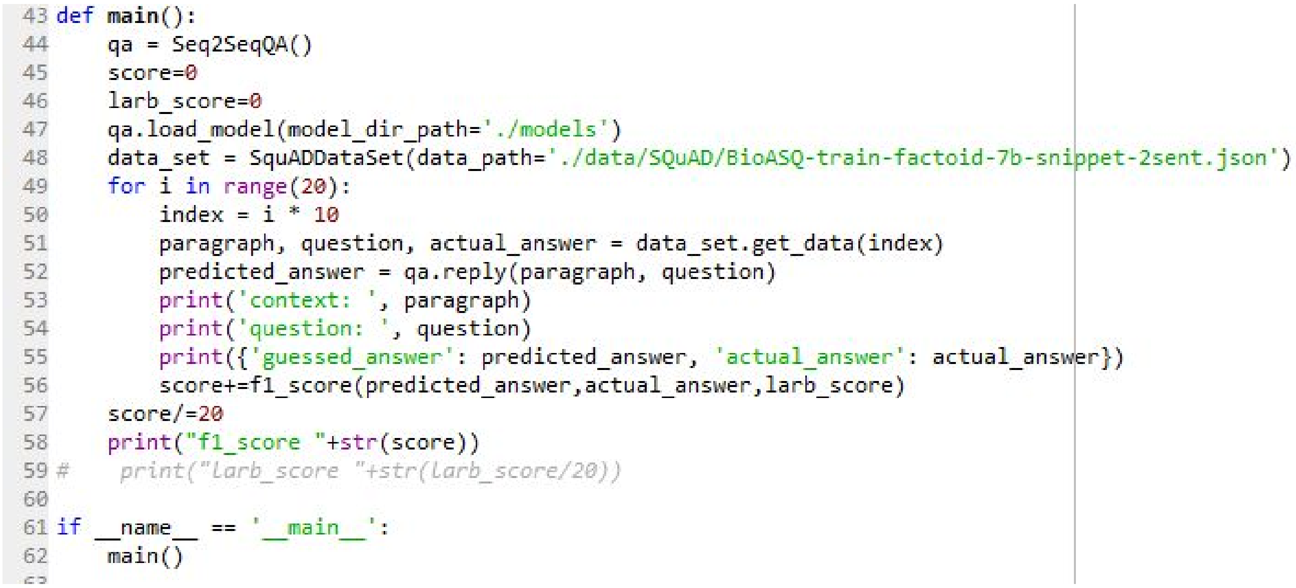
```python
from keras_question_and_answering_system.library.seq2seq import Seq2SeqQA
from keras_question_and_answering_system.library.utility.squad import SquADDataSet
import numpy as np


def main():
    random_state = 42
    output_dir_path = './models'

    np.random.seed(random_state)
    data_set = SquADDataSet(data_path='./data/SQuAD/BioASQ-train-factoid-7b-snippet-2sent.json')

    qa = Seq2SeqQA()
    batch_size = 64
    epochs = 200
    history = qa.fit(data_set, model_dir_path=output_dir_path,
                     batch_size=batch_size, epochs=epochs,
                     random_state=random_state)


if __name__ == '__main__':
    main()
```

Results and Discussion

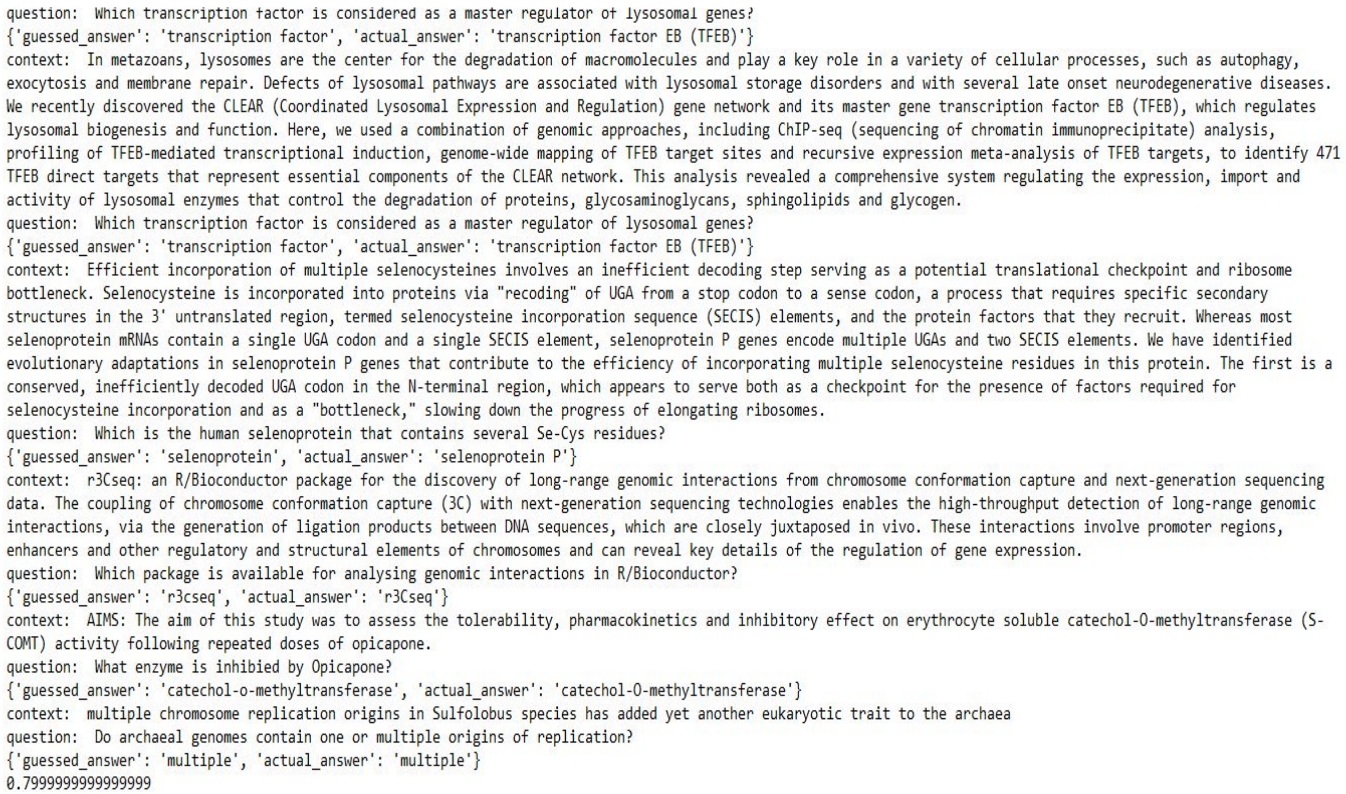
Precision is a measure of exactness or quality, whereas recall is a measure of completeness or quantity. In simple terms, high precision means that an algorithm returned substantially more relevant results than irrelevant ones, while high recall means that an algorithm returned most of the relevant results.

F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it’s better to look at both Precision and Recall. In our case, F1 score is **0.81666667** which outperforms the widely used baseline obtained with Infersent embeddings.

Precision = TP/TP+FP
Recall = TP/TP+FN
F1 Score = 2*(Recall * Precision) / (Recall + Precision)

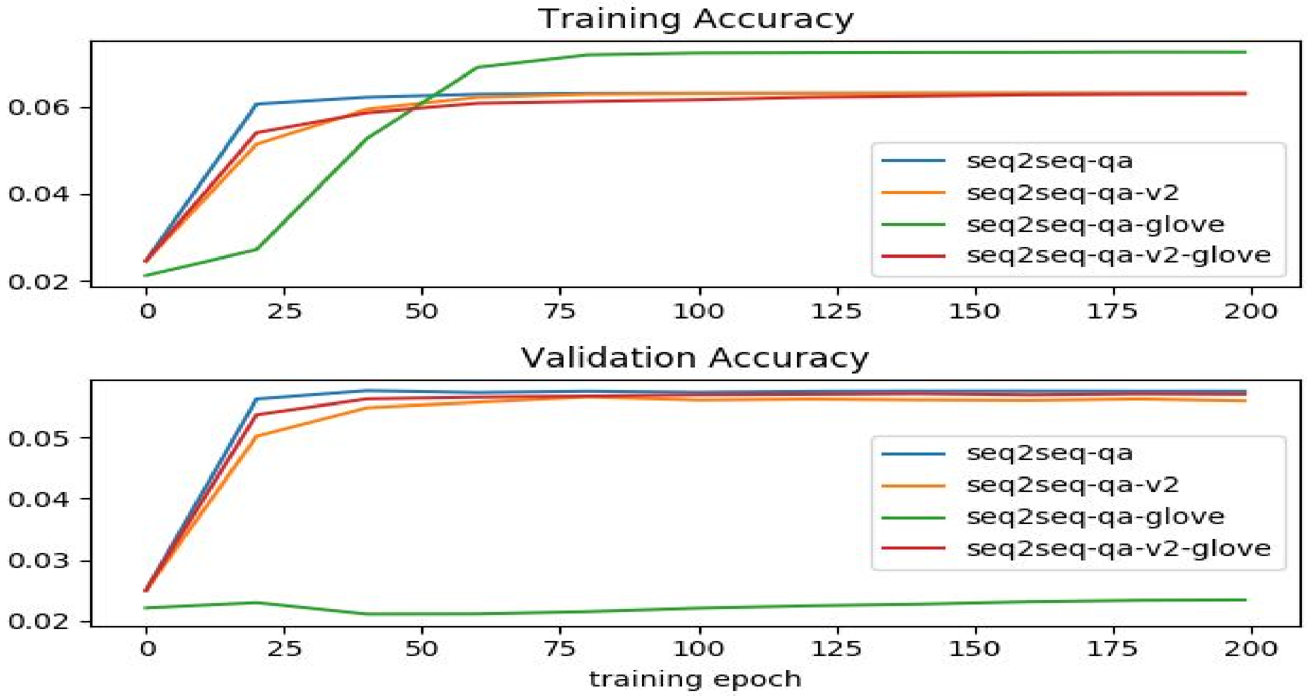


```python
def main():
    qa = Seq2SeqQA()
    score=0
    larb_score=0
    qa.load_model(model_dir_path='./models')
    data_set = SquADDataSet(data_path='./data/SQuAD/BioASQ-train-factoid-7b-snippet-2sent.json')
    for i in range(20):
        index = i * 10
        paragraph, question, actual_answer = data_set.get_data(index)
        predicted_answer = qa.reply(paragraph, question)
        print('context: ', paragraph)
        print('question: ', question)
        print({'guessed_answer': predicted_answer, 'actual_answer': actual_answer})
        score+=f1_score(predicted_answer,actual_answer,larb_score)
    score/=20
    print("f1_score "+str(score))
#    print("larb_score "+str(larb_score/20))

if __name__ == '__main__':
    main()
```



```
question:  Which transcription factor is considered as a master regulator of lysosomal genes?
{'guessed_answer': 'transcription factor', 'actual_answer': 'transcription factor EB (TFEB)'}
context:  In metazoans, lysosomes are the center for the degradation of macromolecules and play a key role in a variety of cellular processes, such as autophagy,
exocytosis and membrane repair. Defects of lysosomal pathways are associated with lysosomal storage disorders and with several late onset neurodegenerative diseases.
We recently discovered the CLEAR (Coordinated Lysosomal Expression and Regulation) gene network and its master gene transcription factor EB (TFEB), which regulates
lysosomal biogenesis and function. Here, we used a combination of genomic approaches, including ChIP-seq (sequencing of chromatin immunoprecipitate) analysis,
profiling of TFEB-mediated transcriptional induction, genome-wide mapping of TFEB target sites and recursive expression meta-analysis of TFEB targets, to identify 471
TFEB direct targets that represent essential components of the CLEAR network. This analysis revealed a comprehensive system regulating the expression, import and
activity of lysosomal enzymes that control the degradation of proteins, glycosaminoglycans, sphingolipids and glycogen.
question:  Which transcription factor is considered as a master regulator of lysosomal genes?
{'guessed_answer': 'transcription factor', 'actual_answer': 'transcription factor EB (TFEB)'}
context:  Efficient incorporation of multiple selenocysteines involves an inefficient decoding step serving as a potential translational checkpoint and ribosome
bottleneck. Selenocysteine is incorporated into proteins via "recoding" of UGA from a stop codon to a sense codon, a process that requires specific secondary
structures in the 3' untranslated region, termed selenocysteine incorporation sequence (SECIS) elements, and the protein factors that they recruit. Whereas most
selenoprotein mRNAs contain a single UGA codon and a single SECIS element, selenoprotein P genes encode multiple UGAs and two SECIS elements. We have identified
evolutionary adaptations in selenoprotein P genes that contribute to the efficiency of incorporating multiple selenocysteine residues in this protein. The first is a
conserved, inefficiently decoded UGA codon in the N-terminal region, which appears to serve both as a checkpoint for the presence of factors required for
selenocysteine incorporation and as a "bottleneck," slowing down the progress of elongating ribosomes.
question:  Which is the human selenoprotein that contains several Se-Cys residues?
{'guessed_answer': 'selenoprotein', 'actual_answer': 'selenoprotein P'}
context:  r3Cseq: an R/Bioconductor package for the discovery of long-range genomic interactions from chromosome conformation capture and next-generation sequencing
data. The coupling of chromosome conformation capture (3C) with next-generation sequencing technologies enables the high-throughput detection of long-range genomic
interactions, via the generation of ligation products between DNA sequences, which are closely juxtaposed in vivo. These interactions involve promoter regions,
enhancers and other regulatory and structural elements of chromosomes and can reveal key details of the regulation of gene expression.
question:  Which package is available for analysing genomic interactions in R/Bioconductor?
{'guessed_answer': 'r3cseq', 'actual_answer': 'r3Cseq'}
context:  AIMS: The aim of this study was to assess the tolerability, pharmacokinetics and inhibitory effect on erythrocyte soluble catechol-O-methyltransferase (S-
COMT) activity following repeated doses of opicapone.
question:  What enzyme is inhibied by Opicapone?
{'guessed_answer': 'catechol-o-methyltransferase', 'actual_answer': 'catechol-O-methyltransferase'}
context:  multiple chromosome replication origins in Sulfolobus species has added yet another eukaryotic trait to the archaea
question:  Do archaeal genomes contain one or multiple origins of replication?
{'guessed_answer': 'multiple', 'actual_answer': 'multiple'}
0.7999999999999999
```

The figure below compare the training accuracy and validation accuracy of various models using the script **squad_compare_models**:



In seq2seq-qa-v2 one hot encoding seq2seq used but it is different from seq2seq-qa as in seq2seq-qa-v2 the paragraph context and the question are added after the LSTM + RepeatVector layer. While in seq2seq-qa-glove data was trained on word-level (GloVe word2vec encoding) with input = paragraph_context + ' Q ' + question) and in Seq2seq-qa-v2-glove the paragraph context and the question are added after the LSTM + RepeatVector layer. To summarize, the **RepeatVector** is used as an adapter to fit the fixed-sized output of the encoder to the differing length and input expected by the decoder.

Model	**F1-Score**
seq2seq-qa	**0.7999999**
seq2seq-v2-qa	**0.8166667**
seq2seq-glove-qa	**0.7499999**
seq2seq-glove-v2-qa	**0.7999999**

Ablation Analysis (merits and demerits)

Base-Line

We have used Infersent, which is a *sentence embeddings* method that provides semantic sentence representations. These embeddings can be used for various downstream tasks like finding similarity between two sentences. We have tried to using sentence embedding for finding the sentence which is most likely having the required answer. Getting the correct answer from answer, we have used these models

1). Unsupervised Learning Model : Firstly tried using euclidean distance to detect the sentence having minimum distance from the question. The accuracy of this model came around 45%. Then, I switched to cosine similarity and the accuracy improved from 45% to 63%. This makes sense because euclidean distance does not care for alignment or angle between the vectors whereas cosine takes care of that. Direction is important in case of vectorial representations.

1). Supervised Learning Model : Here, We have restricted my paragraph length to 10 sentences for simplicity. The idea is to match the root of the question which is which may be same for a sentence in any case to all the roots/sub-roots of the sentence. It is important to do stemming before comparing the roots of sentences with the question root. For example root word for question is appear while the root word in the sentence is appeared. If you don't stem appear & appeared to a common term, them matching won't be possible. The accuracy of multinomial logistic regression is 65% for the validation set. Considering the original model which had a lot of features with an accuracy of 79%, this one is quite simpler. The random forest gave an accuracy of 67% and finally, XGBoost worked best with an accuracy of 69% on the validation set.

Conclusion

In this article, we introduced seq2seq model with repeatvector, which is a Encoder-Decoder based model for biomedical text mining. Requiring minimal task-specific architectural modification, our model outperforms previous models on biomedical QA task. We observed that our seq2seq performs better than baseline model. We have trained the model for BioASq 7b/train/Full_Abstract/BioASq-Factoid dataset as well as for list-7b dataset with 200 epochs and batch size of 64. The predicted answers for the questions are for the same dataset. From error analysis it is evident that among different variants of seq2seq model, the one which is trained with Repeatvector provided the most promising results as it is used as an adapter to fit the fixed-sized output of the encoder to the differing length and input expected by the decoder.