Retrieval of Similar Questions and Answer Recommendation

Manpreet Nanreh*
mnanreh@ryerson.ca
Ryerson University
Toronto, Ontario

ABSTRACT

1 INTRODUCTION

Online question answering (Q&A) communities contain a rich source of knowledge expanding many different topics. These communities such as Quora allow users to retrieve such information. Users have the ability to ask questions or provide answers to existing questions on these websites. However, the main issue these services often face is the constant duplication of already asked questions and difficulty of providing suggestive answer. A user might try to explore the existing collection of questions manually but upon failure they will result to asking a new question. Therefore, this paper focuses on addressing these problems and proposes a model in order to reduce the Q&A clutter.

2 RELATED WORK

There has been a lot work previously done in order to detect text similarity. Especially the work done by Mueller and Thyagarajan (2016) [4] offers reliable Manhattan LSTM (MaLSTM) model as its architecture can be seen in **Figure 1**. Their model consists of two network of LSTM_a and LSTM_b which work in a siamese connection and share the weights among themself such that $\text{LSTM}_a = \text{LSTM}_b$. Each one of the LSTM is fed one of the two pieces of text being compared and the output is the exponent raised negative manhattan distance. This model will be used in this study in order to determine question similarity.

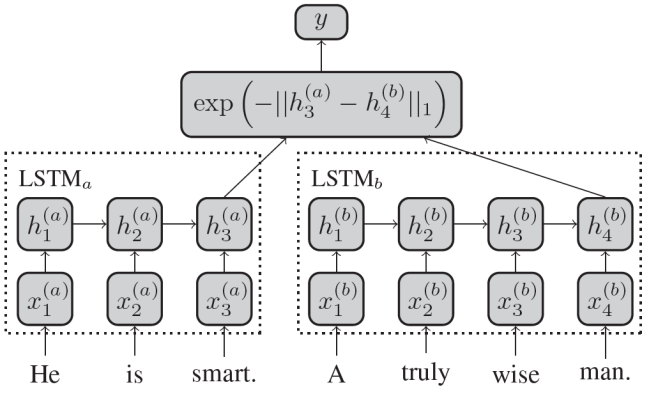

Figure 1: MaLSTM model architecture as defined in the paper by Mueller and Thyagarajan (2016) [4].

*All code can be found at Github.

3 METHOD

3.1 Dataset

For this study, the Quora question pairs dataset was obtained from Kaggle [5]. The dataset consists of train and test sets from which only the train set was used for this study because it provided the ability to evaluate the model's performance. The test set consists of randomly generated question pairs by Quora which did not provide any usability for this study. The train set contains two questions and the output indicating whether the two question are similar for each entry.

The Quora question pair dataset was prepared by applying the following steps for each question:

- tokenized the questions in order to get its token representation
- stripped any line break or space around the token and lower case them
- removed stop words from the tokens
- removed punctuation
- applied Porter Stemming to each token

In order to provide users with answer suggestions, a different dataset was used because a single source of data representing all the features for this study could not be found. Since the answer dataset is from a different source, this model can be used for multiple purposes. The answer dataset was retrieved from MS MARCO [3] and it was prepared using the following steps for each answer:

- tokenized the questions in order to get its token representation
- stripped any line break or space around the token and lower case them
- removed stop words from the tokens
- removed punctuation
- applied Porter Stemming to each token

3.2 Sentence Similarity Model

Whenever a user attempts to ask a new question, it gets compared with all the questions in the database and determined whether a duplicate exists. In the case a duplicate question is found, then the user gets presented with the found question and its respective answer. In order to perform question comparison, the MaLSTM model [4] was used which required the Quora question pair dataset to be embedded. To prepare the dataset, a vocabulary representation was created from the tokens of all the questions which was used to number encode the tokens of each question. The 300-dimensional word2vec embedding [1] alongside the vocabulary was used to create the embedding matrix for the LSTM layers in MaLSTM model. After zero padding all the embedded questions to a uniform length, the dataset was split into training and testing set with the following split ratio: 80% and 20%.

The architecture layout of neural network can see in **Figure 2** including the MaLSTM scheme.

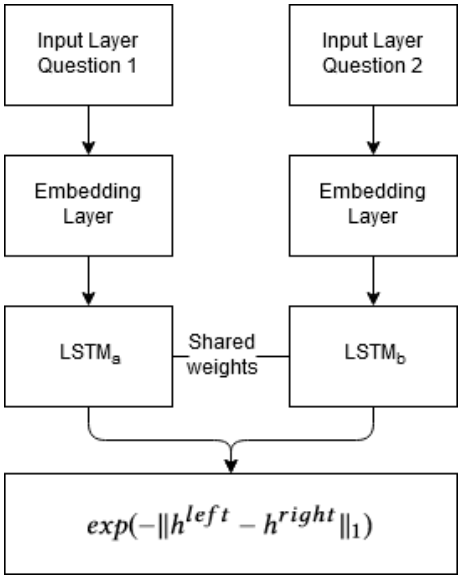

Figure 2: Brief layout of the neural network model for sentence similarity.

During training, the training dataset was split into 80% training and 20% validation. The model was trained for 20 epochs with 50 hidden layers for each LSTM layer and batch size of 1024. The loss function used was mean squared error alongside with the Adam optimizer [2]. A brief summary of the model can be seen in **Figure 3**.

3.3 Answer Recommendation Model

Whenever a duplicate question is not found, the model will suggest answers relating to the asked question. This answer recommendation model attempts to determine the top n keywords for each answer and compare them with the keywords in the asked question. The answers with the most matching keywords were determined to be related to the question and therefore recommended to the user.

In order to provide answer suggestions, the answer dataset had to be further modified. For each token in every answer, its Term Frequency-Inverse Document Frequency (TFIDF) was calculated. The TFIDF in answer recommendation was used to determine top n keywords. The similar steps were taken to find the keywords for the user's query question. Whenever the user had asked a question, the model will compare the keywords from query question and compare them with the keywords from all the answers. The most matching answers were returned. A brief summary of this model can be seen in **Figure 4**.

3.4 Overall Model

The full model first used the Sentence Similarity Model to determine if a duplicate question exists in relation to the query question. Upon finding a duplicate question, the user was recommended the duplicate question along with its answers. If a duplicate was not found, then the Answer Recommendation Model was used to determine answer suggestions. When no suggestions were found, the user had the option to ask the question. In any of these cases,

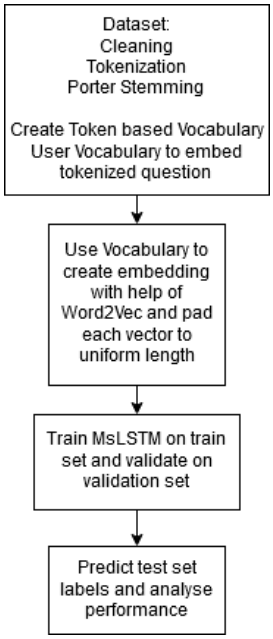

Figure 3: Brief summary of sentence similarity model.

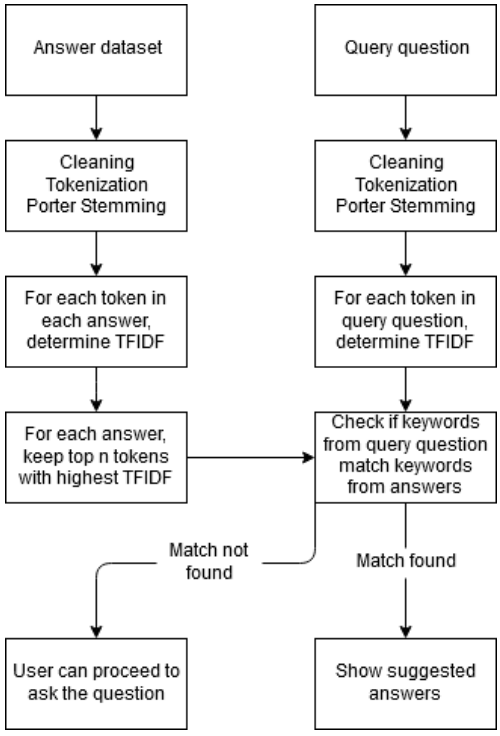

Figure 4: Brief summary of answer recommendation model.

the user always had the option to ask the question regardless of the over model. The layout of overall model can be seen in **Figure 5**.

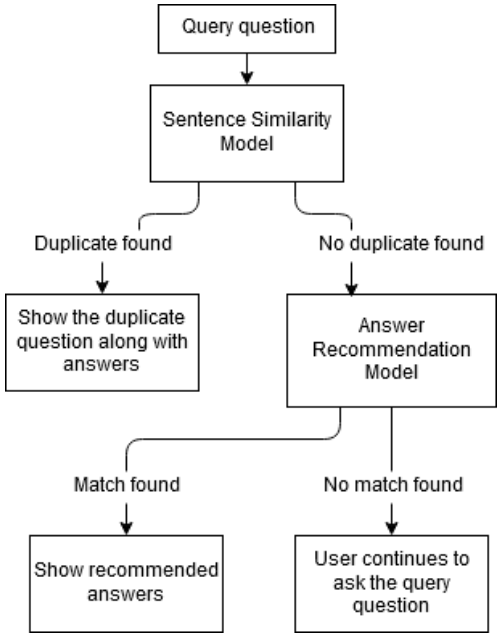

Figure 5: Brief summary of overall model.

4 RESULTS & DISCUSSION

The neural network consisting of MsLSTM was tested on the testing set. It was able to provide 79% accuracy on testing set. During training, the model was improving with the validation loss decreasing with each iteration as seen **Figure 6** and the accuracy increasing as seen in **Figure 7**. The MsLSTM model was able to provide good

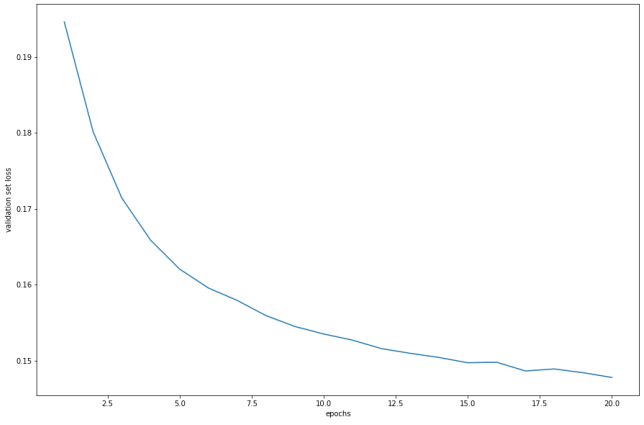

Figure 6: Epoch vs validation set loss.

results and is a good representation of finding duplicate questions.

The Answer Recommendation Model was tested on questions not existing in the dataset. In order for quick testing, the model was tested on a small subset of the full dataset.

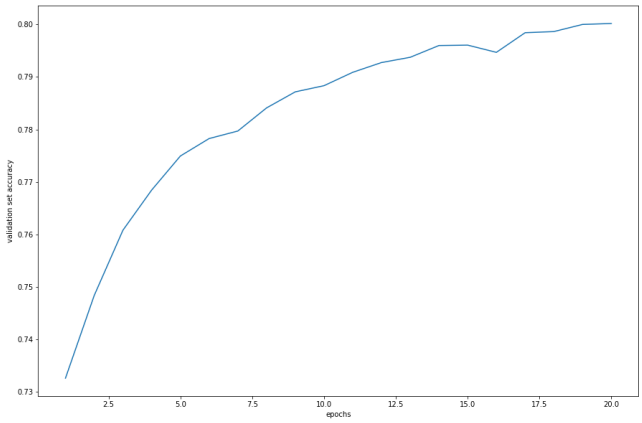

Figure 7: Epoch vs validation set accuracy.

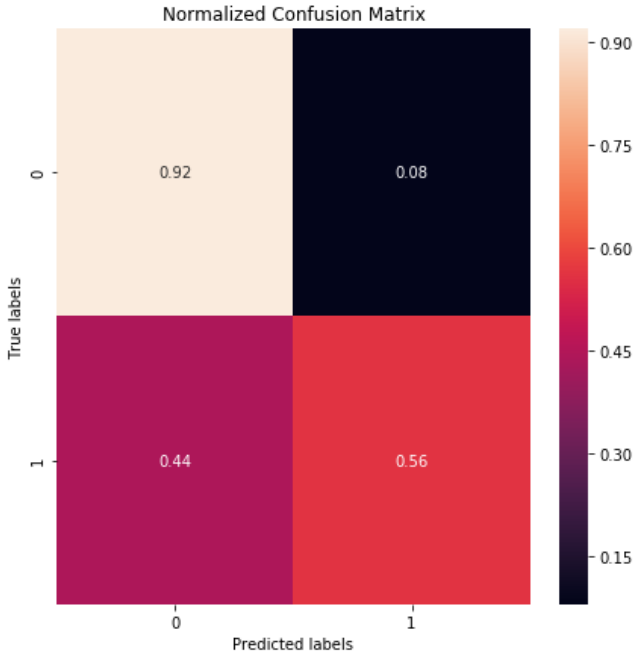

Figure 8: Normalized confusion matrix from test set results where 0 represents not similar and 1 represents similar.

When the overall model was tested on the question "How can I overcome tension?", the Sentence Similarity Model found that a duplicate questions already existed and presented the found question. The duplicated found was: "How do I overcome overcome tension?".

When asked the question: "What is the speed of light?", the Sentence Similarity Model determined that no duplicate existed and Answer Recommendation Model found the following top 3 suggestive answers:

- The speed of light in MPH is 670,616,629 mph.

	precision	recall	f1-score
0	0.78	0.92	0.84
1	0.80	0.56	0.66
accuracy			0.79

Figure 9: Classification report from test set results where 0 represents not similar and 1 represents similar.

- Three factors can limit the speed of photosynthesis: light intensity, carbon dioxide concentration and temperature. Without enough light, a plant cannot photosynthesise very quickly, even if there is plenty of water and carbon dioxide. Increasing the light intensity will boost the speed of photosynthesis.
- Types of rosin for violins are light rosin and dark rosin.

5 CONCLUSION

Both of the models in combination allow to create a system that could help Q&A online communities to reduce question answer clutter. The models in action work very well and are able help amplify the Human Computer interaction and better assist the users into finding answers to their question. In the case a user is not able to find their question manually, the overall model presented in this study offers a way to assist them in finding their answer when the user goes to ask a new question. For future work, Named-entity recognition (NER) can be used in order to extract better keyword from answers and query question.

REFERENCES

[1] Google 2013. *word2vec*. Google. https://code.google.com/archive/p/word2vec/.

[2] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. arXiv:cs.LG/1412.6980

[3] MS MARCO. 2019. *Question Answering V2.1*. Microsoft. http://www.msmarco.org/dataset.aspx.

[4] Jonas Mueller and Aditya Thyagarajan. 2016. Siamese Recurrent Architectures for Learning Sentence Similarity. https://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12195

[5] Quora 2017. *Quora Question Pairs*. Quora. https://www.kaggle.com/c/quora-question-pairs/data.