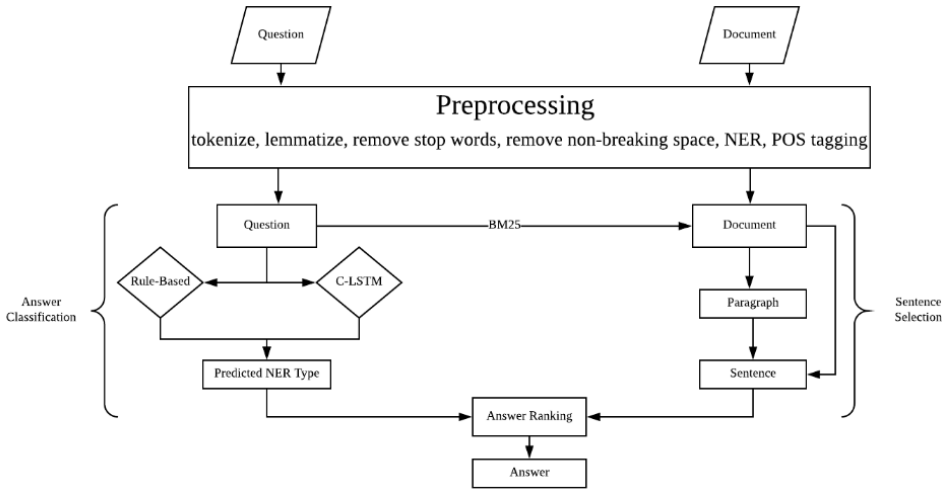
COMP90042 Project 2018 Report
Team Name: LMM Usernames: Fan Li, Jianxing Ma, Shukai Ma.

1. Introduction:

Question Answering ("QA") is the task of automatically determining the answer for a natural language question. In this project, the aim is to develop a QA system based on the given datasets, which includes training set for building model, development set for implementation decision and detailed analysis, and testing set for testing the model. More specifically, given a question and a document, the goal of the QA system is to find the answer to that question within the corresponding document. This report will introduce the developed QA system for this project with reference to the techniques used in implementing the system as well as the rationale behind them, the results of the QA system including results exposition and error analysis, and what future works will implemented in enhancing the performance of the system.

2. Method Description:

The pipeline architecture of the QA system is presented as the graph 2.1 shown below. Generally, the QA system is implemented in three major steps, which are preprocessing, sentence selection and answer classification.

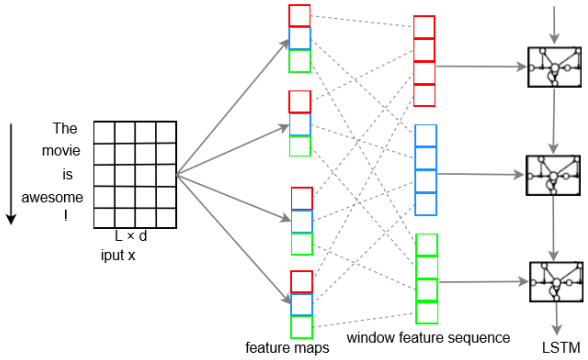


Graph 2-1

In preprocessing, serval preprocessing techniques are used including tokenization, lemmatization, stop words removing, NER, etc. It is worth mentioning that, NER is a vital process in preprocessing because refer to the pipeline architecture above, the rationale of the QA system can be summarized as classify the NER type of a given question and find the corresponding answer in the besting matching sentence, thus, an ad-hoc NER package named Stanford CoreNLP NER is chosen in implementing the NER process for each document. Based on this NER package, various types of tokens are recognized; for the unrecognized token, they all be classified as type "O".

In sentence selection, BM25 is used in finding the best matching sentence based on the content terms in the question and the text. The motivation on why choose BM25 is that -- refer to the background of this project each question will be corresponded with a relevant

document；thus, if the best matching sentence can be found within the relevant document, the probability to find the correct answer will be improved significantly; BM25 can significantly help with this target. In this project, two approaches in finding the best matching sentence using BM25 were tried. The first one is using BM25 to find best matching paragraph within the corresponding document, and then use BM25 to find the best matching sentence within the paragraph; the other one is directly using BM25 to find the best matching sentence within the document.



Graph 2-2

In answer classification, two approaches were implemented in determining the predicted NER type for each question, which are rule-based approach and C-LSTM neural network approach. The motivation of using rule-based is: for some types of questions there are some expected answer types for them. For instance, question type “where is the capital …” would expect an answer with NER type “City”. Graph 2-2 shows the working mechanism of another approach – C-LSTM neural network approach proposed by Zhou (2015). The reason on why choose this approach is based on the experience learned from many academic papers, which is neural network models have been demonstrated to be capable of achieving remarkable performance in sentence and document modelling. When the predicted answer type and the best matching sentence are derived, the last step is using an algorithm based on distance to rank to candidate answers and then return the final answer to the question.

3. Discussion of results

3.1 Error analysis of basic system & Error Identification

There are three types of errors responsible for producing an incorrect answer: sentence retrieval error, system parameters error, and answer classification error. Sentence retrieval error produces wrong answers due to wrongly predict the paragraph or sentence where the correct answer is. The basic system applies BM25 on paragraphs in given document and questions and selects answers from the paragraph with best performance. This may result in incorrect answer locating since there may exist distracter in a long paragraph with numerous sentences and tokens. Table 3.1-1 presents two examples generated by this basic system. Q303 and Q108 are respectively wrongly answered due to incorrect paragraph and sentence locating. So it is necessary to apply BM25 on sentences in the given document to avoid impact of confusing wrong answers.

ID	Predicted para ID	Corr. Para ID	Predicted answer	Corr. answer
Q303	36	47	two	37 percent
Q108	36	36	u.s.	yucca mountain

Table 3.1-1

System parameters error refers to poor quantitative performance of selected model. The effectiveness of BM25 retrieval function is mainly affected by its sub-linear term frequency (TF) normalization component, which is controlled by k1. However, it has so far been unclear how to interpret parameter k1 (Lv and Zhai, 2012). So it is essential to implement different parameters of BM25 compared to the basic system with defaults k1 = 1.2 and b = 0.75. Figure 3.1-2 demonstrates different performance in terms of mean average accuracy of applying different k1and b. Finally, the best performance is achieved when k1 = 0.1 and b = 0.9. However, due to the limited time they were not applied in the model; if so, the accuracy in this step is improved to 0.805825 from 0.75.

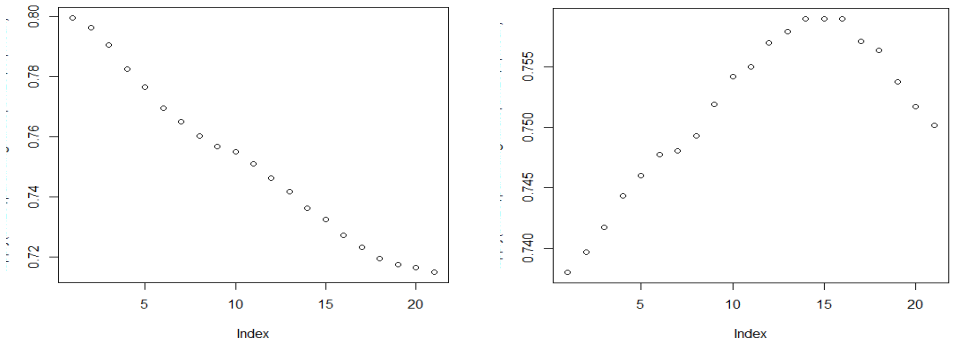


Figure 3.1-2

Answer classification error means wrongly answer questions due to the wrongly predicted answer type. The basic rule-based system identifies the answer types in a less specific way, where possible answer types are classified into LOCATION, PERSON, ORGANIZATION, NUMBER, MONEY, PERCENT, TIME or O based on relatively general keywords identified in questions. Two examples in dev dataset are selected in table 3.1-2. For instance, for question of “Where does uranium rank among elements in terms of its abundance in the Earth's crust?”, the system tags answer type as LOCATION and O based on question keyword of “where”. Apparently, “where” is misinterpreted in this case with ranking. So there are many rules need to be further identified to more precisely predict answer type.

Corr. answer	Previous predicted answer type	Question keywords
51st	LOCATION O	where
french	O PERSON LOCATION NUMBER	what

Table 3.1-3

3.2 Correctness of technique & Enhancements of Results & Results Exposition

For more precise sentence retrieval, improvements are carried out by applying BM25 on paragraphs and then sentences with respect to query rather than only on paragraph. Figure 3.2-1 compares performance of applying BM25 only on paragraphs, on paragraphs and then

sentences, and only on sentences using accuracy in finding correct paragraph and correct answer on the first ten percent of dev dataset as well as F-score on the Kaggle. Here is a great enhancement when applying BM25 only on sentences with respect to the query.

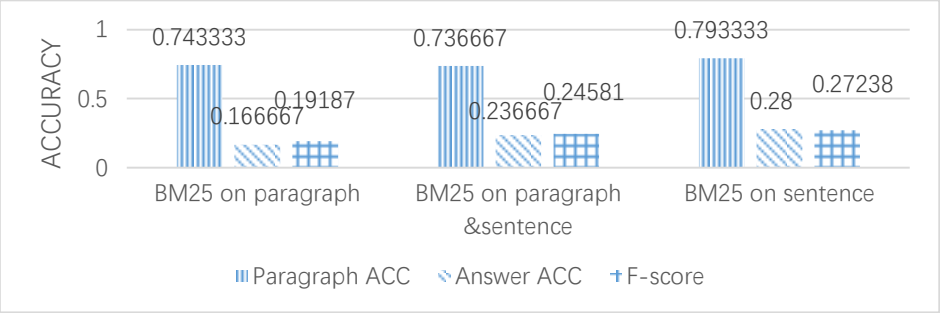


Figure 3.1-1

To more appropriately predict answer types, two approaches are used including manually rule-based method and trainable C-LSTM as stated above. The manual rule-based system defines answer types more precise using more tags. For example, the question will be tagged as ORDINAL according to keyword of "rank" rather than "where" shown in table 3.2-1. Figure 3.2-2 presents different performance using different systems. C-LSTM is trained for better tag the answer type, but performs not well as the manual rule-based system, which is finally adapted.

Corr. answer	Modified predicted answer type	Question keywords
51st	ORDINAL	rank
french	NATIONALITY PERSON	Of what nationality

Table 3.2-1

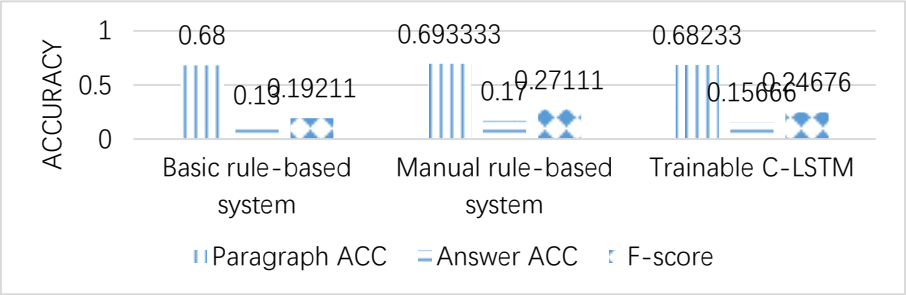


Figure 3.2-2

4. Conclusion

Although the rank in the private Leaderboard improved significantly (from 44th to 19th), a higher rank (top ten) will be accomplished if more time permitted. One solid fact is that the best BM25 parameters can be used in the QA model and then achieve a higher accuracy. In addition, another possible improvement could be that, since the most of the expected answers are NNP, then using a CFG package to obtain all the NNP in each sentence and then using the trained LSF-SCNN model (Guo, 2017) to return the best matching NNP for each question. To be more ambitious, if a knowledge base (based on Wikipedia) can be constructed and then applied in the answer selection, the F-score would be dramatically increased (Diefenbach, 2017).

5. References

Diefenbach, B., et al., 2017, *Core Techniques of Question Answering Systems over Knowledge Bases: a Survey*, viewed at 26 May 2018 at https://www.google.com.hk/url?sa=t&rct=j&q=&esrc=s&source=web&cd=6&cad=rja&uact=8&ved=0ahUKEwj508KU3afbAhUGTLwKHV7iBQwQFghrMAU&url=http%3A%2F%2Fwdaqua.eu%2Fassets%2Fpublications%2F2017_KAIS.pdf&usg=AOvVaw3qMrkAwcneRCJnOvC2OAjP

Guo, J., et al., 2017, *An Enhanced Convolutional Neural Network Model for Answer Selection*, viewed 14 May, 2018 at https://dl.acm.org/citation.cfm?id=3054216

Lv Y., Zhai C. (2012) A Log-Logistic Model-Based Interpretation of TF Normalization of BM25. In: Baeza-Yates R. et al. (eds) Advances in Information Retrieval. ECIR 2012. Lecture Notes in Computer Science, vol 7224. Springer, Berlin, Heidelberg

Zhou, C., Sun, C., Liu, Z., and Lau, F., 2015, *A C-LSTM Neural Network for Text Classification,* viewed 17 May, 2018 at <https://arxiv.org/abs/1511.08630>