Duplicate Question Pair Detection in Natural Language Processing

Raiful Hasan, Mohammad Aminul Hoque
{raiful, mahoque}@uab.edu
Computer Science
The University of Alabama at Birmingham

Abstract

Identifying the duplicate questions is challenging, the sentence composition and word selection vary among persons. There is a way to find the document similarity based on the degree of overlapping in classic natural language processing. However, in duplicate question detection, it not perform well, because most of the questions are too small and the degree of overlapping is insufficient. We have implemented an LSTM based architecture that will detect the duplicate question if they have the same intent.

Background

- Determine the duplicate question is important in the large question-answering platform, such as Quora, Stack-overflow, Reddit, etc.
- There is a possibility that a user asking a question that has already been asked before and answered.
- The applications can suggest previously answered question.
- It could improve the usability of the platform, reduce the waiting time for the user as well as reduce the database space.
- Deep learning techniques in natural language processing (NLP) have made considerable progress in recent years.
- We have used Siamese LSTM based NLP implementation to detect the sentence intent.

Dataset

- **Quora** question pair dataset
- Contains publicly available Quora question
- The dataset has been labeled manually by humans.

id	qid1	qid2	question1	question2	is_duplicate
25	53	54	What is web application?	What is the web application framework?	0
139	279	280	What is the ideal life after retirement?	What's life after retirement?	0
197	395	396	What are some must watch TV shows before you die?	Are there any must watch TV shows?	1
221	443	444	What is my puk code?	What's the PUK for TF64SIMC4?	1

Table 1: Sample data in Quora dataset

Attribute	Values
Total number of entries	4,04,290
Total duplicate pairs	149302 (37%)
Training set	3,04,290
development set	50000
Testing set	50000

Table 2: Attributes of the dataset

Fields	Description
id	unique id for each question pair
qid1	the id for question 1 in the pair
qid2	the id for question 2 in the pair
question1	the full text for question1
question2	the full text for question2
is_duplicate	1- if questions are duplicate or 0 – if questions are not duplicate

Table 3: Database field description

Method

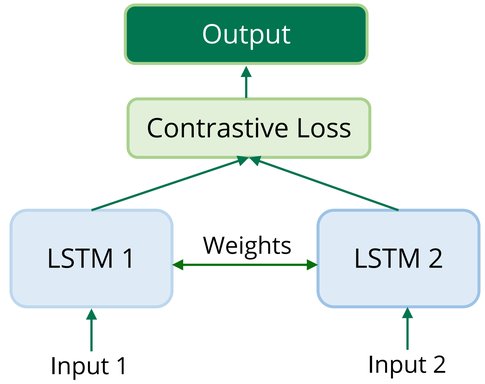

Figure 1: High-level architecture of our implementation

Figure 2: Workflow of the system

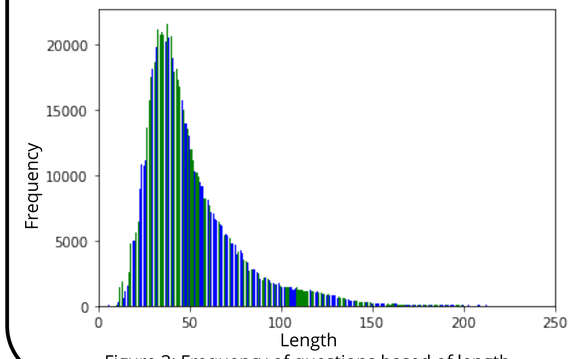

Figure 3: Frequency of questions based of length

Results

- Accuracy – 82.65%

Figure 2: Confusion Matrix

Conclusion

- We have implemented LSTM based architecture to detect the duplicate question.
- We have run 20 epoch and achieved 82.65% accuracy.
- We have used the pre-trained "GoogleNews-vectors-negative300" data for word embedding.
- In the future, we will implement BERT.

References

1. Mueller, Jonas, and Aditya Thyagarajan. "Siamese Recurrent Architectures for Learning Sentence Similarity." *AAAI*. 2016.
2. Sanborn, Adrian, and Jacek Skryzalin. "Deep learning for semantic similarity." *CS224d: Deep Learning for Natural Language Processing. Stanford, CA, USA: Stanford University*(2015)
3. Zarrella, G., Henderson, J., Merkhofer, E. M., and Strickhart, L. (2015). "*MITRE: Seven systems for semantic similarity in tweets.*" In Proceedings of SemEval.
4. Quora question pair; https://www.quora.com/q/quoradata/First-Quora-Dataset-Release-Question-Pairs
5. Stackoverflow; https://stackoverflow.com
6. Keras Documentation; https://keras.io
7. Medium Data science; https://medium.com