Semantic Question Matching - Is That a Duplicate Quora Question?

Megha Arora, Prakruthi Prabhakar
School of Computer Science, Carnegie Mellon University

Motivation

- Building a good knowledge base on Quora
- Text similarity applications: information retrieval, document clustering, topic detection, question answering, machine translation, etc.

Problem Definition

Given a pair of questions, determine the semantic similarity between the pair of questions.

$$f = P(sim(Q_1, Q_2))$$

Data

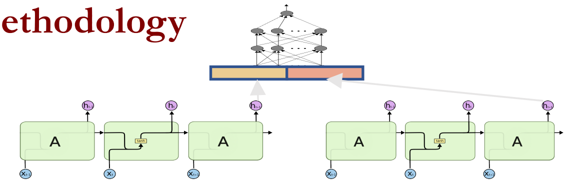
- Quora question similarity pairs Dataset[1] for training and evaluation.

id	Question1	Question2	Y
34	What are some of the best romantic movies in English?	What is the best romantic movie you have ever seen?	1
58	Why did harry become a horcrux?	What is a Horcrux?	0

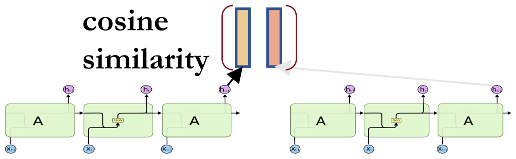
- Data statistics

Dataset size	404,351
#Duplicate questions	149,306
#Non-duplicate questions	255,045
Average question length	11.08

Methodology

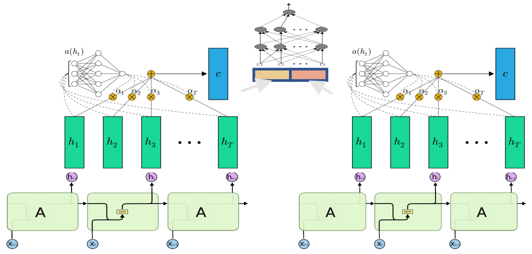


Question 1 Question 2

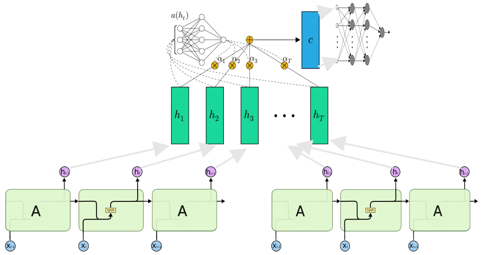


cosine similarity

Question 1 Question 2



Question 1 Question 2



Question 1 Question 2

Results

Method	Accuracy
LSTM with concatenation	0.8262
Bi-LSTM with concatenation	0.8287
Bi-LSTM with cosine similarity	0.7856
Attention-based Bi-LSTM for each question	**0.8320**
Attention-based Bi-LSTM questions concatenated	0.81533
Bi-Multi Perspective Matching (SOTA)*	0.88

Error Analysis

Question1	Question2	Ytrue
Out-of-vocabulary words for embeddings		
What is the Salman Khan?	Who is salman khan?	1
Limited difference in questions		
How do I contact Amazon?	How do I contact Amazon jobs?	0
Erroneous labels?		
what to do in my life, I donno?	What should I do with my life?	0

References

- Wang, Zhiguo, Wael Hamza, and Radu Florian. "Bilateral multi-perspective matching for natural language sentences." *arXiv preprint arXiv:1702.03814* (2017).
- Raffel, Colin, and Daniel PW Ellis. "Feed-forward networks with attention can solve some long-term memory problems." *arXiv preprint arXiv:1512.08756* (2015).

[1] https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs
* Not implemented

Carnegie Mellon University