Learning of Semantically Equivalent Questions

Harini Shreedhar
has099@g.harvard.edu

Siva Rama Krishna Kottapalli
sik546@g.harvard.edu

Section 1: Abstract

Detecting duplicate questions or semantic equivalence between a pair of sentences is an active research area in NLP. New neural network architectures are being tested on this age-old problem to effectively identify the semantic equivalence between pair of sentences/questions.

Much of the motivation for choosing this topic is coming from the usefulness of resorting to Duplicate Question Detection (DQD) to support online question answering community forums, and also conversational interfaces, in general.

DQD falls under the broader task of semantic text similarity (STS). These challenges address a variety of STS subtasks, such as plagiarism detection, comparing machine translation output with a post-edited version, paraphrase detection, among several others. What these STS subtasks have in common is that, when given two input segments, the systems must rate their semantic similarity in some scale, ranging from total semantic equivalence to complete semantic dissimilarity.

For this project, the dataset was provided by Quora. We have implemented 2 models and resorted to a Siamese network in both these models. The first model is based on a Deep Convolution Neural Network and the second is based on a Bidirectional LSTM with Bahdanau Attention. In these two neural networks, various similarity measures such as Manhattan distance, Euclidean distance, Weighted Euclidean distance, Cosine Similarity etc. were used to determine the semantic equivalence between pairs of questions. We have also explored data augmentation technique to generate more positive samples for our project.

Dataset Link: Quora, Kaggle

YouTube Link: here