Abstract: Answer Similarity Analysis using Convolutional Neural Networks

Submitted to: Annajiat Alim Rasel
Senior Lecturer,
BRAC University
CSE713

Submitters:(4 members of Group 17)

Jahid Hasan Mamun	21166047
S M Ibrahim	21166045
Md Tarikul Islam	21166020
Md. Mofaqkhayrul Islam Mim	20266013

Abstract

Day by day online education is gaining much popularity. Moreover, there has been a sudden significant increase in the adoption of online education due to the pandemic. Furthermore, a key part of online education is assessment. Because of the internet and digital data, it is almost effortless to be able to copy answers from different sources, or students may share their responses even to conceptual questions. Thus there can be lots of possibilities of getting similar answers with varying degrees of modifications. For the course instructor, it is very challenging to review the exam papers for such duplicate data, as these data are intentionally modified. This study looks at the similar answer detection in the Kaggle Question Answer dataset. This study extensively explores the dataset and utilizes machine learning models, including neural networks. For word embedding and detecting the semantic similarity of answers, Convolutional Neural Networks (CNN) have been used. A comparative study has been done to find out which neural network model gives the best performance followed by error analysis. These findings can help the evaluators in assessment in the online academic examinations.

Duplicate answers

Answer1: The aim of the system designed in this study, is to detect the similarity of questions in online forums. The data set used in this study is the Question Pairs data set released by the Quora online forum. We explore the data set which include the syntactic structure of the question and also look at the data balance.

Answer2: The purpose of the technique developed in this study is to detect question similarities in online forums. The Question Pairs data collection, which was released by the Quora online community, was used in this study. We look at the data set, which includes the question's grammatical structure as well as the data balance.

Preprocessing and Pretained Word Vectorization

Pretrained Word Embeddings are the embeddings learned in one task that are used for solving another similar task.

Text to word vectorization tools:

1. Google’s Word2Vec
2. Stanford’s GLoVe ✓

Word vector will be used to format text documents with varying sizes according to the dimensions of the word vectors.

CNN

1. Convolution layer - Used CNN to create the model
2. Pooling layer - Max Pooling method for finding most important features.
3. Multi-layer Perceptron layer - The output of this becomes the input, then classified against the other.