CS584: DETECTION OF DUPLICATE QUESTIONS

Raj Shah[1], Kush Jani[1]
[1]Stevens Institute of Technology
rshah97@stevens.edu, kjani1@stevens.edu

ABSTRACT

In this project, we are going explore the different methods to determine the duplication between the pairs of questions. Generally two questions asking the same thing, there are high chances that they can be too different in terms of vocabulary and structure. In the online QA forums, people ask questions but a huge number of questions ends up having the same answer but the expression and words used makes it another question and as a result the question becomes duplicate of the first asked questions.

1 Introduction

There are many QA forums and community sites like Yahoo, Quora and Stack overflow. Detecting duplicate type of questions has become very difficult and challenging for these online forums to keep the answers same for same questions. We can define two questions the same if that two questions have or they can be answered in the same manner. This issue comes up with the huge number of users visiting these websites, making it very hard to have a similar worded questions. Effectively detecting duplicate questions not only saves time for the users but also gives the best answer possible for the same type of questions. This also gives benefit to the person or the expert who is writing the answer in terms of answering the questions multiple times. Therefore we need to build a model that is needed to detect the same type of questions in these websites. In this project, we will try to determine how to best use Neural Networks to identify the duplicate questions.

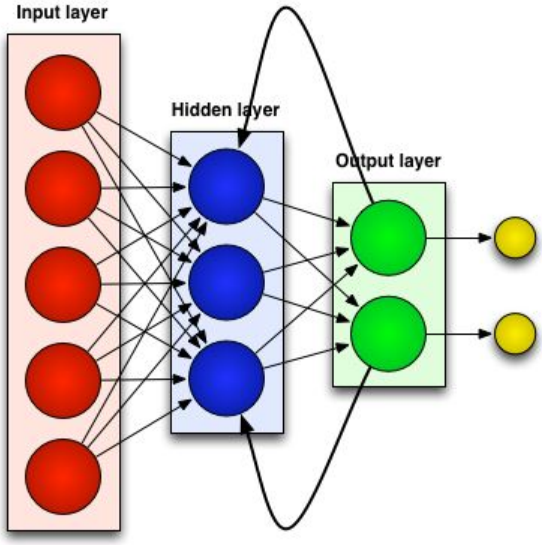
2 Related Work

Detecting duplicate questions is a very major problem in NLP. For this project we have taken into reference two papers which are related to our work. the first paper we have referred is by Zihan Chen et al. where they have used traditional Machine learning algorithms such as Support Vector Machine (SVM) and many other features and extensively pre processed the data. They also argued that the performance of the deep learning methods heavily limited to small and noisy data trained on. Nowadays, because of the recent increase in Deep Learning techniques they got a result that most of these type of topics if we want to solve it, it relies on "Siamese" neural network architecture where the input is taken as two sentences and encodes them using the same neural network. The second paper we have take into reference is of Bogdanova et al. They used two output vectors and used some distance mertic. This approach they found that when pairing a Convoluted neural Network (CNN) with a cosine similarity distance measure gave far more better results as compared to the old fashioned methods like Jaccard Similarity. They noticed this pattern in a Stack overflow dataset. The third paper that we have taken into refernce is the paper of Wang et al. which is the only result oriented publication till date on the Quora dataset.They observed that the Siamese encoding network does not provide interaction between the input sentences and hence they have instead proposed a multi perspective LSTM model.

3 Dataset

The dataset that we are going to use is a Quora Duplicate Question Pairs Dataset. The dataset consists of six columns like id, question1, question2, pair or not. The dataset consist of around 400,000 rows. This dataset was taken from kaggle and the entire dataset is labelled and have enough data.The dublicate columns has two values i.e. if the questions are duplicate then it will have a 1 or 0 if the questions are not duplicate.

4 Methodology and Evaluation Plan



The main thing that we are going to use for our project is RNN.We will be using recurrent neural networks for converting our two pairs of questions into hidden layers. After which we will be using some linear transformations and calculate the prediction of the question being duplicate of each other or not. The majority of our project will be of the same layout, but we are planning to also try to add other pairs of questions by combining and augmenting our dataset and adding and creating more non-duplicate pairs of questions by modifying the lines of the existing dataset. Since we have a dataset which is labelled we plan to have enough data that we can split into a training and testing set. We are also planning to evaluate the effectiveness of our models by testing its success on a predesigned set. We will be showing the results in the form of tables with success rate of each algorithm we use.

References

[1] Dasha Bogdanova, Cicero dos Santos, Luciano Barbosa, and Bianca Zadrozny; "Detecting semantically equivalent questions in online user forums.", 2015.

[2] Zhiguo Wang, Wael Hamza, and Radu Floria;"Bilateral multi-perspective matching for natural language sentences",2017.

[3] Zihan Chen, Hongho Zhang, Xiaoji Zhang, Leqi Zhao; "A paraphrase and semantic similarity detection system for user generated short-text content on microblogs",2016.

[4] Image reference form the link: https://towardsdatascience.com/understanding-recurrent-neural-networks-the-prefered-neural-network-for-time-series-data-7d856c21b759