Sentence Similarity on Quora Question Pairs

Ayça Meriç Çelik - *21627103*, Asma Aiouez - *21504074*, Okan Alan - *21526638*

I. Introduction

IN this project, we will work on one of the most popular NLP topics: Sentence Similarity. After the great success of word embeddings and their applications, we made great progress in understanding natural languages. However, the challenge of evaluating the relations between sentences remains. There are a lot of promising approaches to this problem, and we aim to investigate them. We will start with the BERT model as our baseline, then we will try to increase the performance. If we have enough time left, we are planning to experiment with the newest papers and the current state-of-art. During this research, we are expecting to advance our knowledge on this exciting topic. We believe that we will have satisfying outcomes on our dataset at the end of the semester.

II. Dataset

We will use the “Quora Question Pairs” dataset in our research. The dataset consists of more than 400,000 pairs of questions. In each line, there is pair id, the full text for each question, and a binary value that indicates whether the line truly contains a duplicate pair.

The people from Quora explained the importance of this work for their business: “At Quora, an important product principle is that there should be a single question page for each logically distinct question. For example, the queries “What is the most populous state in the USA?” and “Which state in the United States has the most people?” should not exist separately on Quora because the intent behind both is identical. Having a canonical page for each logically distinct query makes knowledge-sharing more efficient in many ways: for example, knowledge seekers can access all the answers to a question in a single location, and writers can reach a larger readership than if that audience was divided amongst several pages.”

As students/researchers, this dataset is a great opportunity to work on organic data. We can also compare the results with others easily since it is a widely used dataset.

III. Related Works

- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
- ALBERT: A Lite BERT for Self-supervised Learning of Language Representations.
- Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.
- RoBERTa: A Robustly Optimized BERT Pretraining Approach.
- Poly-Encoders: Architecture and Pre-Training Strategies for fast and accurate multi-sentence scoring.

IV. Schedule

After deciding upon the dataset, we plan to have our Project Proposal submitted by April 26th. Then, we will run our baseline model on our dataset to come up with the initial results of our work to eventually compare it with more complex and advanced models by the end of the project. After the initial outcomes, we’ll opt to work on increasing performance.