Project Proposal

Defne Tuncer
Dept. of Computer Science
Hacettepe University
Ankara, Turkey
defnetuncer@hacettepe.edu.tr

Kutay Barcin
Dept. of Computer Science
Hacettepe University
Ankara, Turkey
kutaybarcin@hacettepe.edu.tr

Baran Ekin Ozdemir
Dept. of Computer Science
Hacettepe University
Ankara, Turkey
baranekinozdemir@hacettepe.edu.tr

I. Introduction

Recently, there is an explosion in the information on the internet due to popularization of social media, social news platforms and question-answer sites. These kind of platforms produce massive amounts of textual data as questions, answers, news, comments, ideas and so on. A very large amount of this data is in natural language form. This makes NLP techniques crucial to make the use of this amount of data.

Being massive amount, this data is very difficult to store, process and search on. People often can not easily find what they are looking for or have no time to read it all. This situation causes generation of duplicate information as people ask similar questions, give similar answers and write similar comments unaware of other.

Duplication in this data makes storing, processing and searching operations even more difficult and is a problem that needs to be solved by machines since velocity is too high to classify manually.

Sentence similarity is an important area of NLP that has been trying to provide a solution to above problem of duplicate information. Estimating similarity between two sentences has many applications such as semantic search, question answering, document classification, sentiment analysis, and plagiarism. There are many techniques to measure lexical and semantic similarity between sentences.

In this project, we propose a machine learning approach to measure semantic similarity between two questions. Classifying if the one question has the same meaning seeking the same answer can greatly benefit people using question-answer platforms as finding what they are looking for would be easier.

II. Dataset

Quora Question Pairs [1] is a Kaggle Competition, which challenges participants to tackle the natural language processing (NLP) problem of identifying duplicate questions. The goal of this competition is to predict which of the provided pairs of questions contain two questions with the same meaning. The ground truth is the set of labels that have been supplied by human experts.

III. Related Work

In this work [2] the same dataset, Quora Question Pairs, is studied. Siamese Recurrent neural networks (RNNs) and Manhattan Long short-term memory (LSTM) models are applied, and they obtained a log loss score of 0.28446. In this paper [3] an enchanced recurrent convolutional neural network (Enchanced-RCNN) model is proposed for learning sentence similarity, and compared the model's complexity with the BERT model. Another study [4] focuses on Siamese Recurrent architectures for learning sentence similarity and they achieve as high scores as common LSTM methods by using Support Vector Machines (SVM) with MalSTM fetures. Finally, the Kaggle contest winner team of Quora Question Pairs [5] explains the steps they have taken such as the features; embedding features, classical text mining features and structural features, and the models; Siamese and Attention Neural Networks. In the end, they obtained the log loss score of 0.11579.

IV. Schedule

Date	Milestones	Primary Objective
26 Apr	Project Proposal	Research of the topic, related works and datasets
28 Apr	Meeting	Discussion of baseline methods to implement
12 May	Meeting	Implementing baseline methods
21 May	Progress Report	Comparing baseline methods
26 May	Meeting	Discussion of deep learning method(s) to implement
9 June	Meeting	Implementing deep learning method(s)
16 June	Meeting	Improving models
18 June	Final Report	

References

[1] https://www.kaggle.com/c/quora-question-pairs

[2] Chen, Z., Zhang, H., Zhang, X., Zhao, L. (2018). Quora question pairs.

[3] Shuang Peng, Hengbin Cui, Niantao Xie, Sujian Li, Jiaxing Zhang, and Xiaolong Li. 2020. Enhanced-RCNN: An Efficient Method for Learning Sentence Similarity. In Proceedings of The Web Conference 2020 (WWW '20). Association for Computing Machinery, New York, NY, USA, 2500–2506. DOI:https://doi.org/10.1145/3366423.3379998

[4] Mueller, J., Thyagarajan, A. (2016, March). Siamese recurrent architectures for learning sentence similarity. In thirtieth AAAI conference on artificial intelligence.

[5] https://www.kaggle.com/c/quora-question-pairs/discussion/34355