Detecting Duplicate Questions

Abhishek Patria
Georgia Institute of Technology
abhishek.patria@gatech.edu

Kunaal Ahuja
Georgia Institute of Technology
kunaal@gatech.edu

Arunachalam Muthu Valliappan
Georgia Institute of Technology
avalliappan3@gatech.edu

Abstract

Identifying duplicate questions in Quora using Semantic Parsing. Different model architectures like BERT, Attention based LSTMs, Siamese LSTMs and CRFs are explored and compared to build an ensemble.

1 Motivation

Quora is a question and answer platform where 100 million people visit every month. With such a huge user base it is imperative that a lot of similar questions are asked. The problem of interest is to identify questions having similar meaning or intent and to merge the duplicate or similar questions.

This would not only help curb redundancy but help the user to find a collection of similar answers in one go. Moreover, this would also promote Quora's value of developing a creative knowledge base.

In fact, this is a problem which is very evident on other social Q&A platforms like StackOverflow. Solving this problem would enable users to find questions that have already been answered and prevent community members from answering the same question multiple times. The solution can be further extended to implement automatic short answer grading systems, essay grading system and textual entailment detection problems as well.

2 Related Work

Previous work on semantic relatedness of sentences has focused on logical inference and entailment based on the Stanford Natural Language Inference Corpus 2 . The first of these papers which focused on attention methods using LSTMs was Rocktaschel et al. which introduced word-by-word attention methods with the hypothesis attending on the premise.

Classifying questions as duplicate can be quite subjective because the true meaning of a sentence is very difficult to be known with certainty.

Classifying short texts as duplicate is similar to the problem of record linkage, deduplication etc. Databases often have the same records and field values which are not syntactically identical but refer to the same entity. This is known as record linkage and it doesn't let data mining algorithms work efficiently. (Torsten Zesch et al., 2012).

Another widely used approach is to use discriminate models over features produced from minimal edit distances between dependency parse trees. But, the problem of these approaches are that they require a significant amount of feature engineering and require expensive semantic resources.

3 Goal

The overall goal of the project is to identify the intent and meaning of questions in quora to find duplicates among the question pairs and to tag them into different categories. To enable this task, we are planning to explore the following methodologies:

- **Information Extraction** - To extract labels for identifying topics
- **Semantic Parsing** - Extract meaning and intent to identify duplicate questions

The primary challenge with this task is that question pairs which have a lot of the same or similar words might not have the same intent or meaning. For eg. “Are flights made in Seattle?” and “Are there any flights to Seattle?” have many common words but they mean very different things. Models are prone to labelling these questions as similar if there is data leakage or if the model is not complex enough to capture the meaning. Mistakes of the reverse kind also have to be handled, where there are very few common words but still the questions have the same meaning. Some question pairs might be referring to the same entity in different forms or abbreviations. We are planning to use Coreference Resolution methods to identify such entities and use it to help identify the labels and the entity which is being referred to in the question.

In addition to identifying the labels, extracting such information will enable us to cluster the questions into different groups and to track the evolution and the nature of interest among different topics or entities.

Exploring the key features used in identifying duplicates and identifying the circumstances under which the model's performance is unsatisfactory will enable us to interpret the functioning of the model and identify the directions for future improvement.

4 Plan

Data: To focus the majority of our efforts in model development and analysis, we are using a Supervised Kaggle Dataset. The data is instantly available from the given source.

Data Source:
https://www.kaggle.com/c/quora-question-pairs/

Models: We are planning to explore BERT, Bi-LSTMs, Siamese LSTMs, Attention based LSTMs and CRFs

Resources: We are planning to use Google Colab for training our models. Additional computing resources including GPUs in Colab would accelerate the project significantly

We are using open source models and data to ensure availability. Moreover, if the models proposed turn out to be less useful, we will try to improve upon the vanilla version of BERT. We are planning to explore multiple models simultaneously to decrease the chances of failure. We can broadly classify our approach into 3 buckets.

- **New Architectures:** Since BERT is a sentence representation model and most quora questions are typically less than 2-3 sentences, we expect this architecture to be effective in identifying the similarities between questions.

- **New LSTM Variants:** Some advanced variants like Siamese LSTMs which contain two identical networks will be explored, since the problem at hand is to identify duplicates in question pairs. Attention based LSTMs are also effective in processing longer sequences.

- **Simpler and Interpretable Models:** The safest option among the three is to build simpler and more interpretable models. For this, we are planning to build a POS tagger with Conditional Random Fields and use information extraction to fill templates and compare the templates to identify similarities in question. This will enable us to exactly identify where the model is underperforming and provide us with additional structured data to help debugging and improving the models.

5 Timeline

Data Collection: Completed

Model Development: 16th March - 7th April (3 Weeks) - Midterm report

Analysis and Feedback implementation: 7th April to 15th April (1 Week)

Final Analysis and Report: 16th April to 20th April (1 Week) - Final Report

6 Workload Distribution

We are planning to train and evaluate 3 different types of models mentioned above independently for the first phase of the project and identify the pros and cons of the different models. We will compare the results of these three models and see how they can be cascaded to solve the deduplication. In Phase II we are planning to consolidate the 3 models to generate the final outputs.

7 References

[1] Ankur Parikh, Oscar Tackstr om, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In Proceedings of EMNLP

[2] Viswanathan S., Damodaran N., Simon A., George A., Anand Kumar M., Soman K.P. (2019) Detection of Duplicates in Quora and Twitter Corpus.

[3] In: Peter J., Alavi A., Javadi B. (eds) Advances in Big Data and Cloud Computing. Advances in Intelligent Systems and Computing, vol 750. Springer, Singapore

[4] Mueller, Jonas, and Aditya Thyagarajan. ”Siamese Recurrent Architectures for Learning Sentence Similarity.” AAAI. 2016.