Dauphine | PSL
UNIVERSITÉ PARIS

MASTER IASD

Natural Language Processing (NLP) PROJECT

Quora Question Pairs

Realised by:

BAZ Roland

BENCHEIKH LEHOCINE Mohammed Amine

DJECTA Hibat Errahmen

Supervised by:

Prof. ALLAUZEN Alexandre

2022/2023

Table of Contents

List of Figures

1 Introduction

Natural Language Processing (NLP) is an area of computer science and artificial intelligence that focuses on enabling computers to understand, interpret and generate human language. It has many applications, such as machine translation, speech recognition, sentiment analysis, text classification and chatbot development.

One of the main applications of natural language processing is duplicate questions recognition.It is a challenging task because the same question can be expressed in many different ways, using different words or phrasing and can still convey the same meaning. This makes it difficult for traditional keyword-based techniques to identify similar questions accurately. To overcome these challenges, NLP researchers have developed sophisticated algorithms that can analyze the syntactic and semantic features of questions and compare them with a large corpus of existing questions to identify duplicates.

In recent years, the development of large-scale datasets and the availability of powerful computing resources have enabled significant progress in the field of duplicate question recognition. As a result, many companies and organizations are now using NLP-based systems to identify duplicate questions and improve the efficiency and accuracy of their information retrieval and community question answering systems.

The report begins by introducing the data set employed in the project, followed by a comprehensive examination of the objective at hand. Subsequently, the report elaborates on the strategy used in the analysis and ends up by providing a conclusion.

2 Data Description

Quora is a platform that allows people to gain and share knowledge about any topic by asking questions and connecting with individuals who offer unique insights and quality answers. With over 100 million monthly visitors, many users ask similar questions, leading to duplication and confusion. Quora aimed to improve its duplicate question detection process and therefore challenged data scientists, machine learning experts and AI researchers to find advanced techniques to classify whether question pairs are duplicates or not.

The dataset provided was extensive, consisting of over 400,000 question pairs. Each pair was assigned a label indicating whether they are duplicates or not. The fields in the dataset provided by Quora are:

- **id** - the id of a training set question pair
- **qid1, qid2** - unique ids of each question (only available in train dataset)
- **question1, question2** - the full text of each question
- **is duplicate** - the target variable, set to 1 if question1 and question2 have the same meaning and 0 otherwise.

The objective is to train the models using a train dataset and then evaluate the performance on a separate set of question pairs unseen during training. The primary obstacles encountered involve the imbalanced nature of the Quora dataset, where less than half of the question pairs are labeled as duplicates. Moreover, the dataset contains a notable amount of noise, as some question pairs labeled as duplicates do not truly have the same meaning. Additionally, certain duplicate questions exhibit typos or grammatical errors, further complicating the task.

test_id	question1	question2
0	How does the Surface Pro himself 4 compare with iPad Pro?	Why did Microsoft choose core m3 and not core i3 home Surface Pro 4?
1	Should I have a hair transplant at age 24? How much would it cost?	How much cost does hair transplant require?

Figure 1: A snapshot of the test dataset provided by Quora

3 Task Description

Textual matching and semantic similarity refer mainly to the degree to which two pieces of text convey the same meaning. This can be measured at various levels, such as word, sentence, or document level [CM21]. It is a challenging open task since it requires understanding the meaning of words in the context of the sentence and the entire text. This involves dealing with the ambiguity of language, where a word can have multiple meanings depending on the context in which it is used. Furthermore, the meaning of a sentence is not only determined by the meaning of its constituent words, but also by their arrangement, the presence of negation or modifiers, and other syntactic and semantic features.

In the context of **Quora question pairs detection**, the task is to determine if two questions have similar meanings or not. This is crucial because Quora aims to have a single question page for each logically distinct question. Having duplicate questions on the platform can lead to redundant content and make it difficult for users to find the information they are looking for [Quo]. Therefore, identifying duplicate questions is essential for maintaining the quality of the platform and improving the user experience.

3.1 Common approaches

One common approach to measure the semantic similarity is to use vector space models such as **Word2Vec** [MCCD13] or **GloVe** [PSM14], which represent words as high-dimensional vectors based on their co-occurrence with other words in a large corpus of text. To compute the similarity between two questions, we can first represent each question as a vector by taking the average or weighted sum of the vectors of the words in the question. Then, we can use a similarity metric such as cosine similarity to compute the similarity score between the two question vectors. The higher the similarity score, the more similar the two questions are in meaning. However, using vector space models such as **Word2Vec** and **GloVe** for sentence embeddings do not capture the **compositionality of language**, which means that the meaning of a sentence cannot be fully captured by the meanings of its constituent words.

Another approach is to use deep learning models such as **Siamese neural networks** [NVR16] or Transformers like **BERT** [DCLT18], which are adaptable for text similarity tasks. These models can capture context and complex relationships between words and phrases and provide more accurate similarity scores. The basic idea in **Siamese neural networks** is to use the same weights of the network for both inputs, allowing the model to learn a common feature representation for both texts. This common representation can then be used to measure the similarity. For example, in the case of Quora question pairs detection, the Siamese network takes in two questions as input

and produces two vectors that represent the questions. These vectors are then compared using a distance metric. Transformer-based models such as **BERT** are pre-trained on a large corpus of text using a masked language modelling objective. This pre-training allows the model to learn contextualized representations of words and phrases, which capture complex relationships between words and phrases. The pre-trained model can then be fine-tuned on a specific task such as text similarity.

3.2 Evaluation metrics

There are various metrics that can be used to evaluate the performance of models on semantic similarity tasks. These metrics include:

- **Correlation-based metrics**

  Used mainly to evaluate the performance of semantic similarity systems by comparing the system's similarity scores with human judgments of similarity. The basic idea is that if a system is accurately capturing the human notion of similarity, then its similarity scores should be highly correlated with human judgments. examples of such metrics are the Spearman correlation, Pearson correlation and Kendall correlation...

- **Classification-based metrics**

  The idea is to consider the task of classifying if two texts are similar or not semantically. this is the objective of the Quora question pairs. As it is known, **Accuracy, Precision, Recall and F1-Score** are commonly used in classification tasks but can be adapted for similarity tasks as well. In this case, the model's predictions are binarized by selecting a threshold, and the precision, recall, and F1-Score are calculated based on these binary predictions. For the **Kaggle** leaderboard, the score refers to the log of the binary cross entropy loss between predicted classes and targets.

4 Approach and project pipeline :

The project followed a systematic approach to address the problem of duplicate question recognition. The key steps undertaken were as follows:

1. **Data Exploration**:

   During the data exploration phase, we aimed to gain insights and understand the characteristics of the dataset. This involved analyzing the distribution of the target variable, exploring the distribution of important features, and visualizing relationships between variables.

   To start, we examined the distribution of the target variable "is_duplicate," which indicates whether a pair of questions are duplicates or not. Understanding the balance between duplicate and non-duplicate pairs is crucial for modeling and evaluating performance. By analyzing the distribution, we ensured that the dataset was adequately representative of both classes.

   Furthermore, we explored the distributions of key features derived from the preprocessed text data. Features such as the common word count, common stopword count, and common token count were investigated to understand their distribution across the dataset. This exploration helped us identify potential patterns or trends that could be indicative of duplicate or non-duplicate question pairs.

   To gain a deeper understanding of the relationships between features, we utilized visualizations such as pair plots. These plots allowed us to examine the pairwise interactions between important features, such as the minimum token count ratio, common word count ratio, common stopword count ratio, and token sort ratio. By visualizing these relationships, we could identify any correlations or patterns that might exist between the features and the target variable.

In addition to exploring individual features, we also investigated the absolute difference in the number of words between question pairs. By analyzing the distribution of this feature, we gained insights into the variation in length between duplicate and non-duplicate questions.

2. **Data Preprocessing**:

   In the data preprocessing step, various techniques were applied to clean the text data. These techniques included removing HTML tags, converting text to lowercase, replacing contractions and numerical representations, and applying stemming. Stopwords, common words that do not contribute significantly to the meaning of the text, were removed using the NLTK library. These preprocessing steps helped to standardize the text and remove noise, making it more suitable for analysis.

3. **Feature Extraction**:

   Feature extraction involved creating new features from the preprocessed text data to capture different aspects of the question pairs. Token-based metrics such as common word count, common stopword count, and common token count were computed to quantify the overlap between the questions. Ratios of these metrics to the minimum and maximum lengths of the questions were also calculated to provide normalized measures. Additional features included checking the equality of the last word and first word of the questions, computing the absolute difference in length, and calculating the average token length.

   These extracted features provided valuable information about the characteristics and similarities between question pairs, enabling deeper analysis and modeling tasks.

4. **Model Development**:

   The model development phase involved building a machine learning model to predict the labels for question pairs. We designed a model architecture that incorporated the extracted features as input. Depending on the requirements of the task and the available resources, we considered various algorithms such as Logistic Regression, Linear SVM, and XGBoost. For instance, we developed a neural network-based model using the BERT (Bidirectional Encoder Representations from Transformers) architecture, which has shown promising results in natural language processing tasks. The model was trained on the labeled training data, and appropriate hyperparameters were selected to optimize performance.

5. **Model Evaluation**:

   To assess the effectiveness of our developed model, we performed model evaluation using suitable metrics. We measured metrics such as accuracy, precision, recall, and F1-score to evaluate the model's performance in predicting duplicate and non-duplicate question pairs. Accuracy measures the overall correctness of the model's predictions, while precision and recall provide insights into the model's ability to classify positive cases accurately. F1-score combines precision and recall into a single metric that balances both aspects. By evaluating the model's performance using these metrics, we gained insights into its strengths and weaknesses, allowing us to make informed decisions about its deployment and potential improvements.

5 Conclusion

The detection of question pairs is a significant task in natural language processing, and in this project, we explored various techniques to address this challenge. By using methods such as semantic embeddings, syntactic analysis, and ensemble models, we aimed to accurately determine if two questions are duplicates or not. We used the dataset provided by Quora for training and for evaluating our models. We performed necessary preprocessing and feature engineering to prepare the data for model training. Additionally, we employed a separate dataset test the performance of our models, utilizing metrics like accuracy, precision, recall, and F1-score.

In conclusion, the successful identification of question pairs can reduce redundancy, improving the efficiency and effectiveness of search results. However, there are still challenges to overcome,

such as imbalanced datasets, semantic nuances and language ambiguity. Continuous and extensive research and development hold great potential for further enhancing this field of NLP.

References

[CM21] Dhivya Chandrasekaran and Vijay Mago. Evolution of semantic similarity—a survey. *ACM Computing Surveys (CSUR)*, 54(2):1–37, 2021.

[DCLT18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[MCCD13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[NVR16] Paul Neculoiu, Maarten Versteegh, and Mihai Rotaru. Learning text similarity with siamese recurrent networks. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 148–157, 2016.

[PSM14] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[Quo] First quora dataset release: Question pairs. https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs. Accessed: 2023-05-13.