Seeing Double: Detecting Duplicate Questions using Sentence-BERT

Marko Opačić, Javier Salvatierra

University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
{marko.opacic,javier.corchado}@fer.hr

Abstract

Lorem ipsum dolor sit amet lorem ipsum dolor sit amet lorem ipsum dolor sit amet lorem ipsum dolor sit amet lorem ipsum dolor sit amet

1. Introduction

Intro

2. Related Works

The problem of identifying duplicate questions has similarities to other NLP tasks, such as paraphrase detection and semantic similarity detection. Traditional ML approach use algorithms such as SVM with hand-picked features including n-gram overlaps, part-of-speech agreement, verb similarity and others (Dey et al., 2016). More recently, deep learning approaches have proved to be very effective in a variety of NLP tasks. Most approaches for detecting sentence similarities use a Siamese architecture, which involves producing encoded representations for each of the two input sentences, which are subsequently processed for classification, usually through some form of distance metric Some examples include the Siamese GRU (Homma et al., 2016), Siamese MaLSTM (Imtiaz et al., 2020) and Siamese BiLSTM (Fradelos et al., 2023). Convolutional neural networks (CNNs) are leveraged by several authors, also in a Siamese setting (Bogdanova et al., 2015; Prabowo and Herwanto, 2019).

The Quora Question Pairs dataset by Iyer et al. (2017) has been a valuable resource for research in this area. As part of a Kaggle competition using the dataset, a Siamese LSTM architecture with attention was the winning solution (Dadashov et al., 2017). The winning Siamese LSTM architecture is described in Section 4.1.2..

In the past few years, pretrained Transformer models have set a new benchmark on sentence-pair regression tasks (Devlin et al., 2018). Although the original BERT model performs well on such tasks, it has some limitations, such as requiring both sentences as inputs to determine their similarity, which results in a large computational overhead. The Sentence-BERT models significantly reduce the computational load and achieve state-of-the-art results on semantic textual similarity tasks (Reimers and Gurevych, 2019).

3. Data

The Quora Question Pairs (QQP) dataset consists of 404k question pairs, along with labels indicating whether the questions are duplicates or not.

We use the same dataset split as the SBERT paper, which is a 60/20/20 split resulting in 243k training examples, 80k dev examples and 80k test examples (Reimers and Gurevych, 2019). The splits are not stratified, as different percentages of duplicates are present between the sets. The percentage of duplicates per set is presented in Table 1.

Table 1: Dataset split characteristics

Set	No. of examples	Duplicate percentage
Train	243k	37.25%
Dev	80k	35.05%
Test	80k	40.38%

4. Models

In this section, we describe the models used in the experiments. Cosine similarity is used as the simplest baseline, while the winning solution from the 2017 Quora Question Pairs competition on Kaggle serves as the baseline we are attempting to outperform.

4.1. Baselines

4.1.1. Cosine similarity

The input questions are first tokenized using spaCy, after which embeddings are calculated for each token. GloVe embeddings are used for this purpose, specifically the 300-dimensional vectors trained on a corpus 6B tokens from Wikipedia 2014 and Gigaword 5 (Pennington et al., 2014). Aggregate embeddings for each question are then calculated by summing the embedding vectors element-wise. The cosine similarity is then calculated for the aggregate embeddings to obtain a similarity score. The similarity score is fed into a simple binary classifier which produces the final output label.

4.1.2. Siamese LSTM

As a variant of recurrent neural networks (LSTM), it is an approach that compares the similarity between pairs of data sequences. It employs two identical LSTMs that process two input sequences simultaneously and then compares their final vector representations to determine their similarity. We only use the reported results on the dev set for this model for comparison with ou results.

4.1.3. Sentence BERT

Sentence-BERT (SBERT) is a method for creating semantically meaningful sentence embeddings (vector representations) using transformer-based models like BERT. It fine-tunes BERT to generate fixed-size representations of input sentences that capture their semantic meaning. SBERT is trained using a siamese or triplet network architecture, where it learns to map similar sentences closer together in the embedding space.

In our experiments we used the pretrained RoBERTa base model (RoBERTa) described by Liu et al. (2019), as well a distilled version of the same model (DistilRoBERTa). The distilled version uses the same distillation process used for DistilBERT (Sanh et al. (2019).

It is worth noting that with pretrained large language models there is a risk of test set contamination. This means that part of the test set on which the model is evaluated is used during model training. This violates the assumption that the model is evaluated on unseen data, and undermines the integrity of the results. Specifically for the RoBERTa and DistilRoBERTa models used, their training data does not include the QQP dataset, so we have no such concerns.

5. Experiments

In this section we explain the training process for the models used in the experiments. We then provide an overview and discussion of the results.

5.1. Training

5.1.1. SBERT

We used linear warmup as described... Following advice from Gkouti et al. (2024), we focused only on tuning the learning rate. For selecting the learning rate we tested values of 4e-5, 3e-5 and 2e-5, similar to the original BERT paper (Devlin et al. (2018). We then selected the best performing learning rate on the test set.

6. Conclusion

References

Dasha Bogdanova, Cicero dos Santos, Luciano Barbosa, and Bianca Zadrozny. 2015. Detecting semantically equivalent questions in online user forums. In *Proceedings of the nineteenth conference on computational natural language learning*, pages 123–131.

Elkhan Dadashov, Sukolsak Sakshuwong, and Katherin Yu. 2017. Quora question duplication.

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

Kuntal Dey, Ritvik Shrivastava, and Saroj Kaushik. 2016. A paraphrase and semantic similarity detection system for user generated short-text content on microblogs. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2880–2890.

Georgios Fradelos, Isidoros Perikos, and Ioannis Hatzilygeroudis. 2023. Using siamese bilstm models for identifying text semantic similarity. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 381–392. Springer.

Nefeli Gkouti, Prodromos Malakasiotis, Stavros Toumpis, and Ion Androutsopoulos. 2024. Should i try multiple optimizers when fine-tuning pre-trained transformers for nlp tasks? should i tune their hyperparameters? *arXiv preprint arXiv:2402.06948*.

Yushi Homma, Stuart Sy, and Christopher Yeh. 2016. Detecting duplicate questions with deep learning. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS*, pages 25964–25975.

Zainab Imtiaz, Muhammad Umer, Muhammad Ahmad, Saleem Ullah, Gyu Sang Choi, and Arif Mehmood. 2020. Duplicate questions pair detection using siamese malstm. *IEEE Access*, 8:21932–21942.

Shankar Iyer, Nikhil Dandekar, and Kornél Csernai. 2017. First quora dataset release: Question pairs. Accessed: 2024-06-06.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Damar Adi Prabowo and Guntur Budi Herwanto. 2019. Duplicate question detection in question answer website using convolutional neural network. In *2019 5th International conference on science and technology (ICST)*, volume 1, pages 1–6. IEEE.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.