Improve Quora Question Pair Dataset for Question Similarity Task

Huong T. Le
School of Information and Communication Technology
Hanoi University of Science and Technology
Hanoi, Vietnam
huonglt@soi ct.hust.edu.vn

Dung T. Cao
School of Information and Communication Technology
Hanoi University of Science and Technology
Hanoi, Vietnam
dungct@soict.hust.edu.vn

Trung H. Bui
Adobe Research, San Jose,
California, USA
bui@adobe.com

Long T. Luong
School of Information and Communication Technology
Hanoi University of Science and Technology
Hanoi, Vietnam
thanhlonghanam1997@gmail.com

Huy Q. Nguyen
School of Information and Communication Technology
Hanoi University of Science and Technology
Hanoi, Vietnam
huyquang.hust@gmail.com

***Abstract*—Automatic detection of semantically equivalent questions is a task of the utmost importance in a question answering system. The Quora dataset, which was released in the Quora Question Pairs competition organized by Kaggle, has now been used by many researches to train the system in solving the task of identifying duplicate questions. However, the ground truth labels on this dataset are not 100% accurate and may include incorrect labeling. In this paper, we concentrate on improving the quality of the Quora dataset by combining several strategies, basing on Bert, rules, and reassigning labels by humans.**

Keywords— Quora, question answering, BERT, question similarity

I. INTRODUCTION

Among the numerous applications of natural language processing, question answering is a hot and attractive research area with a wide range of commercial potentials. With the explosion of the Web, question answering is a relevant direction to address the information overloading problem. The past ten years have witnessed the emergence and rapid growth of community question answering forums such as Quora [5] and Stack Overflow [7]. Through the years, they have accumulated a large number of questions and corresponding answers. It is no surprise that many people ask similar questions. Because of that, there is a need to find questions that are similar to the user's question from the existing question answering dataset, so that the system can return the answer by retrieving answers from similar questions.

Having been a long-standing problem in natural language understanding, the automatic detection of semantically equivalent questions is now a task of the utmost importance in a question answering system.

Bogdonova et al. [1] introduced a definition of semantic equivalence that was reused through many researches (e.g., [4], [6], etc…) on duplicate question detection: *"Two questions are semantically equivalent if they can be adequately answered by the exact same answers."* For example, this pair of questions is considered duplicated: *"What is the most populous state in the USA?"/"Which state in the United States has the most people?"*

Introduced in 2017 as their first public dataset, the Quora question pairs dataset has been widely used for training models and benchmarking the duplicate question detection methods. It consists of 404351 question pairs labeled by human to indicate whether they are logically duplicate or not. If two questions are similar, they are marked with the label 1, otherwise, label 0 will be assigned. However, several labels in this dataset disagree among humans. As a result, the ground truth labels on this dataset are not 100% accurate and may include incorrect labeling. For example, these two questions in the Quora dataset are quite similar. However, the answers to these questions are not the same:

How do I get Mac Donalds franchise?/How do I get INOX Franchise?

The dataset imperfection may affect the task of training models for recognizing the semantic equivalence, as well as the reliability of the evaluation of any proposed method. To have a more reliable dataset for the question similarity task, we concentrate on improving the Quora dataset by automatically figuring out the incorrectly labeled question pairs and correcting them.

Recently, Bert [3] is a powerful semantic representation model. The prediction of a system using Bert sometimes can be more accurate than a human's assignment. For example,

This work is supported by Adobe Research, Adobe Systems Incorporated.

978-1-6654-0435-8/21/$31.00 ©2021 IEEE

the following question pair is considered as duplicated by humans, but "not duplicated" by a BERT-based system:

> *Is universal health care good? Why or why not?/Should the U.S. implement a universal health care? Why or why not?*

These questions do not ask the same problem. Therefore, the BERT-based system's result is correct in this case.

Therefore, we use Bert to generate a semantic representation of questions and to do the classification task for each question pair. By analyzing the disagreement between Bert-based results and manually annotated labels, we propose rules to filter out incorrectly labeled question pairs in the Quora dataset. Also, we propose a method to correct the labels of these pairs.

The rest of the paper is organized as follows. Section 2 introduces our model using in the task of identifying the similarity between a question pair. Section 3 analyzes the output of our Bert model comparing to the human assigned labels in the Quora development dataset. Section 4 introduces our rule-based method to figure out dissimilarity question pairs. The scenario to reassign labels for question pairs in the Quora dataset is described in Section 5. Finally, Section 6 is our conclusion of the paper.

II. QUESTION SIMILARITY USING BERT

Bidirectional Encoder Representations from Transformers (Bert) [3] is designed to pre-train deep bidirectional representations from an unlabeled text by jointly conditioning on both left and right context in all layers. Its input can be a single sentence or a pair of sentences in one token sequence. The representation for each input token is a combination of token embeddings, segment embeddings, and position embeddings. Token embedding is represented as one-hot vector created from the WordPiece dictionary [8] with 30,000 tokens. Segment embeddings indicate whether the current token belongs to – first sentence or second sentence in a pair of sentences. Position embeddings contain information about the positions of tokens in a specific sentence.

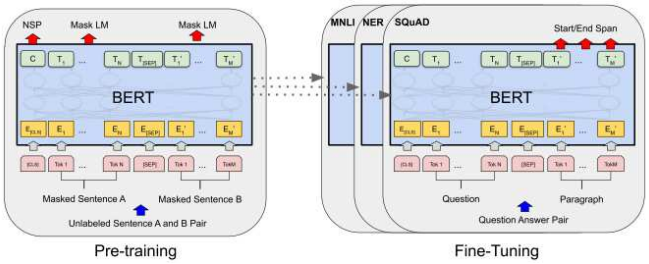

Fig. 1. Overall pre-training and fine-tuning procedures for Bert.

In the original Bert architecture, [CLS] token is added in front of input text. [SEP] symbol is used if it is needed to separate two parts of the input. In the sentence similarity problem, [SEP] is used to separate two questions.

The output of Bert is a sequence that reflects the main features of the two sentences. The features of each sentence are recognized by a pair [CLS] … [SEP]. These two subsequences can be used in a classification model to determine the similarity between two sentences.

The pre-trained Bert model can be used in different tasks by fine-tuning it with the training dataset of that specific task (Fig. 1). To apply the pre-trained Bert for our classification task, we fine-tune Bert with the Quora training dataset (363850 question pairs) with some modifications in the model as follows:

+ Using [CLS] token to mark the beginning of the second question: A question in the Quora dataset can have more than one sentence. If only one [CLS] token is used as in the original model, the [CLS] token in the output can miss some information. Because of that, we represent the input question pair as [CLS] <question1> [SEP] [CLS] <question2> [SEP]. Both [CLS] tokens at the output layer are used in the question classification task.

+ The [CLS] tokens at the last 4 layers of Bert are used as input features for a support vector machine classifier, which returns 1 if two questions have the same meaning, and 0 otherwise.

III. ANALYZING INCORRECTLY ASSIGNED LABELS IN THE QUORA DATASET

The Quora dataset is divided into a training dataset (363850 question pairs) and a development dataset (40431 pairs). The development dataset consists of 255045 negative (non-duplicate) and 149306 positive (duplicate) instances. Each sample in the dataset has the following fields:

1. **id:** id of the sample.
2. **qid1:** A unique number to identify question 1.
3. **qid2:** A unique number to identify question 2.
4. **question1:** Content of question 1.
5. **question2:** Content of question 2.
6. **is duplicate**: get the value of 1 if two questions have the same meaning, otherwise, it is 0

To detect incorrectly assigned labels in the Quora dataset, we analyze cases that the Bert model assigning different labels than that of the Quora. We train the Bert model using the Quora training dataset and test with the Quora development dataset. The results are shown in Table 1 below.

TABLE I. EXPERIMENTAL RESULTS WITH THE BERT MODEL

		Human (is duplicate)	
		Positive	***Negative***
Predicted (Bert label)	Positive	13287	2001
	Negative	1598	23544

The accuracy and F-score of the Bert model, in this case, is 91.10% and 88.07%, respectively.

We filter out question pairs in the Quora development dataset that have different labels than those assigning by Bert. After analyzing these question pairs, we found that there several misclassified labels in the Quora dataset, especially the ones concerning the positive instances. A lot of questions, labeled as duplicates, are just partially or nearly semantically equivalent. Let us consider two sets of question pairs that the Bert model assigns different labels that that of the Quora dataset:

- is_duplicate=0, Bert_label=1 (2001 question pairs). Question pairs in this set are different from each other only a few words. Several cases in this set are similar, but humans assign dissimilar. For example:

  Is unemployment scary?/Why is unemployment bad?

- is_duplicate=1, Bert_label=0 (1598 question pairs). Question pairs in this set look quite similar, but some of them cannot have the same answers as they are different in some important words. For example:

  How do I get Mac Donalds franchise?/How do I get INOX Franchise?

In the above example, two questions ask about different entities (Mac Donalds and INOX), so they should be different. Human assigns incorrect label (1) in this case, and the Bert model gives a correct label (0).

However, the Bert model still assigns label 1 in some cases similar to the above. For example:

What is the Sahara, and how do the average temperatures there compare to the ones in the Gobi Desert?/What is the Sahara, and how do the average temperatures there compare to the ones in the Dasht-e Kavir?

Both labels in the Quora dataset and the Bert model are 1 for the above example. It points out that the Bert model does not deal with such cases completely correct. This is because it learns from the Quora dataset, which is a noisy corpus.

To improve the Quora dataset, we propose a rule-based method to filter out dissimilar questions, using information about named entities appearing in the two questions. This method will be discussed in Section 4.

IV. RULE-BASED METHOD TO IDENTIFY DISSIMILAR QUESTION PAIRS

Many of the incorrect duplicates are due to the model deciding that two questions are duplicates, even if they refer to different named entities. Therefore, we will now define a new label, rule_label, which deals with named entities explicitly, as follows.

Two questions are dissimilar if their named entities are not corresponding. In that case, the rule_label = 0. Otherwise, rule_label = 1, which means their entities are corresponding or the two questions do not contain entities. We use this principle to filter out dissimilar questions.

To detect named entities, we use the named entity recognizer (NER) in the Spacy framework (https://spacy.io/). This tool can recognize the following named entities: GEO (Geographical Entity), ORG (Organization), PER (Person), GPE (Geopolitical Entity), TIME (Time indicator), ART (Artifact), EVENT (Event), NAT (Natural Phenomenon).

The named entities GEO, TIME, ART, and EVENT can be expressed in different ways, which causes the named entity recognizer cannot detect all of these entities. Therefore, we only detect PER, GPE, and ORG, and use these entities to compare two input questions. Another problem with the named entity recognizer is that questions in the Quora dataset are casually written by humans with a lot of errors. For example, abbreviations, ungrammatical words are used freely in text. To deal with this problem, we preprocess input questions automatically to change them to the formal form.

Besides, proper names are not capitalized, whereas normal text is capitalized. Because of that, the named entity recognizer may contain errors in detecting named entities. Due to this reason, we apply some error-tolerant rules to compare entities in the two input questions. Two questions are considered as different only if one question has at least two entities more than the other question.

The remaining part of this section analyzes different situations that can make the named entity recognizer in Spacy having mistakes and our solution to these situations.

Situation 1. The system cannot detect some named entities.

For example:

Which aircraft was superior - the Douglas DC8 or the Boeing 707?/ Was the Douglas DC8 a superior aircraft to the Boeing 707?

The named entity recognizer detected in the above example are: [('Boeing', 'ORG')] in the first sentence, and [('Douglas', 'PERSON'), ('Boeing', 'ORG')] in the second sentence. In this case, the NER assigns incorrect names for the entities that it detected. Also, 'Douglas DC8' in the second sentence should be an entity instead of 'Douglas'. The word 'Douglas' is not detected as an entity in the first sentence, whereas it is recognized in the second sentence.

To deal with such cases, the text in the unmatched named entity of one question is searched in the other question. If it is found, we consider it as a matching case. Also, we only base on the text recognized as a named entity, not base on the name of entities, since the named entity recognizer in Space can misrecognize these names.

Situation 2. A proper name can have several expressions, sharing at least one word (e.g., Barack Obama vs. Obama). When comparing two entities, we do not compare exact matching but based on whether two entities share at least one similar word or not.

Situation 3. The named entity recognizer detects more entities in a question than it has.

For example:

Who is a better Person for office Hillary of Donald?/ Why is Hillary Clinton a better choice than Donald Trump?

The entities recognized by Spacy in the above question pairs are: [('Person', 'ORG'), ('Hillary of Donald', 'PERSON')], and [('Hillary Clinton', 'PERSON'), ('Donald Trump', 'PERSON')], in which 'Person' is incorrectly detected as ORG.

Because of that, our system accepts the cases that one question contains one entity more than another. If both questions have entities and more than 2/3 entities in the first question also appear in the second question or vice versa, we have rule_label = 1. Otherwise, rule_label=0, which means two questions are different. If two questions do not have any entity, rule_label is also assigned to 1. The value 1 of rule_label does not mean two questions are similar but only indicates that we do not have evidence to say two questions are different.

Situation 4. A proper name has several written ways which do not share any common word (e.g., *Bengaluru/Bangalore*). If the two questions have only these two entities, then rule_label = 0. If both Bert_label=0 and rule_label=0, then two questions are considered as dissimilar.

V. THE SCENARIO TO REASSIGN LABELS IN QUORA

Since questions in the Quora dataset contain spelling and grammar errors, both the Bert model and our rule-based model cannot detect all dissimilar questions. Because of that, we base on both models to improve the accuracy in evaluating the similarity of each question pair.

We train the Bert model with the Quora training dataset (363850 samples) and test with the Quora development dataset (40431 samples). We filter our question pairs that both the Bert model and the rule-based model return dissimilar (Bert_label=0 and rule_label=0), but the label in the Quora dataset is similar (is_duplicate=1) (182 pairs in total) to reassign labels. In cases of both the Bert model and the rule-based model return dissimilar, the final label is 0.

To guarantee the correctness of our proposed method, we removed labels of the question pairs mentioned above and asked humans to reassign their labels using a crowdsourcing service from Amazon Mechanical Turk (Mturk) [9]. Mturk is a website where a Requester can hire crowdworkers (Mturker) to perform a particular task such as classification, tagging, or translation. The workers complete the task independently and receive a rate from the employer.

We mixed these question pairs with 90 question pairs whose all the three labels (is_duplicate, Bert_label, rule_label) are 1. Then we crowdsourced the labeling task to five people working independently by removing labels of the question pairs mentioned above and uploading these question pairs to Mturk. The task of Mturk workers was to label 1 for the semantically equivalent question pairs, and 0 otherwise.

To choose Mturk workers participating in the task, we first released the labeling task with a subset of ten questions to twenty high-rated Mturk workers. After getting their results, we selected five people whose labels were most similar to major labels of all workers. Finally, we assigned the labeling task with the whole question set to these five workers.

The final label of each question pairs is taken from the majority labels among five people. Finally, 150 out of the above-mentioned 182 question pairs are assigned with the label 0, agreeing with the Bert and rule models. That means, the accuracy of assigning label 0 for the cases that both the Bert and rule models return the label 0 is 82.41%. It points out that the automatically assigned labels are reliable enough to be applied.

Next, we update the Quora development dataset with labels that have been reassigned in Mturk and used this dataset to retrain the system. The Bert model after trained and the rule-based model is used to reassign labels in the Quora training dataset (363850 samples).

The training dataset is divided into ten smaller datasets of the same size, named block0, block1, …, block9. Before automatically assigning labels for the training dataset, we re-evaluate the accuracy of our method by automatically reassigning labels for block0 and block1 and comparing them with the labels assigned by Mturkers. If the system's accuracy is more than 80% of cases for both block0 and block1, we will automatically assign labels for the remaining blocks.

We filter out question pairs in block 0 with is_duplicate=1, Bert_label=0, rule_label=0 (467 samples) and ask five people to label them. This task was done by using the same way as we did in the previous step. To increase the diversity of the label, 230 question pairs in block 0 with is_duplicate=1, Bert_label=1 , rule_label=1 are added to the dataset and then upload to Mturk for reassigning labels. After getting the results, we reselect 467 samples mentioned above to evaluate the system's accuracy.

The same experiment was carried out with block1. The results are shown in Table 2 below.

TABLE II. MTURK LABELS FOR QUESTION PAIRS IN BLOCK 0 AND BLOCK 1 WITH IS_DUPLICATE=1, BERT_LABEL=0, RULE_LABEL=0

#question pairs with is_duplicate=1		***block0***	***block1***
		467	385
Mturk_label	1	83	60
	0	384	325
Accuracy		82.23%	84.42%

The results with the two blocks shown that more than 80% of cases humans agreeing with our system. The remaining cases often require world knowledge to understand, as in the following example:

What could be the reason behind Arnab Goswami quitting Times Now?/Why is Arnab Goswami resigned as the Editor-in Chief of Times Now and ET Now?

The accuracy of more than 80% indicates that our method is good enough to automatically reassign labels for the remaining blocks of the Quora train set.

The label of a question pair is defined as:

label = Bert_label if Bert_label == rule_label

or rule_label == 1

else label = is_duplicate

The Quora train and development dataset after being reassigned are uploaded in the link

https://drive.google.com/drive/folders/1j5N2htKoM4LMSAuGid9NZEkBANNvdylD?usp=sharing

We have reassigned 150 labels (from 1 to 0) in the Quora development dataset and 22460 labels in the Quora training dataset.

VI. Conclusion and Future Works

This paper has introduced our method of reassigning labels for the Quora dataset, basing on Bert and our proposed rules. Our method has an accuracy of more than 80% when comparing the labels assigned by the system with the ones assigned by Mturkers. We have updated 150 labels in the Quora development dataset and 22460 labels in the Quora train. This dataset has been published so that it can be used by other researchers on question similarity. Future work includes identifying other cases of incorrectly assigning labels in the Quora dataset and investigating methods to solve these cases.

References

[1] Dasha Bogdanova, Cicero dos Santos, Luciano Barbosa, and Bianca Zadrozny. Detecting semantically equivalent questions in online user forums. Proceedings of the 19th Conference on Computational Natural Language Learning, 1:123–131, 2015.

[2] Corinna Cortes; Vladimir N. Vapnik. 1995. Support-vector networks. Machine Learning. 20 (3): 273–297

[3] 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.

[4] 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)

[5] "Quora Question Pairs | Kaggle". Kaggle.com. N.p., 2017. Web. 23 Apr. 2017.

[6] Damar Adi Prabowo, Guntur Budi Herwanto. 2019. Duplicate Question Detection in Question Answer Website using Convolutional Neural Network. 2019 5th International Conference on Science and Technology (ICST)

[7] I.Srba and M. Bielikova, "Why is Stack Overflow failing? Preserving sustainability in community question answering," IEEE Software, vol. 33, no. 4, pp. 80–89, July 2016.

[8] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144.

[9] Amazon Mechanical Turk. www.mturk.com. Last visited Feb. 16th , 2021