Improving Student Forum Responsiveness: Detecting Duplicate Questions in Educational Forums

Manal Mohania(✉), Liyuan Zhou, and Tom Gedeon

College of Engineering and Computer Science, Australian National University,
Canberra, Australia
{manal.mohania,liyuan.zhou,tom.gedeon}@anu.edu.au

Abstract. Student forums are important for student engagement and learning in university courses but require high staff resources to moderate and answer questions. In introductory courses, the content can remain almost unchanged each year, so the questions asked in the course forums do not see a lot of variety over different iterations, which provides an opportunity for automation. This paper compiles a dataset of forum threads and meta-information of the participants from the Web Design and Development course at the Australian National University for the purposes of duplicate question detection in educational forums. A state of the art neural network model is trained on the dataset to measure its usefulness. An accuracy of 91.8% is achieved, which is on par with what is achieved on other datasets with similar features. A high performing neural network for this dataset could potentially be used to create a live system that detects and reuses answers for duplicate questions on course forums.

Keywords: Duplicate question detection · Neural networks · Duplicate question pair dataset

1 Introduction

The use of online forums as a medium for discussion and communication has become widespread in the field of education. One typical use case is for facilitating student discussions during an offering of a course. These forums are generally very rich in micro-collaborations [1] because all users have the ability to ask, answer and rate content. A study [2] reveals that discussions on these course forums promote collaborative learning by enhancing community building, developing self-identity, and improving relational dynamics, which in turn support learning at various knowledge levels and improve the cognitive process in learning.

However, while solving some problems, these discussion forums face problems of their own. While the forums are becoming more and more accessible by making the bar for participating on these forums quite low, it also inevitably leads to a

T. Gedeon et al. (Eds.): ICONIP 2019, LNCS 11955, pp. 410–421, 2019.
https://doi.org/10.1007/978-3-030-36718-3_35

lowering of the overall quality of the forum. In particular, while asking questions, it has been observed that a significant number of questions asked on a forum have previously been asked before. While no formal study that investigates this was found, this has been observed in some of the major web forums such as StackExchange, Quora and Yahoo! Answers.

With this paper, we release an anonymised dataset of questions and answers asked in a course forum for a Web Development and Design course at the Australian National University over the years 2015–2019. This course uses Piazza, a question and answer web service. We thus, henceforth, refer to this dataset as the "COMP1710 Piazza Dataset" (COMP1710 is the course code of the undergraduate version of the Web Design Development course at the Australian National University). This dataset will also feature metadata about the students in the course, such as their overall grade in the course, the mark they got for their participation on the course forum, their gender and ethnicity etc. Moreover, information about the questions that are duplicates of one another are also stored within the dataset.

After construction of the dataset, we perform some experiments by running a state-of-art-model built for natural language sentence matching on this dataset. We find that it achieves an impressive accuracy of 91.8% despite the average "sentence" length being much higher than what was previously used with that model. Upon experimenting with other duplicate question datasets where the average sentence is comparatively longer than the Quora Question Pairs Dataset (the original dataset on which it was tested), similar high accuracies were achieved. Surprisingly, this fact was not noted in the original paper.

2 Related Work

The task of detecting duplicate questions is a sub-task of the more general paraphrase detection task. However, the approaches used to solve the more general task are not always a step in the right direction towards detecting duplicate questions. In fact, it has been found that the performances achieved by different machine learning models on text paraphrase detection was significantly better than the ones achieved on detecting semantically equivalent questions [3].

2.1 Datasets

Numerous datasets related to the fields of question similarity have been published in recent years.

The Qatar Living Dataset. SemEval (Semantic Evaluation) is an ongoing series of evaluations of computational semantic analysis systems. In 2016, one of the tasks (Task 3) [4] in the SemEval workshops was related to answer selection in community question answering forums, which involves both detecting semantically equivalent questions and also selecting the best answer from a range of answers.

For evaluation of the models, they released the Qatar Living data corpus the source of which is the Qatar Living Forum[1]. This dataset contained 317 original questions, 3169 related questions and 31690 comments

The CQADupStack Dataset. CQADupStack is another benchmark dataset in the field of Community Question Answering. It contains threads from twelve StackExchange subforums, annotated with duplicate question information and comes with pre-defined training, development, and test splits, both for retrieval and classification experiments [5].

The dataset contains over 460,000 threads (an average of 38,362 threads per subforum). The percentage of duplicate questions has a high variation between subforums- ranging from 1.52% for the Wordpress[2] subforum to 9.31% for the English subforum[3]. The average number of duplicate questions per duplication, however, has a much smaller range (1.02 to 1.22).

The duplicate question annotations were manually performed by the users in these subforums. As a result, these labels are not guaranteed to be perfect. In fact, a study [6] concluded that the number of duplicates could be increased by around 45%, by annotating only 0.0003% of all the question pairs in the data set.

The Quora Question Pairs Datasets. In early 2017, Quora[4], a question and answer website, published a dataset of over 400,000 potential duplicate question pairs [7].

Questions on Quora differ from the questions on the Stackexchange and Yahoo! Forums in that they do not possess a separate question body. Questions on Quora are limited to a maximum length of 250 characters. This limitation compels the user to ask more general and less detailed questions. This is in contrast to most educational forums where the asker has the ability to explain their current understanding of the topic through the question bodies.

Despite the existence of a multitude of datasets relating to the field of community question answering, we construct another dataset due to the following reasons:

1. **Narrow Scope of Field:** Other datasets published so far are quite general in nature (with the exception of InsuranceQA). Quora Question Pairs for example, is not limited by scope in the types of questions contained. The Stackoverflow dataset, on the other hand, is somewhat more restricted in scope when compared to the Quora Question Pairs dataset. However, its scope (general programming) is still quite large to make it difficult to perform a detailed analysis. We create a dataset with a more restricted scope, one of questions asked during multiple offerings of a web design course at the

[1] www.qatarliving.com/forum.

[2] https://wordpress.stackexchange.com/.

[3] https://english.stackexchange.com/.

[4] www.quora.com.

Australian National University. The scope of this is small enough such that a significant fraction of duplicate question pairs are found, and large enough so that significantly many new questions can be added to it that are not duplicates of existing questions. This helps with the training of neural network models.
2. **Inclusion of Meta-data:** None of the datasets published so far include meta-data about the backgrounds of the users who participate in the forums. We include meta-data such as the ethnicity, gender, grade obtained in the course etc. This can potentially be used to deduce correlations between the quality of the posts made and the backgrounds of the users. A presence of a strong correlation would suggest that the background of the students could act as an effective heuristic measure when deciding the best answer to a given question.

3 The COMP1710 Piazza Dataset

Piazza is a question and answer platform that is used by many universities across the world. Piazza comes with a wide set of features which makes it an indispensable asset for many courses. For instance, the platform allows users to ask questions, post notes or hold a poll. These can be done anonymously, semi-anonymously or with the name visible to everyone. Piazza also provides a good rendering engine for code and $\LaTeX$ snippets.

The COMP1710 Piazza dataset is an anonymised dataset of questions and answers asked in the Piazza course forum for a Web Development and Design course at the Australian National University over the years 2015–2019. This dataset also features metadata about the students in the course, such as their overall grade in the course, the mark they got for their participation on the course forum, their gender and ethnicity etc. Moreover, information about the questions that are duplicates of one another are also provided within the dataset.

3.1 Dataset Format

The dataset is divided into three different files- one for the content of the threads, one for the metadata and one for information about duplicate questions.

One of the files (questions.json) is a JSON file that maps unique thread ids to information about them. The unique ids for the threads were created by hyphen separating the year of posting and the serial number of the thread in that year. For example, $2018 - 141$ refers to the 141^{st} thread in the year 2018. These ids are mapped to information about the threads such as their title, body, answers, comments, votes, anonymous ids of the users that participated etc. Information on the history of the thread is also supplied along with the timestamps. The names used for their keys are self-explanatory. As mentioned previously, the choice of these keys was influenced by the visual structure of a Piazza thread page, and the information available to users.

The metadata is present in a separate JSON file (metadata.json). This file consists of a mapping from the anonymous students ids to their relevant meta information. References to the anonymised student mappings will also be present in the questions.json file.

Finally, the annotations for the duplicate questions are available in a CSV file of its own (duplicates.csv). Each row in the file contains the thread ids of questions that are duplicates of one another. Only the questions that have at least one duplicate have been mentioned in this file. Otherwise if a particular thread id is omitted from the file, it means that either the thread is not a question, or that it does not have any duplicates.

The dataset will be made available in December via www.hcc-workshop.anu.edu.au/comp1710-piazza-dataset.

3.2 Duplicate Question Definition

The definition for duplicate questions that was initially agreed upon by the dataset annotators was the same as the one that is used frequently in literature [8].

> "Two questions are semantically equivalent if they can be adequately answered by the exact same answer."

However, this definition when used directly with the COMP1710 Piazza Dataset is not very useful. The primary reason for that is that the Piazza forums for the COMP1710 course are monitored for quality less rigorously than other real world forums such as StackExchange. In particular, asking multiple questions as part of the same thread is allowed in the former whereas, the latter follows the principle of "one question per thread"[5]. Keeping in mind that the eventual goal was to create a live question answering system, a few constraints were added to make it applicable to the majority of the questions in the dataset. The additional constraints added are listed below.

- If multiple questions are asked in two different threads, the threads would be considered duplicates if the majority of questions in one are duplicates of the majority of questions in another.
- In the case above, if there is no clear majority on the number of questions, the questions are then weighted by their word counts. As a consequence, if a particular thread consists of two questions where one uses a word count of 100 and the other uses 10 words only, the first question is assumed to be the "majority" of the given thread.
- Some questions when asked in different years get different responses. For example, "What is the location for the final exam?" is likely to receive different responses in each year. Such information retrieval questions where the answer may vary across years have still been annotated as duplicates.

[5] https://stackoverflow.com/help/how-to-ask.

Adding these additional constraints to the definition for duplicate questions made the annotation process more robust to human biases when performing annotations for questions that are in the inevitable grey areas due to ambiguities in the questions being considered.

3.3 Forum Statistics

The COMP1710 Piazza Dataset combines data from various different sources. Performing different statistical analyses may thus, reveal better insights about the dataset. Various statistical analyses are performed for the dataset, and the results obtained are then compared to existing datasets and discussed.

Number of Threads. The COMP1710 Piazza Dataset consists of 4,145 threads (inclusive of questions, notes and polls). When compared to other datasets from the domain, this is only larger than the Qatar Living Dataset. However, while the dataset size may seem orders of magnitude smaller relative to the larger ones, the smaller scope for the topic of discussion compensates for fewer threads, and is at the large end of what can plausibly collected from university course forums – the course has grown from 146 students in 2015 to 264 in 2019, and so has been on the medium to large size throughout.

Out of the 4,145 threads in the dataset, 3,262 of them are questions.

Average Length of Questions. On average, a question body contains 66.2 words (all HTML tags are stripped before this figure is calculated). This is on par with the average question length of the StackExchange datasets where the users have the ability to contextualise/describe their thoughts about the problem they are facing.

On the other hand, this statistic is much larger when compared to the Quora Question Pairs dataset, which has an average question length of under 10 words. The reduced length of questions on Quora allows the question to be more focused in scope. A longer question body, while allows the asker to explain the question with more rigour, also carries extra information that is often irrelevant to the question being asked.

3.4 Statistics for Duplicate Questions

The COMP1710 Piazza dataset has some interesting statistics for the duplicate questions present. We discuss these in this section and also compare the statistics with other datasets where possible. Due to the late arrival of the 2019 data, these statistics have been measured for the 2015–2018 subset.

Percentage of Duplicate Questions. There has been little change in the content and assignments of the COMP1710 course with each iteration. Unsurprisingly, a lot of the questions that are asked in a particular year are very similar to those that were asked in other years. In fact, around 42% of the questions that have been asked over the four years in the Piazza forums for this course have duplicates. When compared against question *pairs*, approximately 0.14% of the question pairs in this dataset are duplicates from all possible pairs of questions.

While the low percentage for the percentage of duplicate pairs is explicable, it does imply that there is a heavy imbalance in the labels of the classes of this dataset. This needs to be a consideration when trying to learn latent features from the dataset.

Number of Duplicate Questions per Duplicate Question. When a particular question has at least one duplicate, there are at least 4 of them on average in the COMP1710 Piazza Dataset. This number is significantly higher than the StackExchange datasets where this statistic has a value between 1 and 2 for all forums [6]. The higher number in this case is likely indicative of the fact that certain questions are very popular which push the average up. As an example, 20 various forms of the question "How do I submit my assignment?" were posted over the four years.

It should be noted that the feature of certain questions being very popular is not restricted to this dataset. In the StackExchange dataset, for example, on the webmasters subforum[6], a certain question appeared in 106 different forms. Due to the sheer size of the dataset, the value of the average number of duplicates is not heavily affected by such outliers.

4 Experiments and Results

4.1 The BiMPM Model

Wang et al. proposed the Bilateral Multi-Perspective Matching (Bi-MPM) model for the task of Natural Language Sentence Matching [9]. The model achieves state of the art performance for sentence paraphrase matching, when tested on the Quora Question Pairs Dataset. This model matches each time stamp of each of the two questions with every time stamp of the other question. A Bi-LSTM layer is then used to produce a fixed length matching vector, which is further as used as input for a fully connected layer that makes the final decision. We use this model to conduct our experiments.

4.2 Experiments Performed

Due to the late arrival of the 2019 data, we only utilised the 2015–2018 subset of the data for our experiments. That subset of the data contains 2,300 questions which corresponds to approximately 2.6 million question pairs (Fig. 1).

[6] https://webmasters.stackexchange.com.

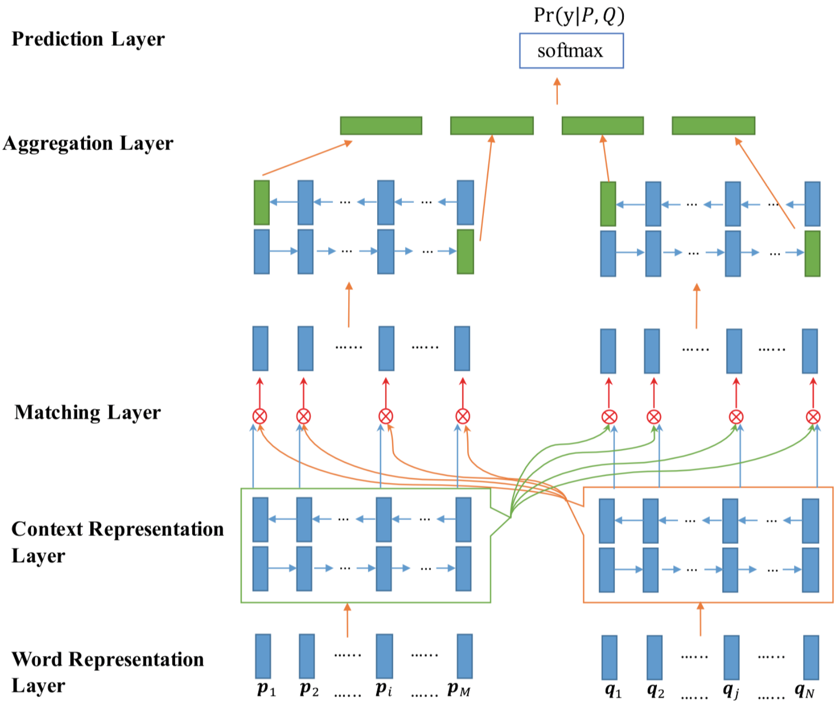
Prediction Layer
Aggregation Layer
Matching Layer
Context Representation Layer
Word Representation Layer
Pr(y|P, Q)
softmax
p1 p2 ...... pi ...... pM
q1 q2 ...... qj ...... qN

Fig. 1. The Bi-MPM model [9]

Preprocessing of the Data. The data was pre-processed in three stages- conversion to lowercase, removal of HTML tags (including images) and the removal of foreign characters. The final step was necessary because even though the course is taught in English, due to the significant number of international students, some question inadvertently contain non-English characters. We also used a list of stop words that contained standard greetings, as they add little value to the semantics of a question. However, it was also observed that removing this stop list did not affect results significantly.

4.3 Samples Generated

The first sample we generated consisted of 7,260 question pairs. This sample was created such that the number of question pairs labelled as duplicates was roughly equal to the number question pairs labelled non-duplicates. We created a 60%-20%-20% split for training, validation and testing respectively.

Another sample that we generated was one containing over 33,000 question pairs. This sample was purposefully created such that there is a heavy imbalance in the number of duplicate and non-duplicate questions. This sample contained around 2,200 duplicate question pairs. The validation and the test files used for this sample were the same as the ones used in the previous sample.

Results. We trained and tested the Bi-MPM model on the samples described above. The experiments were run with a batch size of 60 and trained to a maximum of 20 epoch cycles. The dropout rate used was 0.1 and the learning rate was 0.0005. We made use of the Adam [10] optimiser for the model. The results we achieved are summarised below (averaged over multiple runs) (Table 1).

Table 1. Performance of the Bi-MPM model on balanced and unbalanced samples

Sample type	Validation accuracy	Test accuracy
Unbalanced	77.14%	**76.11%**
Balanced	94.23%	**91.78%**

Discussion. Considering the average performance of various other models on other datasets that we discussed in Sect. 2, the results achieved by the Bi-MPM model on the balanced sample are on the higher end of the spectrum. However, the results are not overly surprising because the model performed remarkably well on the Quora Question Pairs dataset. We also validate its performance on other datasets, which we investigate in Sect. 4.3.

The difference between the performance of the model on the balanced and unbalanced samples is, however, not surprising. The heavy imbalance in the model causes the weights in the model to be trained such that the output is always biased towards marking question pairs as non-duplicates. However, since the test set itself is balanced, the overall accuracy is significantly lower on the test set.

Q1: the mark about assignment2: i have some questions about the mark and the feedback about it. 1. it is said that there is no forum posts nominated. however, apparently my questions are not anonymous. here is the photo about it. 2. it is said that there is no more than 3 links and no labels in the image map. however, there are 5 photos in the image map and each one has a label and a link. 3. it is said that there are no less than 10 photos in the photo gallery. however, i made two photo galleries and each one has five photos. in all, there are 10 photos here is the photo about the feedback. i strongly hope you can check my assignment again and give me a reasonable mark. thank you!
Q2: marking issues: i had included a portfolio page on my website showing some of my music, which was supposed to be my ”something original”, but received no marks for it. also, i accidentally added my new css file to every page rather than just 1, and received no marks. is there any chance of that being taken into consideration for my marks?
Model Output: Non-duplicates
Gold Standard: Duplicates

Fig. 2. Example of a question pair that was incorrectly classified as non-duplicates

A further analysis of questions that were incorrectly classified by the model trained on the balanced sample is performed below.

Figure 2 is an example of a question pair where the intent of both the questions is the same, in that they want to get their assignments remarked. The reasons, however, are very different which is evident from the context. This additional context is likely to be the reason why the model did not consider these two questions to be duplicates. The additional context, however, does not affect the true label of this question pair because all such questions in the dataset had the same answer along the lines of "It is best to bring this up with your tutor directly during your lab".

Q1: how do i make multiple page?
Q2: delete file in partch: hi all, i want to know how to delete files in partch with file name have symbol . or space in it? thanks so much for the answer.
Model Output: Duplicates
Gold Standard: Non-duplicates

Fig. 3. Example of a question pair that was incorrectly classified as duplicates

Finally, Fig. 3 is a question pair where one of the questions is quite short and harder to reason about. While there are not any common keywords between the two questions, they have been likely labelled as duplicates because the training set contained a few questions about "deleting multiple files on Partch" which may have ended up in the model parameters being updated such that the words "multiple", "delete" and "Partch" might be treated as near synonyms of one another resulting in the two questions being classified as duplicates.

With Other Datasets. To confirm that the results obtained by the Bi-MPM model on the COMP1710 Piazza dataset were not the result of an anomaly in the dataset, we ran the Bi-MPM model on a few other datasets, namely, AskUbuntu and Meta StackExchange. They were chosen primarily because the format of the questions in those are more similar to the ones in the COMP1710 Piazza dataset as compared to Quora Question Pairs.

The configuration of the model was the same as in the above experiment. The train, validation and test splits used for these datasets were the same ones used by Rodrigues et al. [11] to discredit the work of Bogdanova et al. [8]. These splits do not contain the clue in the question texts which had originally been left in by Bogdanova et al. The training, validation and testing sets in both the datasets have an almost equitable distribution for the two labels.

Running the Bi-MPM model for a maximum of 20 epochs on the respective datasets produced the results that are summarised in Table 2.

Discussion

The performance achieved by the Bi-MPM model on the other datasets is in a very similar range to what is achieved with the COMP1710 Piazza dataset.

Table 2. Performance of the Bi-MPM model on other duplicate question datasets

Dataset	Test set accuracy
Quora Question Pairs	88.17%[a]
Meta StackExchange	88.95%
AskUbunutu	92.34%
Comp1710 Piazza	91.78%

[a] As reported in the original paper

Surprisingly enough, not all of these results were reported in the original Bi-MPM paper [9]. There are two main points of discussion with these results.

Firstly, the average question length seems to have an inverse effect on the performance of the model. The model performs worst on the dataset with the smallest average question length. This is a bit surprising because a longer question often contains information that is not entirely relevant to the crux of the underlying question.

Secondly, despite major differences between the COMP1710 Piazza dataset and the AskUbuntu and Meta StackExchange datasets, the test accuracy achieved on the datasets is comparable. We believe that there are two conflicting factors at play here. Firstly, the narrow scope of field of the COMP1710 Piazza dataset makes it comparatively easier for the model to learn from the training set as it has a very limited vocabulary and can thus be better analysed. Secondly, the lack of incentive to maintain a high question quality results in the questions in the dataset having a high number of spelling errors and often, bad grammar. Considering these two points, it is not surprising that the accuracy achieved on the COMP1710 Piazza dataset is similar to the StackExchange datasets.

5 Conclusion and Future Work

In the previous section we saw that the Bi-MPM model performs very well on questions where the length is longer. This fact was surprisingly not noted in their original paper. In absolute terms, an accuracy of 91.8% is achieved on our test set. This is a lot higher than the test accuracies on other datasets by other models that we found during our research. However, this figure is on par with the accuracy achieved by the Bi-MPM model on other datasets, including those of StackExchange, where the questions are longer on average.

A substantial perceived benefit of the automated system is responsiveness, in being able to provide an answer essentially instantaneously to the large majority of questions. This work was used to create a pilot automated question answering system that automatically reuses answers from previous years if a new question that is semantically equivalent to a question from a previous year is asked. On limited testing by 11 students (9 males, 2 females with an average age of 22.18 and a standard deviation of 2.52) on a live version of the course, the average vote (on a scale from 1 to 5) on the ability of the bot to reuse answers from previous years was valued at 2.67. One of the reasons that the value was low was likely that the validity and the quality of the old answer was not taken into account when reusing it in another year, which should be possible given the 91.8% accuracy on duplicate question detection.

Finally, the inclusion of the metadata in the dataset makes it a useful dataset outside the fields of artificial intelligence and machine learning. As an example, one could use the metadata to study the correlation (and potential causation) between forum participation and the grade achieved in the course.

References

1. Shachaf, P.: Social reference: toward a unifying theory. Libr. Inform. Sci. Res. **32**(1), 66–76 (2010)
2. Blooma, M.J., Kurian, J.C., Chua, A.Y.K., Goh, D.H.L., Lien, N.H.: Social question answering: analyzing knowledge, cognitive processes and social dimensions of micro-collaborations. Comput. Educ. **69**, 109–120 (2013)
3. Rodrigues, J., Saedi, C., Branco, A., Silva, J.: Semantic equivalence detection: are interrogatives harder than declaratives? In: Proceedings of the 11th International Conference on Language Resources and Evaluation (2018)
4. AlessandroMoschitti, P.L., AbedAlhakimFreihat, W.H., Glass, J., Randeree, B.: Semeval-2016 task 3: community question answering. Proc. SemEval 525—545 (2016)
5. Hoogeveen, D., Verspoor K., Baldwin T.: CQADupStack: a benchmark data set for community question-answering research. In: Proceedings of the 20th Australasian Document Computing Symposium, p. 3. ACM (2015)
6. Hoogeveen, D., Verspoor F., Baldwin T.: CQADupStack: gold or silver. In: Proceedings of the SIGIR 2016 Workshop on Web Question Answering Beyond Factoids, vol. 16 (2016)
7. Quora dataset. https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs. Accessed 7 Oct 2018
8. Bogdanova, D., dos Santos, C., Barbosa, L., Zadrozny, B.: Detecting semantically equivalent questions in online user forums. In: Proceedings of the 19th Conference on Computational Natural Language Learning, pp. 123–131 (2015)
9. Wang, Z., Hamza, W., Florian, R.: Bilateral multi-perspective matching for natural language sentences. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence, pp. 4144–4150 (2017)
10. Kingma, D. P., Ba J.: Adam: a method for stochastic optimization. In: Proceedings of the 3rd International Conference for Learning Representations (2014)
11. Rodrigues, J.A., Saedi, C., Maraev, V., Silva, J., Branco, A.: Ways of asking and replying in duplicate question detection. In: Proceedings of the 6th Joint Conference on Lexical and Computational Semantics, pp. 262–270 (2017)