# Supporting Creation of FAQ Dataset for E-learning Chatbot

Yasunobu Sumikawa, Masaaki Fujiyoshi, Hisashi Hatakeyama, and Masahiro Nagai

Tokyo Metropolitan University, Japan
{sumikawa-yasunobu, fujiyoshi-masaski, hatak, mnagai}@tmu.ac.jp

**Abstract.** Recently, many universities provide e-learning systems for supporting classes. Though the system is an effective and efficient learning environment, it usually lacks a dynamic user support systems. A chatbot is a good choice to support a dynamic Q&A system; however, it is difficult to collect the large number of Q&A data or high-quality datasets required to train the chatbot model to obtain high accuracy. In this paper, we propose a novel framework for supporting dataset creation. This framework provides two recommendation algorithms: creating new questions and aggregating semantically similar answers. We evaluated our framework and confirmed that the framework can improve quality of an FAQ dataset.

**Keywords:** Dataset creation, e-learning, FAQ, chatbot

## 1 Introduction

Thanks to growing IT infrastructures, many universities provide e-learning systems for supporting classes. For example, in the viewpoints of teachers, they can share resumes, assignments, and notifications with students through the system when and where they want to do so. However, unfortunately, many e-learning systems lack a dynamic Q&A system. In other words, it is impossible for users to ask any questions they may have after the system engineers' working time ends. This problem may reduce usability, especially for users who are not familiar with using computers. To provide 24-hour support for the users, a chatbot is a good choice because it automatically answers FAQs any time. Indeed, in industry and government, chatbots are already used to support customers and civilians, respectively, in order to enhance their experiences [1].

As the chatbot is usually implemented by machine learning models, we must prepare the high-quality dataset for training it to obtain highly accurate answers. This requirement has two challenges. First, it is expensive in terms of both the time and cost spent to build the dataset. Second, to collect Q&A data, we must listen to and record the difficulties users had; however, ways to find the people who encounter difficulties in order to ask them questions, either face-to-face or through an email, are few. This

[1] E.g., Facebook bot on Messenger https://developers.facebook.com/videos/f8-2016/introducing-bots-on-messenger/, Yokohama city's bot to support how to trash garbage https://soranews24.com/2017/08/17/yokohama-government-trash-helper-app-gives-poignant-philosophical-advice-to-depressed-citizens/

challenge indicates that it is difficult to collect a large amount of Q&A data to create FAQ datasets and to train chatbots.

**Contributions.** In this paper, we propose a novel framework for supporting chatbot dataset creation specifically for an e-learning system. The core contribution of this study is to provide recommendations that are applicable to small sized datasets. Compared with previous studies on dataset creation, our framework uses two unsupervised learning algorithms: supporting creation of new questions and finding semantically similar answers. We make assumptions as follows:

- It is difficult to automatically create FAQ datasets from small Q&A datasets.
- We can manually create FAQ datasets from small Q&A datasets.
- Supporting manual creation is beneficial to decrease the costs even though we can create the dataset without any tools.

If we have enough data, we can apply supervised learning algorithms that automatically create FAQ datasets as well as [7]. After obtaining many questions from chatbots equipped with small datasets, we can apply the supervised learning algorithms to improve scalability for dataset creation.

The contributions of this study are summarized as follows:

1. To the best of our knowledge, we are the first to create a chatbot to enhance e-learning system used in a Japanese university in practice.
2. We propose a novel framework to create FAQ dataset.
3. We evaluated a chatbot trained on a dataset that is created with the framework and obtained over 81% in terms of macro-average $F_1$-score.

## 2 Related Works

Analyzing Q&A data has been performed by many researchers. Many of the studies seek to improve user experiences [4, 6] or results of classification. As the objective of this study is to support dataset creation for improving the accuracy of a chatbot, which is essentially a multi-class classifier, we focus on comparing the studies trying to improve results of classification with this study.

Finding similar questions to exploit FAQ data is a popular way to improve the accuracy of the classifier for Q&A. One of the most popular approaches is to train language or translation models by probability-based-estimation or neural network [2, 3, 5]. This kind of approach is powerful; however, it assumes that a large amount of data is available to be applied to their models. In contrast, we assume that we can use small Q&A data, and therefore, it is difficult to employ methods for estimating language models. To support creating FAQ dataset from the small size of dataset, we design our framework as an unsupervised learning using a lexical analysis and an entropy-based method.

Supporting dataset creation is another study related to this study. Behúň *et al.* propose an automatic annotation tool for collecting ground truth to a purely visual dataset by Kinect [1]. Rodofile *et al.* design a modular dataset generation framework for cyber-attacks [7]. These studies make assumptions that they can use large datasets or it is easy to create large datasets; the targets are different from our study.

## 3 Data Collection

We first collected raw data from logs of users of the e-learning system introduced in Tokyo Metropolitan University and recorded the questions they asked and answers provided by system engineers who managed the e-learning system in practice. We collected the data from April 1, 2015 to July 31, 2018. The dataset includes 200 Q&A pairs in total.

## 4 Categorization

In this section, we introduce our categorization scheme for the collected raw Q&A based on features of the e-learning system. The objective is to organize answers; this is useful for analyzing the kinds of features users often have difficulties with and understanding the feature we should focus on when preparing FAQ data. From the collected data and manual investigation thereof, we propose 11 categories as shown in Tab. 1.

**Table 1.** Categories for answers.

| Category | Name | Description |
|---|---|---|
| **C1** | **Documents** | Answers related to any questions on documents. For example, ways of showing files to students. |
| **C2** | **Assignments** | Answers focusing on assignments, e.g., settings for an opening duration for students and downloading the results. |
| **C3** | **Test/Questionnaire** | Answers for both test and questionnaire such as re-use of problems in different classes. |
| **C4** | **Contents** | Answers in general for broad contents in e-learning that do not fall under any specific type (e.g., how to keep all data files, assignments, and tests in order to use next year). |
| **C5** | **Uploading** | Answers focusing on processes of uploading any data. For example, answering ”the maximum file size user can upload at a time” question. |
| **C6** | **Registration** | Answers related to processes of registering to classes and such as how a teacher invites another teacher to a class for collaborative team teaching. |
| **C7** | **Aggregation** | Answers for how to combine several classes on the e-learning system. |
| **C8** | **Login** | Answers for any questions related to how to log into the e-learning system (e.g., how to obtain a new password). |
| **C9** | **Contact** | Answers regarding ways about how to communicate between teachers and students such as sending an e-mail to students via the system, using a bulletin board, and so on. |
| **C10** | **Students** | Answers focusing on how students use the e-learning system. In this category, all answers are for only students. |
| **C11** | **Basic Usage** | Answers for how to use the e-learning system. For example, system requirements and operating hours. |

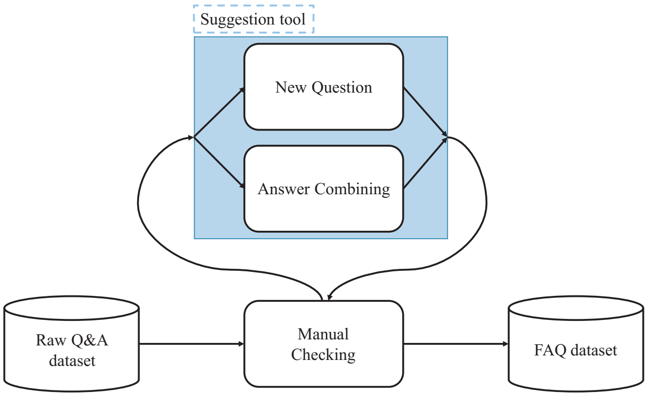

**Fig. 1.** Process overview.

## 5 Dataset Creation using Supports

Fig. 1 shows an overview of processes for creating an FAQ dataset. We assume that transforming the Q&A into FAQ is performed by manual processes. During this process, our framework suggests words to create new questions and combinations of semantically the same answers. The two recommendations play key roles to improve the accuracy of a chatbot, as the first one increases the number of labeled data whereas the second one decreases redundant labels. As our framework is designed for manual creation, users can choose one of the two algorithms when they want to use it. In the remainder of this section, we detail the algorithms of the two suggestions.

### 5.1 Supporting Creation of New Questions

Increasing the number of questions for each answer is one of the most important processes to improve the accuracy of classifications. However, creating new questions is challenging as we must come up with new suitable words that should not be used in other answers to distinguish them. Our framework automatically finds important words that are missed in questions when characterizing their answers. This framework first exploits words from answers if they are not used in current question texts. It then calculates the importance of the exploited words to find words characterizing an answer from other ones. We measure the importance by TF-IDF, which is formally defined as follows:

$$NewWord(a) = \{w \mid w \in W(a) \setminus W(Q_a), TFIDF(w, a) \geq t_{nw}\} \quad (1)$$

$$TFIDF(w, a) = tf_{w,a} * \frac{\mid A \mid}{\mid \{a' \in A \mid w_i \in W(a')\} \mid} \quad (2)$$

where $W(a)$ is a set of words included in answer $a$, $Q_a$ is a set of questions for an answer $a$, $TFIDF$ is a function calculating a score of TF-IDF for a given word $w$, $A$ is a set of answers, and $t_{nw}$ is a threshold used to suggest the words as keywords. In this study, we regard an answer as a document.

**Table 2.** Statistics of dataset created by our approach.

| | |
|---|---|
| Total Num. of answers | 79 |
| Total Num. of questions for training in baseline | 155 |
| Total Num. of questions for training in proposed dataset | 367 |
| Total Num. of questions for test | 44 |
| Ave. len. of questions | 76.9 |

Finally, the framework outputs a list of the top-$k$ important words as a ranking style. The ranking function just sorts results of Eg. 2. The top-ranked words may help us with creating new questions by combining or paraphrasing them.

### 5.2 Combining Answers

If there are more than two answers that are semantically the same as each other, we can combine them to be an answer. We perform mutual information (MI) to find similar answers.

$$MI(a_1, a_2) = \sum_{w_1 \in W(Q_{a_1})} \sum_{w_2 \in W(Q_{a_2})} p(w_1, w_2) \log \left( \frac{p(w_1, w_2)}{p(w_1)p(w_2)} \right) \quad (3)$$

$$Combine(A, t_{ca}) = \{(a_1, a_2) \mid a_1, a_2 \in A, MI(a_1, a_2) \geq t_{ca}\} \quad (4)$$

The Eq. 4 shows pairs of answers whose MI scores are over a given threshold. Showing pairs is enough because we can incrementally use our framework; in other words, even if we can combine more than three answers, we can apply the Eq. 4 to the dataset more than twice. From this simple way, we can combine two or more similar answers as an answer.

## 6 Experimental Evaluation

### 6.1 Setup

**Classification Algorithm.** We used the IBM Watson to implement the chatbot program.

**Data Collection for Evaluation.** Tab. 2 shows the statistics of the dataset used for this evaluation. We used 79 answers and 44 questions to measure the accuracy of the chatbot. Note that the 44 questions were not used to train the chatbot. Tab. 3 details how many answers and questions were prepared for each category.

**Comparisons.** In this paper, we used only the classification algorithm (Watson), as our framework is designed for dataset creation. To evaluate the effectiveness of the framework, we used the following two datasets.

- **Proposed dataset**: This dataset is created with our framework [2].
- **Baseline**: This dataset has the same answers as above dataset; however, this dataset excludes questions created by our framework.

[2] The proposed dataset is available on a public repository server: `https://doi.org/10.5281/zenodo.2557319`.

**Table 3.** Numbers of answers and questions for test.

| | **C1** | **C2** | **C3** | **C4** | **C5** | **C6** |
|---|---|---|---|---|---|---|
| Num. of answers | 2 | 3 | 2 | 2 | 2 | 4 |
| Num. of questions | 4 | 6 | 3 | 4 | 4 | 9 |
| | **C7** | **C8** | **C9** | **C10** | **C11** | *Total* |
| Num. of answers | 2 | 1 | 2 | 1 | 2 | *23* |
| Num. of questions | 4 | 1 | 4 | 1 | 4 | *44* |

**Measurements.** Usually, multi-label classification (MLC) studies use two kinds of measurements: label-based measures and example-based loss functions [8]. As for the label-based measurement, we use macro-average precision, recall, and $F_1$. The macro-average measurements treat all labels equally; in other words, they compute the metrics independently for each label and then take the average. The formal definitions of macro-average precision, recall, and $F_1$ are as follows:

$$P_i = \frac{TP_i}{TP_i + FP_i} \tag{5}$$

$$R_i = \frac{TP_i}{TP_i + FN_i} \tag{6}$$

$$F_1 = \left( \sum_i \frac{2\ P_i\ R_i}{P_i + R_i} \right) / \mid \mathcal{L} \mid \tag{7}$$

where *TP*, *FP*, and *FN* mean true positive, false positive and false negative, respectively, and $\mathcal{L}$ is a set of the label defined in Tab.1. Note here that the precision is defined as the proportion of predicted labels that are truly relevant. The recall is defined as the proportion of truly relevant labels that are included in predictions. The trade-off between precision and recall is formalized by their harmonic mean, called $F_1$-score. In the label-based measurements, the higher these scores are, the better the performances of the model are.

Regarding the example-based loss functions, hamming loss (HL), ranking loss (RL) and log loss (LL) are popular measurements. HL calculates the fraction of the wrong labels to the total number of labels. RL means a proportion of labels' pairs that are not correctly ordered. Finally, LL calculates scores from probabilistic confidence. This metric can be seen as cross-entropy between the distribution of the true labels and the predictions. Their formal definitions are given as follows:

$$HL = \frac{1}{NL} \sum_i^N \sum_l^L [\![ y_{i,l} \neq \hat{y}_{i,l} ]\!] \tag{8}$$

$$RL = \frac{1}{N} \sum_i^N \sum_{y_j > y_k} ( [\![ \hat{y}_i < \hat{y}_j ]\!] + \frac{1}{2} [\![ \hat{y}_i = \hat{y}_j ]\!] ) \tag{9}$$

$$LL = - \sum_i^L y_i \log(p_i) \tag{10}$$

**Table 4.** Scores for both baseline and our approaches. The abbreviated names of measurements are for: macro-average precision (maP), macro-average recall (maR), macro-average F-score (maF), hamming loss (HL), ranking loss (RL), log loss (LL)

| | maP | maR | maF | HL | RL | LL |
|---|---|---|---|---|---|---|
| Baseline | 67.4% | 54.5% | 57.3% | 0.02 | 0.45 | 0.27 |
| Proposed dataset | **93.1%** | **75.0%** | **81.2%** | **0.01** | **0.25** | **0.11** |

**Table 5.** Numbers of answers and questions in training data.

| | | **C1** | **C2** | **C3** | **C4** | **C5** | **C6** |
|---|---|---|---|---|---|---|---|
| Baseline | Num. of answers | 9 | 9 | 15 | 3 | 3 | 11 |
| | Num. of questions | 13 | 18 | 23 | 3 | 12 | 41 |
| Proposed dataset | Num. of answers | 9 | 13 | 17 | 3 | 3 | 11 |
| | Num. of questions | 50 | 68 | 89 | 15 | 19 | 66 |
| | | **C7** | **C8** | **C9** | **C10** | **C11** | *Total* |
| Baseline | Num. of answers | 5 | 1 | 8 | 3 | 3 | *70* |
| | Num. of questions | 12 | 5 | 21 | 3 | 4 | *155* |
| Proposed dataset | Num. of answers | 5 | 2 | 8 | 3 | 5 | *79* |
| | Num. of questions | 25 | 11 | 44 | 15 | 25 | *427* |

In the example-based loss functions, the smaller these scores are, the better the performances of the model are.

### 6.2 Discussion of Results

Tab. 4 compares all measurements of our framework with that of the baselines. The conclusion is that using our framework improves all measurements. Especially, macro-average precision is improved over 25% compared with the baseline. The main reason is that we can increase the number of questions. Looking at Tab. 5, the proposed dataset has twice as many questions as the baseline does.

We then performed error analysis. Figs. 2 and 3 show confusion matrices of our approach. The former one shows what answers the chatbot outputs for test questions whereas the later one shows the result by mapping answers to their categories. In Fig. 2, we use indexes for answers; for example, if we use a question whose answer is the second one, we use `A2` in the figure. The index numbers start in order from 1 to 79, as our dataset has 79 answers as shown in Tab. 5. From Fig. 2, we can see that the chatbot sometimes performs ”mis-answering” for several questions. On the other hand, Fig. 3 shows that the chatbot wrongly predicts only for two categories. For a better understanding of the results, we measured inner- and inter-category similarity by using the Jaccard index. This measurement calculates the similarity by counting the number of unique words shared by given two sets after normalizing their sizes. The formal definition is given as follows:

$$Jaccard(q_1, q_2) = \frac{\mid W_{q_1} \cap W_{q_2} \mid}{\mid W_{q_1} \cup W_{q_2} \mid} \tag{11}$$

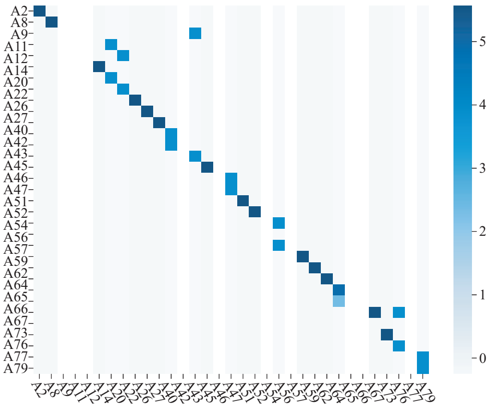

**Fig. 2.** Answer level based confusion matrix of the proposal. The $x$ axis represents correct labels whereas labels predicted by classifier are on the $y$ axis.

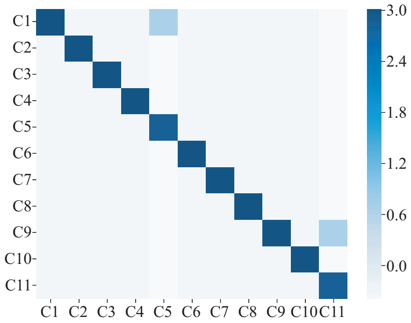

**Fig. 3.** Category level based confusion matrix of the proposed dataset. The $x$ axis represents correct labels whereas labels predicted by classifier are on the $y$ axis.

**Table 6.** Inner-category similarity.

| | **C1** | **C2** | **C3** | **C4** | **C5** | **C6** |
|---|---|---|---|---|---|---|
| Baseline | 0.19 | 0.17 | 0.13 | 0.32 | 0.25 | 0.10 |
| Proposed dataset | 0.11 | 0.09 | 0.07 | 0.06 | 0.22 | 0.08 |
| | **C7** | **C8** | **C9** | **C10** | **C11** | |
| Baseline | 0.20 | 0.27 | 0.12 | 0.30 | 0.23 | |
| Proposed dataset | 0.12 | 0.07 | 0.06 | 0.06 | 0.03 | |

where $W_{q_1}$ indicates a word set of a question $q_1$. Tab. 6 shows scores of the inner-category similarities calculated with the Jaccard index. We can see that relatively high scores occupy this table. In contrast, Fig. 4 shows scores of inter-category similarities scores calculated by the Jaccard index between all combinations of the two different categories. Overall, the scores are lower than that of inner-category similarity. These observations indicate that we should improve the quality of question texts to distinguish in the same category.

In addition, from Fig. 3, we can observe that several questions of **C5** (**Uploading**) and **C11** (**Basic Usage**) are mis-predicted to answers of **C1** (**Documents**) and **C9** (**Contact**), respectively. Mispredictions of **C5** questions as **C1** are understandable as the two categories (**C1** and **C5**) can share file-related words. Indeed, Fig. 4 shows the score of the Jaccard index between the two categories is quite high. Next, to identify why the chatbot wrongly showed an answer of **C9** instead of **C11**, we manually analyzed questions of two categories, **C9** and **C11**. In our dataset, there is a question about *how to make available a function for sending e-mail between teachers and stu-*

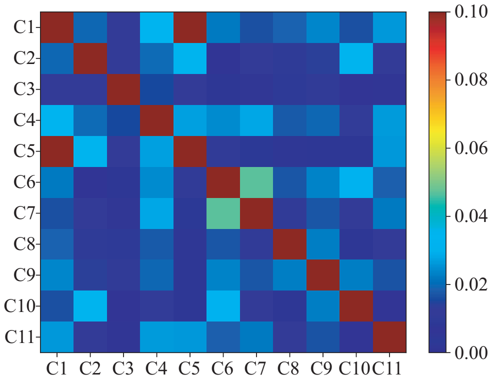

**Fig. 4.** Inter-category similarity of proposed dataset.

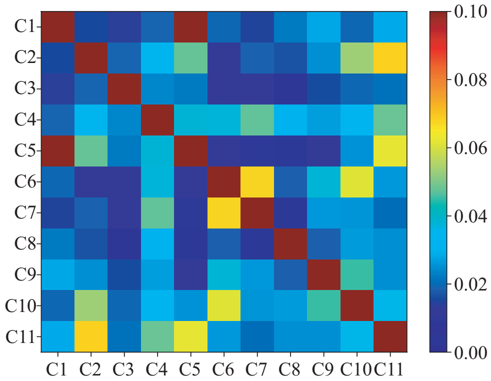

**Fig. 5.** Inter-category similarity of baseline.

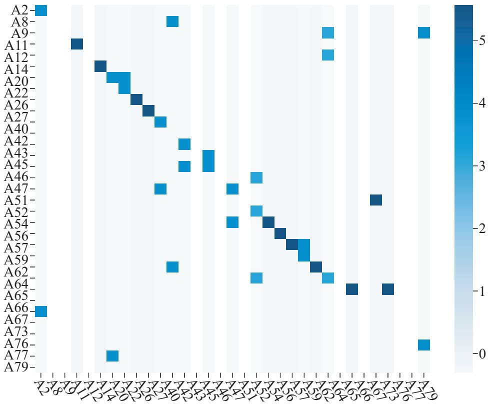

**Fig. 6.** Answer granularity based confusion matrix of baseline. The *x* axis represents correct labels whereas labels predicted by classifier are on the *y* axis.

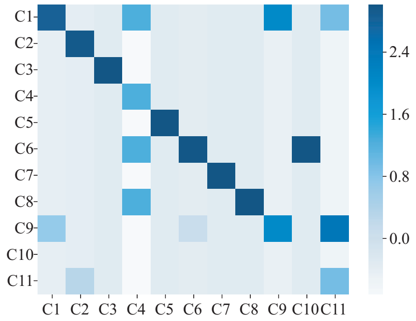

**Fig. 7.** Category granularity based confusion matrix of baseline. The *x* axis represents correct labels whereas labels predicted by classifier are on the *y* axis.

*dents*. This question is similar to **C11** that collect questions related to *how to use the e-learning system*.

Finally, we compared these results of our proposed dataset with that of the baseline. We show inner-category, inter-category, and answer- and category-level confusion matrices of the baseline in Tab. 6, and Figs. 5, 6, and 7. Looking at all similarity scores (Tab. 6 and Fig. 5), they are all higher than that of the proposed dataset. This means that our framework can suggest several kinds of words leading to increasing diversity without decreasing the accuracy of the chatbot because our dataset is better than the baseline.

## 7 Conclusions

In this paper, we introduce a novel framework for supporting chatbot dataset creation specifically for an e-learning system. This framework has two methods: suggesting new words for new questions and aggregating answers that are semantically similar to each other.

In the future, we plan to analyze *a) which questions users tend to have for each month*. In this paper, we assume that all Q&A data can occur independently of time. However, there are some temporal questions regarding registration to classes that may occur early in a semester and questions about tests that users may have late in a semester. This temporal question analysis may improve the effectiveness of chatbots. The future work also includes *b) qualitative evaluation*. This paper focuses on quantitative evaluations; however, analyzing what users feel and think about using chatbots is also important for practical usage.

**Acknowledgments.** We would like to appreciate Mr. Okamura (Tokyo Metropolitan University), and Mr. Kouda, Mr. Suzuki, and Mr. Toya (Alpha Computer, Ltd) for their support collecting and initially organizing Q&A data. This work was supported by JSPS KAKENHI Grant Number 18H01057.

## References

1. Behúň, K., Herout, A., Páldy, A.: Kinect-supported dataset creation for human pose estimation. pp. 55–62. SCCG '14, ACM, New York, NY, USA (2014)
2. Jeon, J., Croft, W.B., Lee, J.H.: Finding similar questions in large question and answer archives. pp. 84–90. CIKM '05, ACM, New York, NY, USA (2005)
3. Leveling, J.: Monolingual and crosslingual sms-based faq retrieval. pp. 3:1–3:6. FIRE '12 & '13, ACM, New York, NY, USA (2007)
4. Morris, M.R., Teevan, J., Panovich, K.: What do people ask their social networks, and why?: A survey study of status message q&a behavior. pp. 1739–1748. CHI '10, ACM, New York, NY, USA (2010)
5. Otsuka, A., Nishida, K., Bessho, K., Asano, H., Tomita, J.: Query expansion with neural question-to-answer translation for faq-based question answering. pp. 1063–1068. WWW '18, Republic and Canton of Geneva, Switzerland (2018)
6. Pinto, G., Torres, W., Castor, F.: A study on the most popular questions about concurrent programming. pp. 39–46. PLATEAU 2015, ACM, New York, NY, USA (2015)
7. Rodofile, N.R., Radke, K., Foo, E.: Framework for scada cyber-attack dataset creation. pp. 69:1–69:10. ACSW '17, ACM, New York, NY, USA (2017)
8. Tsoumakas, G., Katakis, I., Vlahavas, I.: Mining Multi-label Data, pp. 667–685. Springer US, Boston, MA (2010)