An FAQ dataset for E-learning system used on a Japanese University

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

Tokyo Metropolitan University, Japan

A R T I C L E   I N F O

Article history:
Received 1 March 2019
Received in revised form 2 May 2019
Accepted 9 May 2019
Available online 24 May 2019

A B S T R A C T

In this data article, we present an FAQ dataset written in Japanese and its translation to English in order to train chatbot models for e-learning systems. We first collected raw Q&A data reported as the difficulties from April 2015 to July 2018 by users of the e-learning system introduced at Tokyo Metropolitan University. We then divided them into 11 categories according to features provided by the e-learning system. Finally, we integrated questions with the same answers in order to create the FAQ form. The dataset contains 427 questions and 79 answers that were examined by experts with experience in using the e-learning system for more than three years. Using this dataset, we performed statistical analyses to evaluate the qualities of the FAQ dataset. The proposed applications of the dataset include not only academic research but also activities; for example, translating from Japanese to another one like Chinese, adapting/modifying our dataset for another e-learning system, and developing language models to obtain highly accurate responses from chatbots.

© 2019 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

* Corresponding author.
E-mail addresses: sumikawa-yasunobu@tmu.ac.jp (Y. Sumikawa), fujiyoshi-masaki@tmu.ac.jp (M. Fujiyoshi), hatak@tmu.ac.jp (H. Hatakeyama), mnagai@tmu.ac.jp (M. Nagai).

https://doi.org/10.1016/j.dib.2019.104001
2352-3409/© 2019 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
1. Data

The published dataset (See metadata in Table 1) is organized to train chatbot models specifically for an e-learning system. The dataset contains questions asked by students or teachers and their answers in practice. In the dataset, there are Q&A data in Japanese (Answers.csv, Questions.csv, and Categories.csv) and English (Answers_english.csv, Questions_english.csv, and Categories_english.csv), categorization data for answers (Answer2Category.csv) and questions (Answer2Tag.csv).

The dataset also includes the results of statistical analyses, which are described in Section “Statistical Analysis” and Table 4, for comparing our study performed in Ref. [1] with subsequent new research on chatbots or making a new dataset without reducing the quality of the dataset specifically for another e-learning system by checking their scores.

2. Experimental design, materials, and methods

2.1. Data Collection

Raw data were collected by logs that recorded/collected the questions asked by users of our e-learning system and the answers provided by system engineers who managed the e-learning system for more than three years. The collection processes were performed from April 1, 2015 to July 31, 2018. The dataset contains a total of 200 pairs of questions and answers.

2.2. Creating FAQ

After collecting Q&A data, FAQ data were created by combining the raw answers that are semantically similar to each other. If the number of questions for an answer is less than five, then several new questions were added by paraphrasing using suggestion tools proposed in Ref. [1]. When the new questions were added, four experts with experience in using and managing the e-learning system for more than three years examined whether the results were correct or not.
### Table 1
Database files.

| File name               | Content                              | Columns                                      |
|-------------------------|--------------------------------------|----------------------------------------------|
| Answers.csv             | Texts of answers.                    | AID: IDs for answers. Text: Texts of answers in Japanese. |
| Questions.csv           | Texts of questions.                  | Text: Texts of questions in Japanese. AID: IDs of questions corresponding to a question in the same line. |
| Categories.csv          | Titles of categories.                | CID: IDs for categories. Title: Titles of the categories in Japanese. |
| Answer2Category.csv     | Pairs of answers and categories.     | AID: IDs for answers CID: IDs for categories. |
| Answer2Tag.csv          | Titles of answers.                   | AID: IDs for answers CID: IDs for categories. |
| Answers_english.csv     | Texts of answers in English.         | AID: IDs for answers Text: Texts of answers in English. |
| Questions_english.csv   | Texts of questions in English.       | Text: Texts of questions in English. AID: IDs of answers corresponding to a question in the same line. |
| Categories_english.csv  | Titles of categories and their English names. | CID: IDs for categories. Title: Titles of the categories in Japanese. English Title: Titles of the categories in English. |

### Table 2
Statistics of the whole dataset published in this paper.

|                          |                              |
|--------------------------|------------------------------|
| Number of categories     | 11                           |
| Average Number of answers per category | 6.58                       |
| Average Number of questions per answer | 30.58                      |

### Table 3
Numbers of the questions and answers in each category.

| Category | C1 | C2 | C3 | C4 | C5 | C6 | Total |
|----------|----|----|----|----|----|----|-------|
| Questions| 9  | 13 | 17 | 3  | 3  | 11 |       |
| Answers  | 50 | 68 | 89 | 15 | 19 | 66 |       |
| Questions| 5  | 2  | 8  | 3  | 5  | 79 |       |
| Answers  | 25 | 11 | 44 | 15 | 25 | 427|       |

### Table 4
Files including scores of statistics.

| File name                        | Content                                                                 |
|----------------------------------|-------------------------------------------------------------------------|
| basic_stats_chvals.tsv           | Scores of basic statistics. We used Eq. 1 in “ch val” raws.             |
| inner_class.tsv                  | Scores of four measurements listed above from 2 (MI) to 5 (JS). In this file, we ignore all tags. |
| inner_class_inter_tag.tsv        | Scores of the four measurements listed above from 2 (MI) to 5 (JS). This file calculates the scores between different tags for each category. |
| inner_tag.tsv                    | Scores of the four measurements from MI to JS. We calculate them between all combinations of data for each tag. |
| inter_class.tsv                  | Scores of the four measurements from MI to JS. This file calculates them between all combinations of different categories. |
| inter_class_inter_tag.tsv        | Scores of the four measurements from MI to JS. In this file, we create combinations of tags from different categories. |
| inter_tag.tsv                    | This file calculates the four measurements for all combinations of tags including same and different categories. |
From this FAQ data, we trained a chatbot from these Q&A data and then confirmed that the chatbot correctly predicted over 81% in terms of macro-average F1-score [1]. Note that this result is confirmed only from the Japanese version’s dataset. To increase usability and to check the validity of the dataset, we translate the dataset to English. As the English dataset is a translation, the number of Q&A data and the organization of them are the same as the Japanese dataset.

2.3. Categorization

We defined the categories and tags for answers for ease in analyzing their quality. We first defined the categories according to the features of our e-learning system. Then, we defined tags to explicitly write semantic meanings of answers; for example, ways of adding other teachers to the class and ways of removing old teachers. From this process, we finally defined 11 categories of answers: Documents (C1), Assignments (C2), Test/Questionnaire (C3), Contents (C4), Uploading (C5), Registration (C6), Aggregation (C7), Login (C8), Contact (C9), Students (C10), and Basic Usage (C11). Tables 2 and 3 show statistics of the whole published dataset and numbers of questions and answers for each category, respectively.

2.4. Statistical Analysis

For quality control, we performed statistical analyses to evaluate how well the dataset is organized to train chatbot models. Note that the tags for answers can be seen as categories for questions as each answer has several questions. We use the following measurements that are widely used to evaluate the qualities of datasets.

1. Calinski and Harabaz (CH). CH is one of the most popular measures of cluster quality [2]. Intuitively, if all the data in a cluster are close to each other, but the data in different clusters are not close, then we can say that the quality of the clusters is high. The formal definition is given as follows:

\[
CH(k) = \frac{(n - k) B(k)}{(k - 1) W(k)}
\]  

where \(B(k)\) and \(W(k)\) are the sums of squares for the \(k\) clusters of the inner- and inter-clusters, respectively, and \(n\) is the number of clustered data. Thus, the higher the score, the better the quality of the cluster.

2. Mutual Information (MI). MI is a popular information-theoretic measure. This measures the dependence between two sets; in other words, how much information values of a given category generates about the other given category. Thus, the lower the score, the better the quality of our dataset. The formal definition of this measure is as follows:

\[
MI(A, B) = \sum_{b \in B} \sum_{a \in A} \frac{P(a, b)}{P(a)} \log \frac{P(a, b)}{P(a)P(b)}
\]

where \(A\) and \(B\) be the categories, \(P(a)\) and \(P(b)\) be the marginal probabilities, and \(P(a, b)\) be a joint probability. To evaluate the published dataset, we use a variant of MI called Adjusted MI (AMI) [3] defined as follows:

\[
AMI(A, B) = \frac{MI(A, B) - E(MI(A, B))}{\max(H(A), H(B)) - E(MI(A, B))}
\]

where \(E(MI(A, B))\) is the expected MI between the given two categories, \(\max\) is a function returning the largest value of the given ones, and \(H(A)\) is the entropy of the category \(A\).
3. Jaccard Index (Jaccard). This method calculates similarity by the number of common unique data shared by two given sets. To reduce errors caused by too large/small sizes of the given sets, this measure normalizes the results by dividing the size of the intersection of the two given sets by the size of their union. The formal definition is given as follows:

$$Jaccard(S_A, S_B) = \frac{|S_A \cap S_B|}{|S_A \cup S_B|}$$

where $S_A$ and $S_B$ be sets of the given two categories.

4. TF-IDF + Kullback–Leibler (KL) divergence. TF-IDF is one of the most popular methods to measure the importance of a word in a document. KL divergence is an entropy-based method for measuring the similarity between two probability distributions. Their formal definitions are:

$$TF-IDF(w, d) = tf_w, d \frac{|D|}{\left|\{d' \in D \mid w_i \in W(d')\}\right|}$$

$$KL(A, B) = \sum_{a \in A, b \in B} a \log \frac{a}{M}$$

where $tf_{w, d}$ be a term frequency of $w$ in a document $d$ and $W(d)$ be a set of words of $d$. After creating TF-IDF vectors for all elements of $A$ and $B$, we apply Eq. 2 for all combinations of the vectors.

5. TF-IDF + Jensen–Shannon (JS) divergence. JS divergence is an extension of KL divergence to be symmetric form.

$$JS(A, B) = \frac{1}{2} KL(A, M) + \frac{1}{2} KL(M, B)$$

where $M$ be $\frac{1}{2}(A + B)$.

We attached the scores of the five measurements to our database (See metadata in Table 4). We believe that these scores can be useful to compare our study performed in Ref. [1] with subsequent new research on chatbots to investigate differences in detail or to allow transfer or extension of the dataset to another one.

As our e-learning system extends Sakai [4] without removing major functions that are usually implemented in e-learning system or Learning Management System (LMS) like distributing materials used in class, managing assignment, grades, and communication between teachers and students, it is able to use the questions about the functions in the case that a new dataset is created for another e-learning system. However, answers should be changed as there are several unique implementations of the functions and management by human e.g., how to restrict setting new passwords to respect our university's policy, which email address users should send if they cannot resolve the confront difficulties, and so on. Note that the provided statistical results are performed on questions' texts; thus, changing answers' texts does not affect the results.

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.

Transparency document

Transparency document associated with this article can be found in the online version at https://doi.org/10.1016/j.dib.2019.104001.
References

[1] Yasunobu Sumikawa, Masaaki Fujiyoshi, Hisashi Hatakeyama, Masahiro Nagai, “Supporting creation of FAQ dataset for E-learning chatbot”, intelligent decision technologies, smart innovation, IDT’19, smart innovation, systems and technologies, Springer, 2019.

[2] T. Calinski, J. Harabasz, A dendrite method for cluster analysis, Commun. Stat. 3 (1) (1974) 1–27.

[3] N.X. Vinh, J. Epps, J. Bailey, Information theoretic measures for clusterings comparison: Is a correction for chance necessary? ICML’09 (2009) 1073–1080.

[4] Sakai, https://www.sakailms.org.