A Question Answering System in Response to the COVID-19 Crisis

1 Introduction

In March 2020, both federal and provincial governments of Canada took urgent measures to limit the spread of the coronavirus disease (COVID-19) and put in place various programs to deal with the pandemic and its repercussions on people. Governments needed to adapt their infrastructure quickly to assist the population during this crisis. In particular, the telephone helplines had an average wait time of hours at the beginning of the crisis. However, the answers to many questions asked on the helplines were already available on various government websites. The problem was to efficiently find the right answer to a specific question in pages of information. To solve this problem, we developed an automatic question-answering system to assist users. The system asks users to enter a free-form question and returns, whenever possible, the relevant answer from a government website.

2 Model architecture

Our model consists of two main components: a BERT [1] model that ranks all COVID-19 FAQ questions from the Quebec government's website [2] according to their relevance with a user's question, and an out-of-distribution (OOD) detector that eliminates questions that cannot be answered with the available information from the website. This detector operates on the embeddings generated by BERT.

A Mechanical Turk (`https://mturk.com`) data collection was organized to obtain paraphrases for the FAQ questions from the 2020-04-16 scrape of the Quebec government website pertaining to the coronavirus crisis. We fine-tuned the pre-trained $\text{BERT}_{\text{BASE}}$ model to learn to match paraphrases with their corresponding FAQ questions. To do so, we fed each paraphrase and FAQ question separately to BERT and averaged the resulting word piece embeddings to obtain the question representation. Then, we minimized the cross-entropy loss for the i'th (paraphrase, matching FAQ question) pair in the training set:

$$\mathcal{L}_i = -\sum_{k=1}^{K} c_{i,k} \log \frac{\exp(Y^\top x_i)_k}{\sum_{j=1}^{K} \exp(Y^\top x_i)_j},$$

where $x_i \in \mathbb{R}^L$ is the embedding for the i'th paraphrase question, $Y \in \mathbb{R}^{L \times K}$ is the matrix of embeddings of the candidate FAQ questions, L is the dimensionality of the embeddings, K is the number of candidate FAQ questions, and $c_{i,k} = 1$ if the k-th candidate FAQ question is the ground truth question matching the i'th paraphrase question, and 0 otherwise.

We used the Local Outlier Factor (LOF) algorithm [3] for OOD detection, with the default parameter settings in scikit-learn [4]. This algorithm is trained on in-distribution examples. Given that the COVID-19 content of the Quebec government website changes on a regular basis, the OOD detector is updated each time a new scrape is organized. The FAQ questions from the latest scrape are used as the in-distribution examples. We also use the paraphrase questions collected for the 2020-04-16 scrape that match FAQ questions available in the latest scrape as additional in-distribution examples. A general schematic of the proposed solution is presented in Figure 1.

3 Summary of results

Two Mechanical Turk data collections for the English COVID-19 FAQ section of the Quebec government website were organized. The first data collection provides 20 paraphrases for each of the 313 FAQ questions from the 2020-04-16 scrape (15 for training and 5 for validation) to fine-tune $\text{BERT}_{\text{BASE}}$. The second data collection provides $\sim$4 paraphrases for each of the 586 FAQ questions from the 2020-05-22 scrape for our test set. The test set also contains 2500 OOD examples from the HealthTap dataset [5]. Finally, we translated the English FAQ questions, their corresponding paraphrases, and the HealthTap OOD examples to French with the DeepL translator [6]. The translations' quality was verified by manual inspection of a sample. We fine-tuned the pre-trained CamemBERT model [7] using the same procedure as for English. The results are given in Table 1, and example paraphrases are shown in Figure 2.

As we see in the table, the model's accuracy is higher when there is an overlap in corresponding FAQ questions between the training and test data. As new scrapes are organized and new FAQ questions become available, the performance of the model can be improved by periodically organizing new data collections or by augmenting the training set with user questions whose answers led to positive user feedback in the deployed system. Our question answering system was recently deployed in a platform that solicits user feedback.

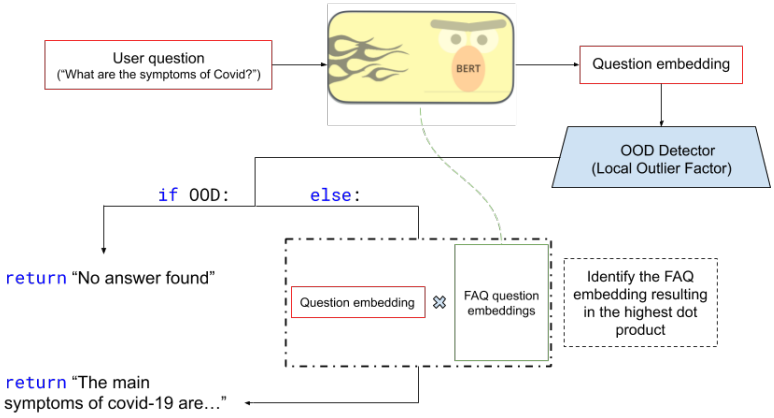

Figure 1: General workflow of the model

```
FAQ entry: I no longer have any income. What must I do?
Paraphrase 1: How much is CERB providing for workers in Canada due to the pandemic?
Paraphrase 2: Are there any resources available to me if I have lost my job?
Paraphrase 3: I have been laid off from my job, are there any programs to help?
Paraphrase 4: What can I do if i have no more money left from a job?
```

Figure 2: Example showing the Mechanical Turk paraphrases collected for a particular FAQ entry.

Table 1: Accuracy of the question answering system on the 2020-05-22 test set. This set includes 2500 HealthTap OOD examples (3rd column), 876 paraphrases corresponding to the 219 FAQ questions which also exist in the 2020-04-16 scrape of the Quebec government website (4th column), and 1467 paraphrases corresponding to the 367 new FAQ questions which don't exist in the 2020-04-16 scrape (5th column). Pre-trained BERT was fine tuned on the 2020-04-16 data. The LOF algorithm's in-distribution examples correspond to all the FAQ questions from the 2020-05-22 scrape as well as all paraphrases from the 2020-04-16 scrape that correspond to FAQ questions that appear in both scrapes.

Language	**Overall accuracy**	**OOD accuracy**	**ID accuracy corresponding to existing FAQ questions**	**ID accuracy corresponding to new FAQ questions**
English	81.83%	90.48%	82.42%	66.73%
French	71.32%	70.60%	80.65%	66.55%

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[2] Quebec.ca covid faq. `https://www.quebec.ca/en/health/health-issues/a-z/2019-coronavirus/answers-questions-coronavirus-covid19/`. Accessed: 2020-07-01.

[3] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. Lof: Identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, SIGMOD '00, page 93–104, New York, NY, USA, 2000. Association for Computing Machinery.

[4] Scikit-learn lof manual. `https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.LocalOutlierFactor.html`. Accessed: 2020-06-30.

[5] Healthtap dataset. `https://github.com/durakkerem/Medical-Question-Answer-Datasets`. Accessed: 2020-06-30.

[6] Deepl translator. `https://www.deepl.com/translator`.

[7] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. Camembert: a tasty french language model, 2019.