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

Mirko Bronzi[1*], Joumana Ghosn[1*], Jeremy Pinto[1*], Cem Subakan[1*], Xing Han Lu[2], Siva Redy[1,2,3], Prakhar Sharma[4]
1. Mila, Quebec Artificial Intelligence Institute 2. McGill University 3. Facebook CIFAR AI Chair 4. IIT Kharagpur, *Equal Contribution-Alphabetical ordering

Introduction

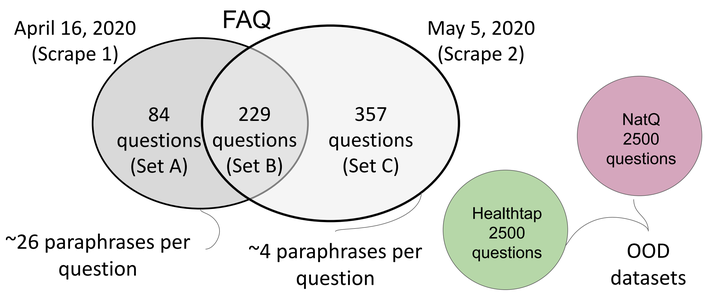
- We developed a COVID-19 question-answering (Q&A) system to alleviate pressure on government helplines.
- Users enter free-form questions and obtain the relevant up-to-date information from vetted websites. If the information is unavailable, the system tells the user that the question is out-of-distribution (OOD).
- **Challenge**: Unlike standard Q&A systems, this system is meant to deal with regular data updates from websites (i.e., data is not I.I.D.).

Data Collection

- We developed a scraper that parses the information available on the COVID-19 FAQ sections of quebec.ca. The scraper runs every 4 hours in production to ensure the Q&A system remains up-to-date with the latest information.
- A Mechanical Turk data collection was organized to obtain paraphrases for the questions from the FAQ sections to mimic user questions.

FAQ question: I no longer have any income. What must I do?
Paraphrase 1: How much is CERB providing for workers in Canada?
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?

- Natural Questions (NatQ) and Healthtap Q&A subsets were used to measure OOD performance. We did not have access to OOD COVID-related questions.

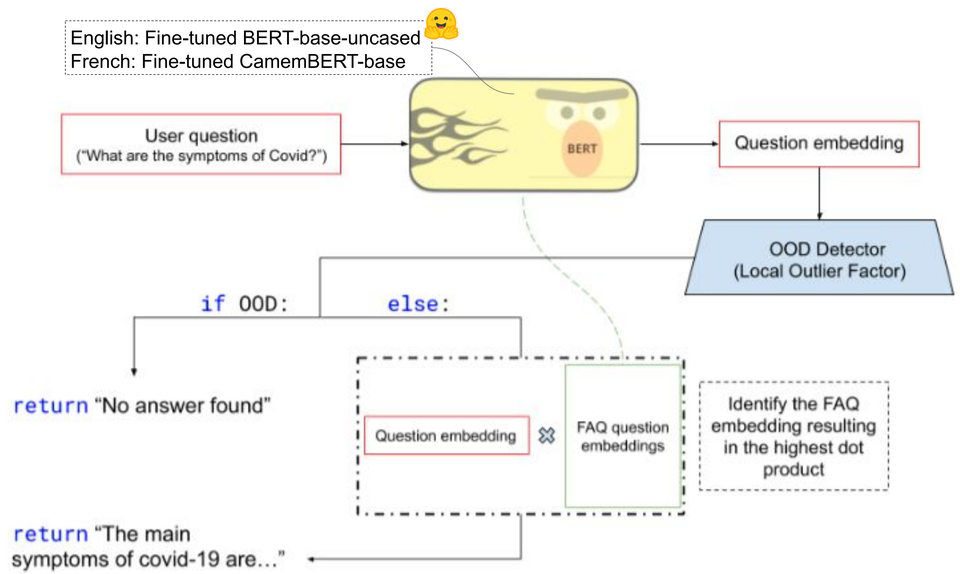


Model Architecture

Our model consists of **two** main components:

1. A BERT model that ranks all COVID-19 questions from the FAQ of quebec.ca according to their relevance with a user question. Each question is represented by its averaged word-piece embeddings at the output of BERT.
2. An OOD detector that determines which questions can be answered with the available information from the website. This detector operates on the BERT question representations and implements the Local Outlier Factor (LOF) algorithm.

General Schematic of the Proposed Solution



Training Details

- BERT is fine-tuned on paraphrased questions from *Scrape 1 (Set A+B)* using a cross-entropy loss to maximize the dot-product between each paraphrase and its corresponding question from the FAQ.
- The OOD detector is fit on the questions from the FAQ of *Scrape 2 and the paraphrases from scrape 1 (Set B)* only. We do not use any OOD samples to train the OOD detector.

Results

- The system is evaluated on the paraphrased questions from *Scrape 2 (Set B+C)*. All paraphrases from Scrape 2 (Set B+C) and all questions from the FAQ of Scrape 2 (Set C) are unseen during the BERT fine-tuning. This experimental setup mimics conditions in production. English performs better compared to french. French data was directly translated from the english data.

	English			French		
Data	**LOF acc.**	**BERT acc.**	**Overall acc.**	**LOF acc.**	**BERT acc.**	**Overall acc.**
Scrape 2 (Set B)	91.16%	89.10%	81.22%	91.41%	87.99%	80.43%
Scrape 2 (Set C)	84.30%	78.22%	65.94%	89.39%	74.69%	66.76%
Healthtap	91.44%	N/A	N/A	74.52%	N/A	N/A
NatQ	95.68%	N/A	N/A	63.08%	N/A	N/A
Entire dataset	91.46%	82.68%	**86.65%**	75.62%	79.98%	**69.86%**

- Fine-tuning BERT increases significantly the performance of question retrieval but does not help OOD detection.

	BERT without Fine-tuning			BERT Fine-tuned		
	LOF acc.	**BERT acc.**	**Overall acc.**	**LOF acc.**	**BERT acc.**	**Overall acc.**
Entire dataset (En)	90.66%	45.83%	75.36%	91.46%	82.68%	86.65%
Entire dataset (Fr)	75.32%	1.29%	47.13%	75.62%	79.98%	69.86%

- Fitting the LOF classifier only on FAQ results in a less strict classifier (less OODs).

LOF acc.	Scrape 2 (Set B)	Scrape 2 (Set C)	Healthtap	NatQ
English	94.00%	94.60%	81.32%	95.36%
French	92.42%	90.72%	18.12%	49.40%

Future Work

- The system is currently live in a production environment. We hope to collect user questions as well as positive and negative feedback to improve the model.
- The system can also be extended to cover multiple official COVID-19 sources.

Acknowledgements