Domain- and Task-Adaptation for VaccinChatNL, a Dutch COVID-19 FAQ Answering Corpus and Classification Model

Jeska Buhmann[1] [0000−0001−6659−6247], Maxime De Bruyn[1] [0000−0002−2801−2391], Ehsan Lotfi[1] [0000−0002−1820−8008], and Walter Daelemans[1] [0000−0002−9832−7890]

CLiPS Research Center, University of Antwerp
Stadscampus L, Lange Winkelstraat 40-42, 2000 Antwerp, Belgium
`{jeska.buhmann, maxime.debruyn, ehsan.lotfi, walter.daelemans}@uantwerpen.be`

Keywords: Dutch FAQ corpus · Dutch paraphrases · FAQ chatbot · domain adaptation · task adaptation

1 Introduction

FAQs are important resources to find information. However, especially if a FAQ concerns many question-answer pairs, it can be a difficult and time-consuming job to find the answer you are looking for. A FAQ chatbot can ease this process by automatically retrieving the relevant answer to a user's question. We present VaccinChatNL, a Dutch FAQ corpus on the topic of COVID-19 vaccination[1]. This is the first publicly available Dutch FAQ answering data set of this size with large groups of semantically equivalent human-paraphrased questions for 181 different intents.

2 FAQ Corpus and Chatbot Development

We used the RASA open source platform for conversational AI[2] to develop our VaccinChat chatbot[3]. The default pipeline was used, including the RASA DIET classifier. The first version of the chatbot was trained with fifty QA-pairs from the official Flemish governmental website (status at the beginning of 2021). After making the trained model public, people were able to use the chatbot by typing in their questions in the online user interface. At the same time, three annotators checked and corrected the predicted intents for the incoming user questions. By iteratively collecting data, accepting or adjusting model predictions, and retraining the model on the improved and increased dataset, we were able to

[1] The paper with the same title was published in Proceedings of the 29th International Conference on Computational Linguistics, pages 3539–3549, Gyeongju, Republic of Korea. (https://aclanthology.org/2022.coling-1.312)

[2] https://rasa.com/

[3] https://vaccinchat.be/

grow the corpus, while simultaneously improving the performance of the chatbot. Afterwards, the complete corpus was checked for privacy sensitive information, resulting in a cleaned corpus of 12,883 user queries and 181 answers.

3 Classification Experiments

We studied the effect of training set size on the classification performance in a 10-fold cross-validation setup with the default RASA pipeline. Results showed a poor performance with a one question to one answer mapping (QA-pairs as training data), and performance only started to converge from 100 user queries per intent, i.e., if the intent had that many train items available.

In addition, we studied whether we could improve classification performance by using a language specific model for Dutch (e.g., BERTje) and applying task- and/or domain-adaptation. All the Dutch language models performed better than the RASA baseline. The task adaptation involved further pre-training with a masked-language-learning approach with the VaccinChat data itself, resulting in improved performance for in-domain user questions that contain domain-specific terminology. The domain adaptation was done by further pre-training a Dutch language model with COVID-19 related tweets, resulting in an improved performance for the slightly of-topic user questions in the non-FAQ intents in our dataset (fallback and chitchat intents).

4 Conclusion

This data set is a main contribution since (i) it concerns Flemish Dutch, a low-resource language, (ii) it contains a large amount (12,883) of question-answer pairs, (iii) it represents a one-to-many mapping of multiple questions labeled with the same answer, and (iv) the questions are real user data containing e.g., typical writing errors, checked for privacy issues.

Another main contribution of this paper is that we show the importance of needing large groups of semantically similar questions, rather than a few examples per intent, to obtain a well-performing intent classification model.

Finally, we show the importance of domain- and task-adaptation before fine-tuning a classifier. We conclude that further pre-training with domain-specific data (in our case COVID-19 related tweets) mainly benefits out-of-domain user questions, whereas extra pre-training with task-specific data mainly improves the in-domain COVID-19 related user questions about vaccination in Flanders, that contain domain-specific terminology.

Acknowledgements This research received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme.