Proceedings of the DataNinja sAIOnARA 2024 Conference 16-18 DOI: 10.11576/dataninja-1159

Question Answering from Healthcare Fora

David M. Schmidt DAVID.SCHMIDT@UNI-BIELEFELD.DE
Bielefeld University, Germany

Philipp Cimiano CIMIANO@CIT-EC.UNI-BIELEFELD.DE
Bielefeld University, Germany

Abstract

Assessing the quality of life of cancer patients is an important aspect of patient-focused drug development and real-world evidence generation. Specialized quality of life questionnaires exist for this purpose, and different types of cancer, such as breast cancer or lung cancer, can be assessed. However, conducting these surveys is a time-consuming process for both patients and clinical staff. At the same time, many patients discuss their experiences with and symptoms of their specific diseases in online healthcare fora. These forum posts may contain information that could be used to answer quality of life questions. Our objective is to determine whether forum posts can be used to answer quality of life questionnaires and, if so, whether this process can be automated successfully.

Keywords: Question Answering, Quality of Life, Healthcare Fora

1. Introduction

In cancer treatment, survival time is not the only important factor for evaluating the success of different therapies. The quality of life of a cancer patient is also an important part of patient-focused drug development in order to ensure patients do not only live as long as possible but also as well as possible considering their respective disease. For evaluating the various aspects that affect the quality of life of cancer patients, specialized questionnaires such as those provided by the *European Organisation for Research and Treatment of Cancer (EORTC)*[1] exist.

However, those filling those questionnaires and evaluating the results is tedious and time-consuming for both patients and clinical experts, although they are of key importance for both patient-focused drug development as well as real-world evidence generation.

At the same time, many cancer patients also frequently use social media and in particular specialized healthcare fora like *Inspire*[2], *Breast Cancer Now*[3] or *Macmillan Cancer Support*[4]. In these fora, patients and relatives of patients discuss how their disease affects their lives, which symptoms they are experiencing, or they support each other getting through those tough times. Many of those topics include aspects of their lives which are relevant to quality of life questions.

2. Research Questions

Considering the mayor investment of time and resources which is necessary to evaluate the quality of life of patients using standard tools like EORTC questionnaires, this observation raises the question whether at least an approximation of the results could be (automatically) extracted from those forum posts. This leads us to the following research questions:

1. *Feasibility:* Is it possible to extract answers to quality of life questions from those forum posts?
2. *Automatization:* Can we extract answers to the questionnaires automatically using AI methods? With which level of accuracy?

3. Research Plan

In order to answer those research questions, first some ground-truth data needs to be collected, consisting of a number of forum posts on the one hand as well as filled *EORTC QLQ-C30* and *EORTC QLQ-BR23* questionnaires on the other hand. We plan to conduct this data by posting a survey in one such healthcare forum. More precisely, we currently cooperate with

1. See https://www.eortc.org/
2. See https://www.inspire.com/
3. See http://breastcancernow.org/
4. See https://www.macmillan.org.uk/

Inspire to realize that data collection. Taking into account the sensitivity of that (partially medical) data, multiple steps were necessary in advance to ensure the compliance of the planned survey with ethical and data privacy standards as well as the consent of the respective patients. These steps include:

1. Plan survey structure and quality of life questionnaires to use (EORTC QLQ-C30 and EORTC QLQ-BR23) ✓
2. Ensure compliance with *General Data Protection Regulation (GDPR)* ✓
3. Get approval by the Ethics Review Board of Bielefeld University ✓
4. Contact healthcare fora about cooperation in posting and conducting the survey ✓
5. Conducting the study (ongoing)

We plan to conduct data from 100 Inspire community members as an initial dataset. In order to answer the first research question, we already started developing annotation guidelines for the data. After the data collection has been completed, the whole dataset will be annotated by three different people, followed by an evaluation whether the data in the forum posts is rich enough to approximate answers to quality of life questions from them.

If this is the case, the annotated data will be used to develop different approaches for extracting the quality of life information from those posts. As the title of this work already suggests, we plan to frame this task as a question answering problem, pointing to different parts of the input context, i.e. the forum posts, which are relevant to the asked question, i.e. the respective quality of life question.

4. Methods

In parallel to preparing the data collection, we already worked on multiple lines of research which are relevant to the second research question. The current state and relevance to the addressed problem are briefly described in the following.

4.1. Baseline Method - Information Extraction from Clinical Trials

The extraction of *PICO (Patient, Intervention, Comparison, Outcomes)* [1, 2] information from *Randomized Controlled Trials (RCTs)* is an important part of creating systematic reviews, which form the foundation of the evidence-based medicine paradigm. Our work [3–5] aims to automatically extract PICO elements in form of nested templates from RCT abstracts. These information extraction approaches could be especially well-suited to act as a baseline for the extraction of quality of life information from forum posts due to the shared medical domain and similar structure of the task.

4.2. Question Answering using *Dependency-based Underspecified Discourse REpresentation Structures (DUDES)*

DUDES [6, 7] are a compositional approach to representing meaning of words, phrases and sentences which can be used to generate SPARQL queries for textual questions given as an input. This DUDES-based approach differs a lot from recent purely machine learning or LLM-based approaches as it is an white box approach by construction. By relying both on fixed composition rules as well as neural dependency parsers etc., DUDES offer an explainable alternative combining the best of both worlds in contrast to the various black box machine learning approaches presented in recent times. This way, it also might be useful for solving the presented question answering task. This work will soon be published in a paper about the DUDES question answering approach.

5. Conclusion

All in all, automated prediction of the quality of life of cancer patients bears the potential to automate parts of the time-consuming quality of life evaluation process and this way potentially allows both clinical experts and cancer patients to focus more on the success of the corresponding therapy.

The necessary data is currently being collected and will soon allow interesting insights into the potential use of social media data for clinical purposes and automation of tedious processes.

References

[1] Connie Schardt, Martha B Adams, Thomas Owens, Sheri Keitz, and Paul Fontelo. Utilization of the pico framework to improve searching pubmed for clinical questions. *BMC medical informatics and decision making*, 7:1–6, 2007.

[2] W Scott Richardson, Mark C Wilson, Jim Nishikawa, and Robert S Hayward. The well-built clinical question: a key to evidence-based decisions. *ACP journal club*, 123(3):A12–A13, 1995.

[3] Christian Witte, David M Schmidt, and Philipp Cimiano. Comparing generative and extractive approaches to information extraction from abstracts describing randomized clinical trials. *Journal of Biomedical Semantics*, 15, 2024.

[4] David M Schmidt and Philipp Cimiano. Grammar-constrained decoding for structured information extraction with fine-tuned generative models applied to clinical trial abstracts. *Frontiers in Artificial Intelligence*, 2024. (in review).

[5] Christian Witte and Philipp Cimiano. Intra-template entity compatibility based slot-filling for clinical trial information extraction. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 178–192, 2022.

[6] Philipp Cimiano. Flexible semantic composition with dudes (short paper). In *Proceedings of the Eight International Conference on Computational Semantics*, pages 272–276, 2009.

[7] Philipp Cimiano, Christina Unger, and John McCrae. *Ontology-based interpretation of natural language*. Springer Nature, 2022.