Consumer Health Information and Question Answering: Helping Consumers find Answers to their Health-Related Information Needs

Anuja Tayal
657111812
UIC,Chicago
atayal4@uic.edu

REVIEW

On searching the internet for any minor or major health issue, instead of getting any concrete answer, consumers stumble upon so many different answers and make them more confused and concerned towards the issue. To resolve this, the authors (Dina Demner-Fushman, Yassine Mrabet, Asma Ben Abacha,NLM) in [1] wants to understand these questions, build dataset to find reliable answers, build and evaluate the consumer health question answering system.

Introduction

The authors develop CHiQA-Consumer Health Information and Question Answering System[1] which has a responsive web interface and a backend module consisting of spelling correction preprocessing module, a question understanding module, 2 Answer Retrieval Module where 1 of them is Traditional IR module and other Question Entailment Module and a final answer generation module.

Question Understanding Module

The Authors in their previous works have proved that knowing the focus and type of question is sufficient to find upto 65% of answers consumers face on daily base. Therefore, they extracted the focus and type of question using 3 modules consisting of Knowledge Based, Traditional Machine Learning Approach and Deep Learning Approach of Question Frame Extraction Approach. To understand question type, the authors hypothesise that .* is sign of .* is a statement about symptoms. In Machine Learning approach, they curated 1400 consumer health questions by taking help of the GARD Dataset(Genetics and Rare Disease Information Center) and trained a multi-class classification system which returned a single candidate. Similar approach was taken to understand the question focus whereby the authors used MetaMap Lite to extract medical entity as their knowledge based approach, trained a Multi-Class SVM using GARD dataset to return a candidate focus for their machine learning approach.

For their deep learning based question frame extraction approach, the authors used Named Entity Recognition annotated with BIOES token format. They developed Bi-LSTM model to build character-level embeddings, which were combined with word embeddings using another BiLSTM layer. The authors considered different word embeddings namely- glove embeddings, binary vector encoding, different vector weights (TF-idf, raw word frequency) and even exploited UMLS. Finally a CRF layer was used to generate tokens. The model was trained for 200 iterations with different embeddings.

Answer Retrieval Module

To retrieve answers, they used two modules-IR based and Entailment based method. For information retrieval based methods, the authors indexed the MedQuAD collection which consisted of Consumer oriented web pages using Apache Solr. Solr BM25 similarity based ranking was used to retrieve answers where question focus and types words were given large weights as compared to other words.

The authors also used entailment to retrieve the answers. If every answer to B is also a complete or partial answer to A, then authors speculate that A entails B. A feature based classifier using logistic regression was trained by considering a total of 9 features where 5 were similarity measures between the questions (word overlap, dice coefficient, cosine distance, levenshtein distance, Jaccard similarity) along with maximum and average values, question length ratio and number of common noun and verbs between the questions. MedQuAD collection of 47,000 question answer pairs were used to train the model and to reduce the search space, Terrier Search Engine was used to retrieve top 100 questions.

Answer Generation Module

The authors combined Information Retrieval and Entailment answers using team draft interleaving methos whereby top 5 answers were shown to the users along with the related questions.

Results

The overall system was evaluated using MEDI-QA collection which consisted of LiveQA-Med 2017 (covering consumer health questions received by NLM) and Alexa MedlinePlus Collection (consisting of 104 short simple template based questions). The answers were classified as Correct and Complete

ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

Answers, Correct but incomplete, Incorrect but Related and Incorrect. The authors considered answers classified as "Correct and Complete Answers" or "Correct but Incomplete" as correct answers. The authors used Mean Average Precision(MAP) which is defined as the mean of the average precision scores for each question and Mean Reciprocal Rank(MRR) which is defined as the average of the reciprocal ranks of results for each question.

Pros and Cons

One of the important things I liked was the motive behind solving the problem. I too have faced this issue and if this problem is resolved it will impact millions of consumers. Also, I liked that while giving the answer, they are also providing with related questions which could be very helpful. Lastly, while reading the paper, I loved the discussion session which was very detailed. On the other hand, the CHiQA system has so many modules and failure points. Even when using word embeddings, they have tried so many different word embeddings. The paper was not self explainable as they had done some prior work but they could have explained their modules a bit. Lastly, their diagrams were too large with labels not self explainable. Also in one of the diagram, equations were written which were neither used in the paper nor explained which could be improved.

Conclusion

To conclude, the authors used different approaches to resolve very common problem faced by the consumers regarding their health.

REFERENCES

[1] Dina Demner-Fushman, Yassine Mrabet, and Asma Ben Abacha. 2020. Consumer health information and question answering: helping consumers find answers to their health-related information needs. *J. Am. Med. Inform. Assoc.* 27, 2 (Feb. 2020), 194–201.