Improved Keyword Extraction using Semantics for Question Retrieval in Healthcare
J.Praveen Chandar¹, S.Sai Meera², A.Sindhuja Komalam³, S.Subalakshmi Santhoshi⁴
Velammal Institute of Technology, Panchetti, India

Abstract:
In the healthcare systems, the criticality of health related information requires careful consideration of how to present the information for sociability. Our main aim is to reduce the gap that prevails between layman and experts in the healthcare domain. This is done by capturing the meaning of key phrases of given questions and retrieving the relevant descriptions by jointly utilizing local mining and global learning approaches. This could be useful to support multiple languages. There needs to be separate server implementations for local mining and global learning. The key concept from the question is identified by using NLP POS tagging technique. If an answer could not be obtained after local mining process, we use resources in global learning. If uncertainty still prevails, an expert answer is to be provided.

Keywords: Healthcare, Mining, Natural language processing, Question retrieval

I. INTRODUCTION
Nowadays, in the era of internet, around eight out of ten users have, at some time searched for health info on the web, because they feel their healthcare needs are not being adequately addressed by increasingly busy physicians. Despite limited research on health benefits, there are many situations where online health communities appear to aid patients. These benefits include improved quality of life, better decision making, making patients feel less alone and more empowered. Indeed, the availability of online health communities is very much useful for individuals with impaired mobility, embarrassing medical conditions, and other conditions that prohibit them from receiving adequate face-to-face medical and emotional support. In this paper we propose a system where a health seeker can post questions and get answers. The keywords from the question are extracted and relevant answers are provided. We combine both local mining and global learning approaches to retrieve the answers. Community question answering has been in trend in the recent days. They are seldom moderated, rather open, and thus they have few restrictions, if any, on who can post and who can answer a question. On the positive side, this means that one can freely ask any question and expect some good, honest answers. On the negative side, this takes effort to go through all possible answers and to make sense of them. For example, it is not unusual for a question to have hundreds of answers, which makes it very time consuming to the user to inspect and to winnow. The challenge we propose may help automate the process of finding good answers to new questions in a community-created discussion forum. We do this by finding similar questions through natural language processing. To achieve question answering, a knowledge base with collection of documents needs to be maintained. As both local mining and global learning approaches are adopted here, separation between set of files needs to be maintained. Key concept detection was used to enhance the Question retrieval by Wei-Nan Zhang et al. [1]. Key concepts can be seen as the real refined intent in user queries. Not all detected key concepts are suitable for paraphrasing. For example, for human names, product names, location names and organization names etc., we could not obtain diverse forms by paraphrasing. Hence, for the key concept paraphrase generation step, the above name entities are not considered. We recognize them by using named entity recognizer of the Stanford core-nlp toolkit [2]. Meanwhile, the concepts, including noun phrase and verb phrase, were extracted by using the openNLP[3] chunking tool. In this study, we only consider the verb phrases and noun phrases of the chunking results as key concept candidates.

Figure 1. The framework of key concept paraphrase based question retrieval. It is constructed by three modules, first, the ranking based key concept detection for query refinement, second, the translation based approach to paraphrase mining, by using multiple languages, for query expansion, and third, the unified question retrieval model to integrate key concept and the corresponding paraphrase. [1]

II. EXISTING SYSTEM
It was shown that the inconsistency of community generated health data greatly hindered data exchange, management and integrity [4]. Even worse, it was reported that users had encountered big challenges in reusing the archived content due
to the incompatibility between their search terms and those accumulated medical records [5]. Therefore, automatically coding the medical records with standardized terminologies is highly desired [6]. There have been several efforts dedicated to research on automatically mapping medical records to terminologies. Most of these efforts, however, focused on hospital generated health data or health provider released sources by utilizing either isolated or loosely coupled rule-based and machine learning approaches. Compared to these kinds of data, the emerging community generated health data is more colloquial, in terms of inconsistency, complexity and ambiguity, which pose challenges for data access and analytics. Further, most of the previous work simply utilizes the external medical dictionary to code the medical records rather than considering the corpus-aware terminologies. Their reliance on the independent external knowledge may bring in inappropriate terminologies. Constructing a corpus aware terminology vocabulary to prune the irrelevant terminologies of specific dataset and narrow down the candidates is the tough issue we are facing. In addition, the varieties of heterogeneous cues were often not adequately exploited simultaneously. Therefore, a robust integrated framework to draw the strengths from various resources and models is still expected.

III. PROPOSED SYSTEM

Firstly, we process all the files in csv format for local mining. Processing here means NLP processing where the part of the speech is tagged for all the words. Local mining terminologies may suffer from various problems. The first problem is incompleteness. This is because some key medical concepts not explicitly present in the medical records are excluded. The second one is the lower precision. This is due to some irrelevant medical concepts explicitly embedded in the medical records, and is mistakenly detected and normalized by the local approach. For global learning, we process the set of resources in pdf format. The pdf files are indexed and kept for keyword matching. This process is managed by a admin who is in charge of the medical files. Then users can login to post their queries. Once a user wants answers for the question after he logs in, he just has to post the queries for which relevant answers would be provided. The higher the number of resources we have, the higher is the possibility of getting relevant answers.

Local mining searches for similar questions in previous records of answered queries and returns the result. If relevant answers are not found, we search for answers in the resources. Local mining database gets updated every time, global learning approach is used. This reduces the time of retrieval for the same question the next time because relevant answer can be obtained in local mining itself.

An expert answer might be needed in few cases when both the mining approaches don’t provide us a relevant answer. The inter-expert relationships will be viewed stronger if the experts are professionals in the same or related specific medical areas. This is reflected by their historical data, i.e., the number of questions they have co-answered.

An expert is an authorized doctor who has to register in our application. He can choose his area of specialization. A user who is not satisfied with the answer provided can also post a direct question to doctor. But, an expert cannot be available all the time. So a user needs to wait for a doctor’s answers if needed.
describes the question title in more detail. The question title field contains the type of questions that we would expect to receive from users and this data is the basis for the question retrieval experiments. Machine learning in our approach is achieved by the use of local mining and global learning techniques. Local mining database gets updated by the global learning data's once user posts a newer kind of query to the answering system. The Global learning comprises a large collection of medical related resources in its backend which helps to retrieve a related resource to the Query based on terminology keywords. This Search is completely indexed and thus the retrieval time is faster. In case of resource insufficiency, the Query and the Question will be left in pending state till an expert arrives. Once Experts reviewed the query the answers not only dispatches to the Medical Seekers and also updates the Local Mining Database for future instant retrieval to the related Query from other Users.

Figure 3. the proposed system architecture where both local mining and global learning approach is used for answering user questions.

IV. CONCLUSION

This paper presents a medical terminology assignment scheme to bridge the vocabulary gap between health seekers and healthcare knowledge. The scheme comprises of two components, local mining and global learning. The former establishes a tri-stage framework to locally code each medical record. However, the local mining approach may suffer from information loss and low precision, which are caused by the absence of key medical concepts and the presence of the irrelevant medical concepts. This motivates us to propose a global learning approach to compensate for the insufficiency of local coding approach. The second component collaboratively learns and propagates terminologies among underlying connected medical records. It enables the integration of heterogeneous information. Extensive evaluations on a real-world dataset demonstrate that our scheme is able to produce promising performance as compared to the prevailing coding methods. More importantly, the whole process of our approach is unsupervised and holds potential to handle large-scale data.

V. REFERENCES

[1]. Wei-Nan Zhang, Zhao-Yan Ming, Yu Zhang, Ting Liu, and Tat-Seng Chua, “Capturing the Semantics of Key Phrases Using Multiple Languages for Question Retrieval”


[9]. http://wordnet.princeton.edu/
