Intelligent Ontology based Question Answering System for Medical Domain

D.N. Koggalahewa, J.L. Amararachchi, K.G.D Tharanga *and* S.U. Pilapitiya
Department of Information Technology, Sri Lanka Institute of Information Technology
Malambe, Sri Lanka

Introduction

Irrespective of the domain, the main aim of a Question Answering system is getting a question from the user, comprehending it, searching the answer in an efficient way and presenting the answers to the user. Many methods have been devised for this purpose. This basic idea is using ontology for representing the knowledge and developing the knowledge base. Although the ultimate aim of question answering is finding the exact answer to any question in any context. In today's world of automated content processing, this is inherently a hard task because without a restriction imposed either on the question type or on the user's vocabulary, the question answering process gets a big hit even at the question interpretation phase. The published medical literature and online medical resources are important sources to help physicians make decisions in patient treatment Cimino *et al.*, 2003.

Question answering is a rapid-developing technique that automatically analyses thousands of articles to generate a short text, ideally, in less than a few seconds, to answer questions posed by physicians. Such a technique provides a practical alternative that allows physicians to efficiently seek information at point of patient care. Physicians usually have limited time to browse the retrieved information. For example, studies found that physicians spend on average two minutes or less seeking an answer to a question, and that if a search takes longer, it is likely to be abandoned (Radomski, 1986). Although there are a number of annotated medical knowledge databases available for physicians to use, studies found that most of the resources are not frequently used by physicians in large hospitals due to busy work schedule in their lives (Sackett *et al.*, 2000). Physicians often need to consult literature for the latest information in patient care (Siang *et al*,. 2001). Information retrieval systems (e.g., PubMed) are frequently used by physicians. Another evaluation study showed that it took an average of more than 30 minutes for a healthcare provider to search for answer from the PubMed, which means "information seeking is practical only 'after hours' and not in the clinical setting" (Wikipedia, 2010).

Materials and methods

In this study, we propose natural language processing techniques, adapt existing natural language processing tools, implement, and evaluate MediLink. The system consist of four main components

Question Classifier

Query Generator

Knowledge manager

Text Summarizer

Question classifier

Question Classifier automatically classifies a question posed by a physician/ user into a question type for which a specific answer strategy will be developed. It assigns a question posed by a physician to a specific category for which specific answer strategy is developed. Research has shown that medical questions can be classified by physicians into finite categories namely, Clinical or Non-clinical. The Clinical questions are further divided into General versus Specific; General questions are divided into Evidence and No-evidence.

Once the NLP component generates the semantics Question Classifier will categorize the posed question under one of the above mentioned categories.

Query generator

Query Generator analyses the question to extract noun phrases as the query terms. The sub module Document Retrieval applies the query terms to retrieve documents from either the Web documents or the locally-indexed ontology repository. The module identifies noun phrases from medical questions and then applies the noun phrases as the query terms to retrieve relevant documents. To retrieve definitions that appear in the Web documents, we used a separate medical terms repository and the module will matched for abstract definitions of each terms in the content.

Knowledge manager

Knowledge manager automatically identifies the sentences that provide answers to questions. The sub module Text Summarization removes the redundant sentences and condenses the sentences into a coherent summary. The summary is then presented to the user who posed the question.

We automatically identified lexico-syntactic patterns, the patterns that incorporated both lexicon and syntax information, for identifying definitional sentences. Our strategy is to obtain an exhaustive list of lexico-syntactic patterns that has been generated from a set of definitional sentences. Specifically, we applied the word net definitions with the Google: Definition service. Google: Definition provides definitions that seem to mostly come from web glossaries. Query pattern within a clause is used to extract every noun phrase in both collections of relevant and irrelevant texts.

Text summarizer

We have used our own summarization techniques in the solution. It attempts to condense a stream of text into a shorter version while preserving its information content. Medilink builds on previous summarization and information retrieval techniques. It clusters sentences based on sentence similarity and selects the most representative sentence from each cluster. Medilink clusters sentences using the hierarchical clustering algorithms that have been evaluated in the biomedical domain

Discussion

The Research introduced a methodology of medical knowledge extraction by using ontology based knowledge retrieval approach. Medilink system is a user-driven approach where physician posed the question and system extracts the answer for the given question. For example, when a physician asks "What is the bronchial asthma?" he wants to know not only the definition of this term, but also how to diagnose it and manage it. Essentially, a definition question (i.e., "What is X?") requires answers beyond definitions

(e.g., "what causes X?" and "How to treat X?"). The system should be further tested with different physicians in order to refine the level of extraction of the system.

Conclusions

Physicians are suffering from inefficient medical information retrieval when they require additional information over the medical treatments or diagnosis. Medilink is an effort made to bridge the gap between medical knowledge available and the extraction of knowledge from the available knowledge. The system introduces a methodology of knowledge retrieval from the medical ontology and medical documents by posing real time natural language queries. The system Is capable of making intelligent response over the posed questions. The research covers several techniques on natural language processing (information extraction and text summarization), ontology manipulation and the knowledge representation.

References

Cimino J.J., Li J, M. Graham, L.M. Currie, M. Allen and S. Bakken 2003. Use of online resources while using a clinical information system. AMIA Annu Symp Proc. 2003:175–9

Gosling A.S. and J.I. Westbrook 2004. Allied health professionals' use of online evidence: a survey of 790 staff working in the Australian public hospital system. Int J Med Inform. 73(4):391–401.

Radomski A. M. 1986. Application of System-Process-Goal Approach for description of TRIGA RC1 System", Proceedings of 9th.European TRIGA Nuclear Reactor Users Conference, Oct., 1986,Roma:GA Technologies, TOC-19, USA.

Sackett D, S. Straus, W. Richardson, W. Rosenberg and R. Haynes 2000. Evidence-Based Medicine: How to practice and teach EBM. Edinburgh: Harcourt Publishers Limited.

Siang L, Liu C, Xu Y-Q, Guo B, Shum H-Y. 2001. Real-time texture synthesis by patch-based sampling. ACM Trans Graph. 20(3):127–50.

"Teaching and Learining", 26 February 2010 [Online], Available :http://en.wikipedia.org/wiki/Teaching, Accessed: 20 February 2010