A System for Rapidly Assembling Clinical Knowledge through Curation of Automatically Extracted Knowledge from Medical Knowledge Sources

Deepika Kakrania, MSc, IBM (Presenter); John T.E. Timm, PhD; Jose Morey, MD; Prasanth Prasanna, MD; Tanveer Syeda-Mahmood, PhD

Hypothesis

Automated text-analytic based methods can be used for rapid extraction and correction of clinical knowledge from structured/unstructured medical sources.

Introduction

Clinical decision support systems have predominantly been rule-based and built on fixed pre-determined associations from clinical knowledge. Their efficacy and accuracy depends on the extent of knowledge captured in these systems in the form of facts and assertions. The current approach to populating these knowledge sources are primarily manual through careful handcrafting and often assisted by tools with graphical user interfaces to allow editing. This process is very slow and painstaking, with clinicians only able to enter 10 – 20 assertions per day. Furthermore, it can introduce duplications unless these assertions are crafted using careful underlying knowledge schema that is enabled to search for such duplications. Finally, the extent of knowledge entered is primarily limited by the expertise of the knowledge provider.

Recently, approaches have been developed that go to the other extreme, namely, completely automatic extraction of clinical information from biomedical literature sources and populating knowledge bases automatically. In such approaches, natural language understanding and text mining techniques are often employed to extract essential clinical information [3]. However, the biomedical literature text is not necessarily the best content to infer basic relationships between clinical entities such as diseases and symptoms, which are more easily available in textbook-like sources of online educational material in authentic web references [1, 2]. Furthermore, without a curation process in place, such knowledge generated cannot be treated as authentic clinical knowledge for purposes of reasoning and other uses in a clinical decision support system.

Methods

In this paper, we take an intermediate approach to allow rapid and large-scale assembly of authentic clinical knowledge using a carefully guided, semi-automated method for assembling knowledge.

Specifically, we organize the collection of knowledge around a focus area such as a disease and analyze the most common topical dimensions used in medicine to describe the focus area. For example, in the case of diseases, the dimensions can be symptoms, past medical history, physical exam findings and vitals, etc. We then have clinicians identify authentic textual sections of knowledge corresponding to instance concepts in the focus area and its associated topics. A relationship extraction algorithm is then used to automatically infer relations between clinical concepts found in text and the target concept. The concept extraction process uses authentic vocabularies derived from reference sources such as UMLS and RADLEX to initially
spot clinical concepts after accounting for negations and family history references as described in [4]. Candidate relations are then proposed between concepts detected in a focused text area usually corresponding to a sentence within a paragraph along the chosen dimensions for the disease.

To represent this combined reference knowledge, a lightweight knowledge model as reported in [5] was used. Specifically, the relationships are represented as assertions linking clinical concept nodes in a graph-based representation. To allow for corrections of assertions, a staging version of knowledge graph is used and the automatically extracted assertions are distributed among clinicians for validation. The validation is a simple process of checking off assertions that are found to be clinically meaningful while the rest are dropped. The verified assertions get added to the authentic reference knowledge base for persistence. Since all the verification is done through a web-based interface, a distributed team of clinicians can work collaboratively to quickly validate a large number of assertions.

![Figure 1. Web interface for verification of automatically generated assertions.](image)

**Results**

The above semi-automatic methodology was successfully used to rapidly assemble a large knowledge base of facts and assertions in cardiology covering 54 cardiac diseases. Using several authentic knowledge sources identified by experts, we extracted 1350 documents that contained information about these diseases along dimensions of topics such as symptoms, procedures, treatments, etc. The results of our analysis are described below.

Of the nearly 4000 assertions generated from the automated analysis, 2722 assertions were deemed correct by the clinicians. The overall running time of the algorithm for these assertions was 40 minutes over 1350 documents. In comparison, the average knowledge creation for de novo assertions generated directly by clinicians using their a priori knowledge took an average of 17.2 minutes per assertion which by comparison would have taken 32 days to create the same number (2722) of assertions, thus indicating a tremendous time savings in the creation of knowledge.

**Conclusion**

In this paper we have presented a methodology of semi-automatic assembly of knowledge using automated text analysis and curation by clinicians as a hybrid approach to fast knowledge generation. The evaluation of the system was done for automatic generation of knowledge for over 50 cardiac diseases resulting in over 2700 valid assertions in about 40 minutes of processing time. Finally, the time savings achieved by the semi-automatic process enables a scalable way to build large knowledge bases in future for other disease specialties in medicine.
References


Keywords

Automated Text Analysis, Semi-automatic process, Knowledge Base