Available online at www.sciencedirect.com

ScienceDirect

Data & Knowledge Engineering 66 (2008) 3–17

DATA & KNOWLEDGE ENGINEERING

www.elsevier.com/locate/datak

Assessing clinical trial eligibility with logic expression queries ☆

D.W. Lonsdale *, C. Tustison, C.G. Parker, D.W. Embley

Brigham Young University, Provo, UT 84602, USA

Available online 6 September 2007

Abstract

This paper introduces a system that processes clinical trials using a combination of natural language processing and database techniques. We process web-based clinical trial recruitment pages to extract semantic information reflecting eligibility criteria for potential participants. From this information we then formulate a query that can match criteria against medical data in patient records. The resulting system reflects a tight coupling of web-based information extraction, natural language processing, medical informatic approaches to clinical knowledge representation, and large-scale database technologies. We present an evaluation of the system and future directions for further system development.

Keywords: Clinical trials; Predicate logic; Clinical data; Virtual medical record

1. Background and overview

As electronic texts become more available to researchers (and humans in general), an interesting dichotomy has emerged. On one hand, Web texts cater to users' abilities to read and analyze that information; Web publishers design the data's structure to be easy for humans to digest. Hence it must adhere to conventional syntactic and semantic constraints of the users' natural language. On the other hand, humans have very limited computational capacity for analyzing the vast amounts of electronic information now available. Information extraction research focuses on helping humans access and process large quantities of Web data. Often this work involves devising new strategies and algorithms to convert electronic natural language text into various formats that feed subsequent automatic processing.

The task is complicated by several types of textual layout formats. Text is often classified into one of three categories: unstructured (or free), structured, and semistructured [1]. Unstructured text is the most natural for humans to process, but treating the information automatically is nontrivial. Structured text is stored in a very rigid format (e.g. a database or a table) and hence more readily processed automatically, but is often less

☆ This work was partially funded under National Science Foundation Information and Intelligent Systems grant IIS-0083127. See also http://www.deg.byu.edu.

* Corresponding author.

E-mail addresses: lonz@byu.edu (D.W. Lonsdale), tustison@gmail.com (C. Tustison), craigparker@cgpmd.com (C.G. Parker), embley@cs.byu.edu (D.W. Embley).

doi:10.1016/j.datak.2007.07.005

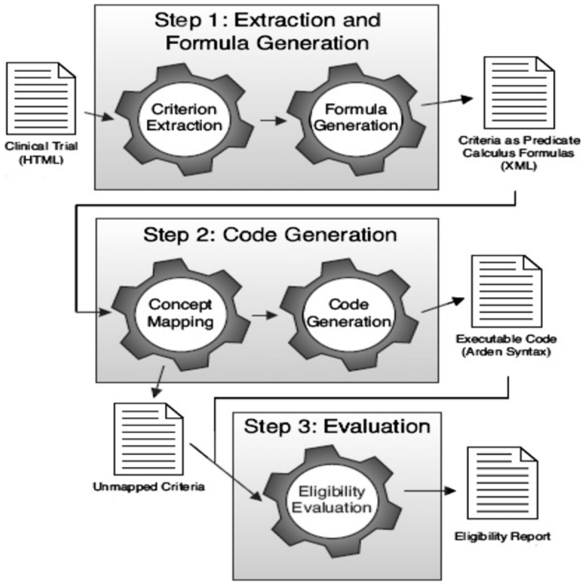

Fig. 1. System processing stages including data formats (input, intermediate, and output).

natural for humans to work with. Semistructured text falls somewhere in between; some structure is imposed—often just enough to render it not quite grammatical—though not enough to help in automatically processing the contents. With the advent of various markup languages and other annotation conventions, Web text often includes other extratextual information that may or may not aid in extracting information. In this paper we discuss processing a repository of semistructured medical text.

Researchers design information extraction systems to perform various tasks, and these tasks require various levels of linguistic processing. Some systems are only concerned with parsing out the extracted information and therefore only require the use of a syntactic parser. Others need more in-depth processing and include a semantic component that can give some meaning to the extracted information. Yet other systems are dependent on real-world knowledge and require a pragmatic component to relate the data gathered from the system to outside information.

One area receiving recent attention is the medical domain. Much of the natural language processing (NLP) research done with medical literature has involved developing systems that extract different types of relationships from text. For example, NLP techniques have been used to extract interesting and novel relationships from Medline[1] abstracts. The Medline repository contains vast amounts of useful information about various disease- and health-related issues. Many researchers have succeeded in extracting various types of relationships found in this repository, including gene relations [2], protein relationships [3,4], acronym–meaning pairs [5], abbreviation definitions [6], and molecular binding relationships [7].

For its part, the field of medical informatics has produced large-scale resources, largely in database format, that specify the vast knowledge required for medical research and patient services. Highly specialized tools for representing clinical information and patient data have also been developed. Unfortunately, there has been only a modest amount of crossover between the NLP and medical informatics fields. The topic of information extraction is a salient one for demonstrating how applications can leverage the developments from both fields.

This paper describes our approach to identification, extraction, and query formulation of information regarding medical clinical trials. Fig. 1 shows an overview of the system. In Step 1, extraction and formula generation, we extract patient criteria from a web-based natural language description of qualifications for

[1] See http://www.medlineplus.gov.