Clinical Trials Text Mining

A dissertation submitted to The University of Manchester for the degree of Master of Science by Research in the Faculty of Engineering

Jonathan Ho-Hio-Hen
Msc Project

Supervisors :

Uli Sattler :

Professor in the Information Management Group within the School of Computer Science of the University of Manchester.

Cassie Gregson :

Senior Informatics Scientist, Biomedical Informatics AstraZeneca.
Research & Development | Clinical Development.

School of Computer Sciences

This project is sponsored by AstraZeneca Research

2011

Abstract

The success of a clinical trial depends partly on the design choice made prior to the trial. To support the AstraZeneca research and development team design choice, a tremendous amount of trial records can be consulted via diverse databases. However, some part of the trial records are unstructured raw text and the number of trials is too overwhelming to be processed manually.

To avoid repeating trials and to prevent adverse events to occur during a trial, specific information requests are formulated for each new trial initiated. The query system must be able to mine the text to answer the formulated information requests. However, the range of treatments introduced in each new trial is vast, thence, gathering all knowledge about the treatment is compulsory to build the query and time-consuming. To overcome this obstacle, some information systems offers taxonomies to build queries. Taxonomies are used to store information about a domain via a hierarchy of preferred terms. Using the list of synonyms attached to a preferred term reduces the time required to gather information. However, knowledge can be represented in a much more expressive and formal way than taxonomies such as OWL ontologies.

OWL ontologies are a set of logical axioms depicting a class hierarchy and class relations. Inferring logical consequences from asserted facts (axioms) in a OWL ontology with a reasoner leads to the discovering of)new knowledge. Querying over OWL ontologies populated with information extracted from the text can improve the queriability, the recall and precision of a query. Indeed, the knowledge discovered by the reasoner can be integrated to the query system to retrieve more facts (recall) and only relevant facts (precision) while providing an easier way to build query (queriability) through the use the class hierarchy and class relations asserted in an existing OWL ontology capturing the pharmaceutical domain. Eventually, the phase of collecting information related to the treatment can be drastically shortened by using existing OWL ontology.