Pancreatic Cysts Identification Using Unstructured Information Management Architecture

Saeed Mehrabi[1], C. Max Schmidt[2], Joshua A. Waters[2], Chris Beesley[3], Anand Krishnan[1], Joe Kesterson[3], Paul Dexter[3], Mohammed A. Al-Haddad[4], Mathew Palakal[1]

[1]*School of Informatics, Indiana University, Indianapolis, IN, USA;* [2]*Department of Surgery, Indiana University, Indianapolis, IN, USA;* [3]*Regenstrief Institute, Indianapolis, IN, USA;* [4]*Department of Medicine, Division of Gastroenterology, Indiana University, Indianapolis, IN USA.*

Pancreatic cancer is one of the deadliest cancers, mostly diagnosed at late stages. Patients with pancreatic cysts are at higher risk of developing cancer and surveillance of these patients can help with early diagnosis. Much information about pancreatic cysts can be found in free text format in various medical narratives. In this retrospective study, a corpus of 1064 records from 44 patients at Indiana University Hospital from 1990 to 2012 was collected. A natural language processing system was developed and used to identify patients with pancreatic cysts. The input goes through series of tasks within the Unstructured Information Management Architecture (UIMA) framework consisting of report separation, metadata detection, sentence detection, concept annotation and writing into the database. Metadata such as medical record number (MRN), report id, report name, report date, report body were extracted from each report. Sentences were detected and concepts within each sentence were extracted using regular expression. Regular expression is a pattern of characters matching specific string of text. Our medical team assembled concepts that are used to identify pancreatic cysts in medical reports and additional keywords were added by searching through literature and Unified Medical Language System (UMLS) knowledge base. The Negex Algorithm was used to find out negation status of concepts. The 1064 reports were divided into sets of train and test sets. Two pancreatic-cyst surgeons created the gold standard data (Inter annotator agreement K=88%). The training set was analyzed to modify the regular expression. The concept identification using the NegEx algorithm resulted in precision and recall of 98.9% and 89% respectively. In order to improve the performance of negation detection, Stanford Dependency parser (SDP) was used. SDP finds out how words are related to each other in a sentence. SDP based negation algorithm improved the recall to 95.7%.

Mentors: Mathew Palakal, School of Informatics, IUPUI; C. Max Schmidt, Department of Surgery, Indiana University.