[Inspiratron.org - Natural language processing, machine learning and cybersecurity](https://inspiratron.org)

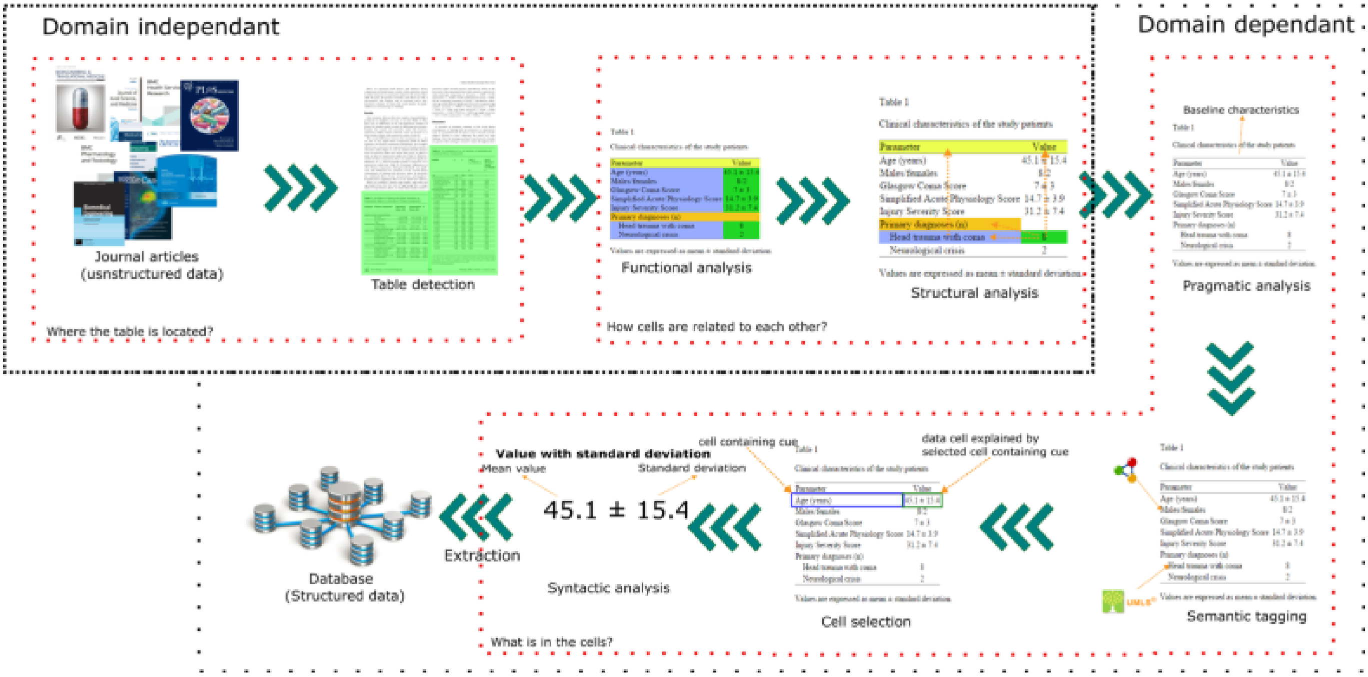
[New Paper] Information extraction from tables in literature

by Nikola Miloševi? - Monday, April 22, 2019

https://inspiratron.org/blog/2019/04/22/new-paper-information-extraction-from-tables-in-literature/

About two months ago (February 2019), a paper that resulted from my Ph.D. work has been published in the International Journal of Document Analysis and Recognition. The paper is titled "A framework for information extraction from tables in biomedical literature". I would like to just outline a couple of things related to this paper and give a bit of background.

Basically, this is a framework paper, that encapsulates and puts together everything that was done during my Ph.D. Previously, I have also published a couple of paper that are more focusing on experimenting with workflow (Extracting Patient Data from Tables in Clinical Literature - Case Study on Extraction of BMI, Weight and Number of Patients) or explaining functional and structural table analysis (Disentangling the Structure of Tables in Scientific Literature). This paper takes the whole task of information extraction and presents a framework for table analysis. This framework consists of table detection, functional, structural, pragmatic analysis, semantic tagging, cell selection and at the end syntactic analysis. The framework is presented in the image below.



The workflow of the framework presented in the paper

The paper evaluates each of these steps and compares rule based and machine learning approaches on several cases of information extraction in biomedical and more specifically clinical domain. The cases that were evaluated was extraction of patient age, gender, adverse event names, and number of patients in a given clinical arm. Demographic information are quite often presented in tables and therefore usual techniques of information extraction from text would not be useful.

The paper concluded, with given a couple of case studies, that some of the tasks are easier to perform using machine learning (such as pragmatic analysis), while some tasks are quite expensive to train and the structure of table and usual naming conventions can facilitate fast creation of rules and heuristics. It was all demonstrated on two case studies. One case study was on clinical papers from PubMedCentral about pulmonary diseases such as COPD and Asthma. We have extracted information about patient demographics and respiratory tests. The other case study was on extracting adverse events that were caused by drug-drug interactions that are reported in drug labels presented in DailyMed.

The presented framework is supported by two tools that I have developed during my Ph.D. The first tool is TableDisentangler. This tool as input takes a folder with XML files downloaded from PubMedCentral and performs functional, structural and pragmatic analysis. The output is stored in the mySQL database. The second tool is called TableInOut. TableInOut takes as input output from the TableDisentangler. It is basically a wizard for creating extraction rules that can be then stored again in the database using an extraction template.

I really hope this paper will help and advance the processing of information from tables in literature. The tools are limited at the moment to XML documents, however, tool contain readers for XMLs used in PMC and DailyMed as well as HTML reader. This makes large majority of XML based formats in which tables are presented. However, new readers can be added as both tools are open source and you can find them on GitHub.

Happy reading, and if you see the ideas from the paper or tools useful, please cite them.

The abstract of the paper is:

> The scientific literature is growing exponentially, and professionals are no more able to cope with the current amount of publications. Text mining provided in the past methods to retrieve and extract information from text; however, most of these approaches ignored tables and figures. The research done in mining table data still does not have an integrated approach for mining that would consider all complexities and challenges of a table. Our research is examining the methods for extracting numerical (number of patients, age, gender distribution) and textual (adverse reactions) information from tables in the clinical literature. We present a requirement analysis template and an integral methodology for information extraction from tables in clinical domain that contains 7 steps: (1) table detection, (2) functional processing, (3) structural processing, (4) semantic tagging, (5) pragmatic processing, (6) cell selection and (7) syntactic processing and extraction. Our approach performed with the *F*-measure ranged between 82 and 92%, depending on the variable, task and its complexity.

Full paper (Open access): https://link.springer.com/article/10.1007/s10032-019-00317-0