Spá: a web-based viewer for text mining in Evidence Based Medicine

Kuiper, J[1]., Marshall, I.J.[2], Wallace, B.C.[3], and Swertz, M.A.[1]

[1] University of Groningen
[2] Kings College London
[3] Brown University

Abstract. When doing sentence extraction or document level predictions on unstructured text the results of trained models are often hard to interpret within their context. Furthermore presenting results to end users can be a considerable user interface challenge. This problem is especially prevalent in Evidence Based Medicine, where most of the research findings are only published as unstructured PDF documents. To this end we present **Spá**(1) [4], a generic web-based visualizer for sentence and document level classifiers on PDFs. Spá allows the results of sentence extractions to be visualized within the PDF document itself, and allows other results to be presented alongside it.

`revision: 176f216, date: 2014-04-14`

1 Introduction

Finding sentences or words with particular characteristics within a larger document is an important task in natural language processing and machine learning. For example, one may wish to identify the most important sentences in a document to automatically generate a summary, or match a certain ontology to impose structure.

Dealing with unstructured text is of particular importance in research areas where most findings are only published in that form. In evidence based medicine, for example, the results of clinical trials, which assess the safety and efficacy of treatments, are often only published as PDF documents.

Furthermore to arrive at informed decisions about a specific clinical question, many clinical trials need to be pooled and summarized. The process of pooling and summarizing clinical trials is called *systematic reviewing*, and forms the corner-stone of current evidence based medical practice. Systematic reviewing consists of specifying an inclusion criteria (i.e., the criteria studies must satisfy to be included in the review), searching the literature, screening the retrieved citations to identify eligible studies and, finally, summarizing the relevant evidence. But achieving this aim is complicated by the massive numbers of trials that are conducted: for example, the Cochrane Library alone indexes 286,418 trials as having been conducted in the last decade (2). While publishing standards are improving and novel tools for systematic reviewing are being created to

[4] freely available under GPLv3 at `https://github.com/joelkuiper/spa`

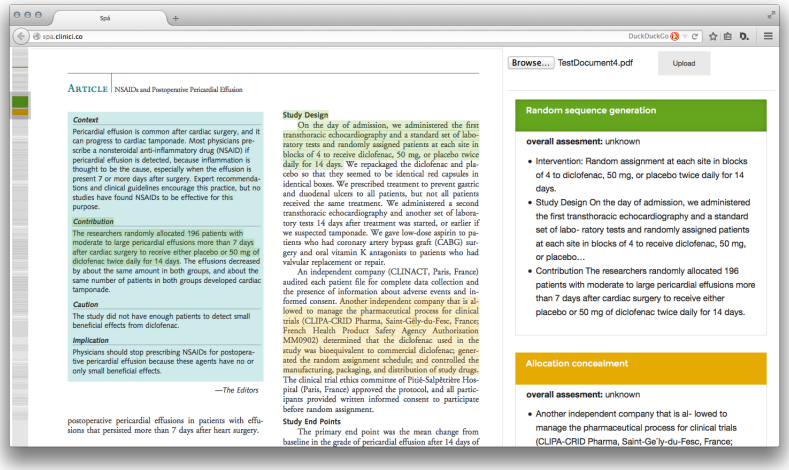

Fig. 1. Screenshot of Spá

address this problem, a lot of legacy publications still only exist as PDF documents. This raises questions about the sustainability of systematic reviews in its current form.

Thus to aid the process of systematic reviewing we created a generic web-based tool that allows the visualization of annotations within a PDF documents, or meta-information alongside it. The aim is to have a pluggable system to allow for semi-automated (machine assisted) screening, data extraction and data summarization.

2 Architecture

Spá relies on Mozilla pdf.js[5] for visualization of the document and text extraction. The results of the text extraction are processed server-side by a variety of (pluggable) pipelines, as outlined in figure 2. Results from these pipelines, which could be complicated machine learning systems, are communicated back to the browser and displayed. For each of the annotations the relevant nodes in the document are highlighted and a minimap, inspired by substance.io, is projected to show where it resides within the document. The user can then interactively activate and inspect the specific type of results. TODO?

3 Case Study

As a case study we evaluate the automatic assessment of Risk of Bias in clinical trial publications. TODO

[5] `http://mozilla.github.io/pdf.js/`

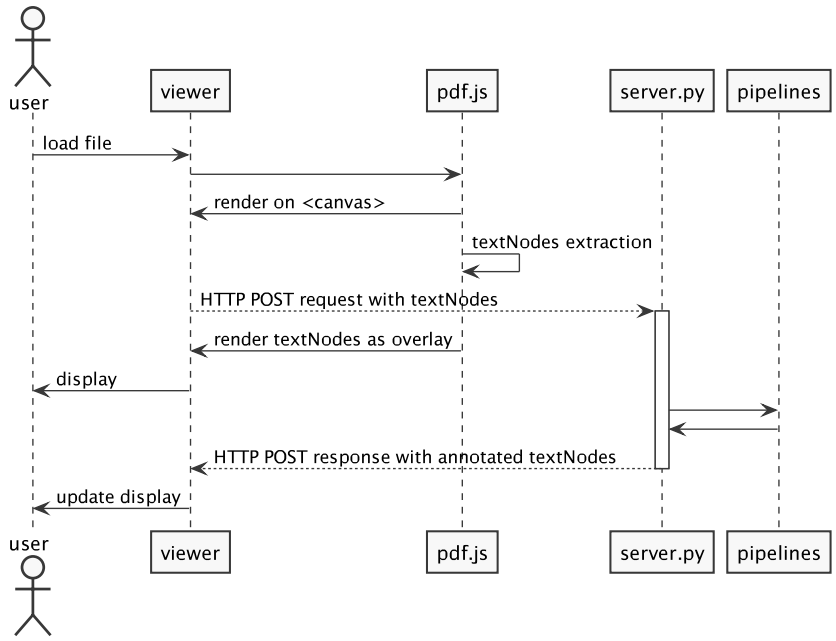

Fig. 2. Sequence diagram of a typical request-response

4 Conclusion & Future work

We present a web-based tool for interactive visualization of annotations and metadata on PDF documents. This allows users to see the results from machine learning systems within the context of a specific document. Moreover, we present a case study for Evidence Based Medicine by automatically extracting potential Risks of Bias, with supporting sentences.

However, we believe the tool to be useful for a much wider range of text mining and machine learning applications. To increase the generality of the tool work is being done to support a consumer/producer pattern for the pipelines, allowing developers to quickly plug-in new systems.

Bibliography

[1] Kuiper, J., Wallace, B.C., Marshall, I.J.: Spa. http://figshare.com/articles/Spa/997707 (April 2014)

[2] Valkenhoef, G., Tervonen, T., Brock, B., Hillege, H.: Deficiencies in the transfer and availability of clinical trials evidence: a review of existing systems and standards. BMC Medical Informatics and Decision Making **12**(1) (2012) 95