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 P.O. Box 30001, 9700 RB Groningen
{joel.kuiper,m.a.swertz}@rug.nl
[2] King's College London, London SE1 3QD, UK
iain.marshall@kcl.ac.uk
[3] Brown University, Providence, RI 02906, USA
byron_wallace@brown.edu

Abstract. Unstructured PDF documents remain the main vehicle for dissemination of scientific findings. Those interested in gathering and assimilating data must therefore manually peruse published articles and extract from these the elements of interest. Evidence-based Medicine provides a compelling illustration of this: many person-hours are spent each year extracting summary information from articles that describe clinical trials. Machine learning provides a potential means of mitigating this burden by automating extraction. But, for automated approaches to be useful to end-users, we need tools that allow domain experts to interact with, and benefit from, model predictions. To this end, we present an web-based tool called Spá[4] that accepts as input an article and provides as output an automatically visually annotated rendering of this article. More generally, Spá provides a framework for visualizing predictions, both at the document and sentence level, for full-text PDFs.

revision: 447664b, date: 2014-04-15

1 Introduction

Finding sentences or words with particular characteristics within a document is an important task in natural language processing and machine learning.

Consider Evidence-based Medicine (EBM) [2], which addresses clinical questions by identifying and synthesizing data from all relevant published articles. When doing this one needs, for example, to assess the *risk of bias* for a particular study across different domains, like bias due to improper blinding of participants and personnel. For this task, one wants to make a summary judgement (e.g., low risk of bias) and simultaneously extract the sentences supporting that judgement.

Extracting such information from the unstructured text of clinical trial articles is a laborious process. Machine learning methods provide the machinery to automate such extractions; as they can effectively impose structures onto PDF's. But if such technologies are to be practically useful, we need tools to visualize

[4] available under GPLv3 at https://github.com/joelkuiper/spa [1]

the predictions. Here we describe Spá which aspires to realize this aim. Spá is an open-source, web-based tool that can incorporate state-of-the-art machine learning predictors to automatically annotate PDF articles. Furthermore we have created machine learning systems on describing clinical trials with risk of bias predictions and extracted sample sizes. This tool is useful for practitioners of EBM and other biomedical researchers.

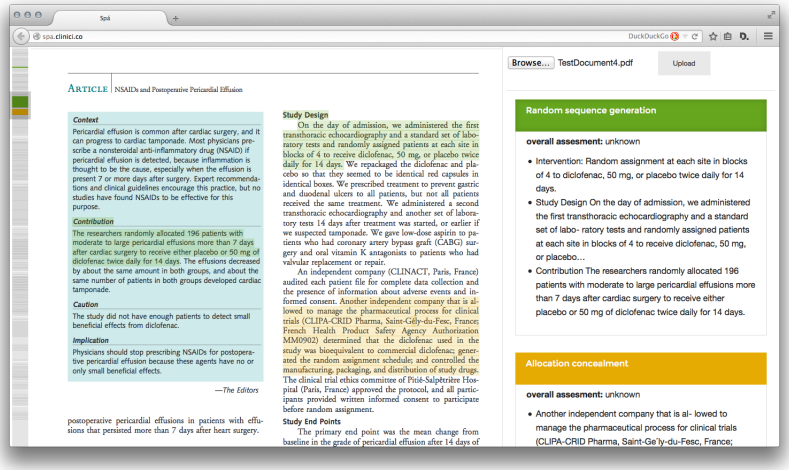

Fig. 1. Screenshot of a PDF with highlighted risk of bias

While our application of interest is EBM, we emphasize that the tool can be used in any domain. Thus the contribution of this work is two-fold, as we present: (1) a practical tool that incorporates machine learning to help researchers rapidly assess published medical articles, and, (2) a general open-source web tool for visualizing the predictions of trained models from full-text articles. These contributions are described further in sections 2 and 3, respectively.

2 Automating Evidence-based Medicine

2.1 Systematic Reviews

The process of pooling and summarizing clinical trials is called *systematic reviewing*, and forms the corner-stone of current EBM. Systematic reviewing consists of specifying an inclusion criteria, 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 [3]. While methods and publishing

standards are improving, a lot of legacy publications will remain only as PDF documents.

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

2.2 Machine Learning Strategies

- Basically show that we're using state of the art
- Briefly talk about cochrane DB / distant supervision
- KDD stuff (briefly)
- Multi-task stuff!

3 Architecture overview

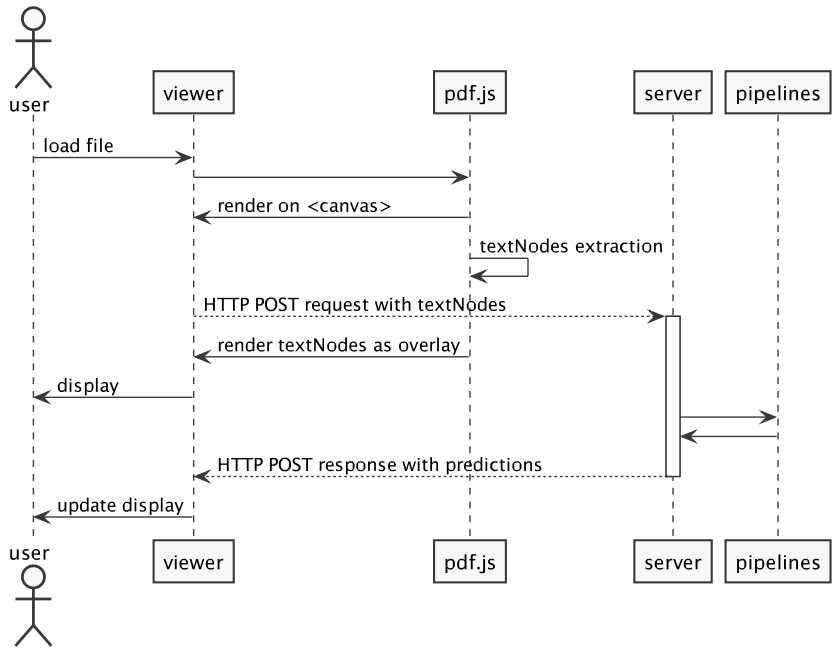

Fig. 2. Sequence diagram of a typical request-response

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

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

variety of 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 using React components[6].

For each of the annotations the relevant nodes in the document are highlighted and a custom scrollbar, inspired by substance.io, that acts as a mini-map is projected to show where it resides within the document. The user can interactively activate and inspect specific results.

4 Conclusion & Future work

We present a web-based tool for interactive visualization of annotations and meta-data 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 EBM 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 protocol for the pipelines, allowing developers to quickly plug in new systems. Furthermore, work is being done on allowing users to save selected annotations, possibly embedded within the document itself, for sharing and off-line use.

Acknowledgments Part of this research was funded by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 261433

References

1. Kuiper, J., Wallace, B.C., Marshall, I.J.: Spa. `http://figshare.com/articles/Spa/997707` (2014)
2. Sackett, D.L., Rosenberg, W.M., Gray, J., Haynes, R.B., Richardson, W.S.: Evidence based medicine: what it is and what it isn't. BMJ: British Medical Journal **312**(7023) (1996) 71–72
3. 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

[6] `http://facebook.github.io/react`