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. Summarizing the evidence about medical interventions is an immense undertaking, in part because unstructured Portable Document Format (PDF) documents remain the main vehicle for disseminating scientific findings. Clinicians and researchers must therefore manually extract and synthesise information from these PDFs to be published in *systematic reviews*. We introduce Spá,[4,5] a web-based viewer that enables automated annotation and summarisation of PDFs via machine learning (ML). To illustrate its functionalility, here we use Spá to semi-automate the assessment of bias in clinical trials. Spá has a modular architecture, therefore the tool may be widely useful in other domains with a PDF-based literature, including law, physics, and biology.

1 Introduction

Imposing structure on full-text documents is an important and practical task in natural language processing and machine learning. *Systematic reviews* are an instructive example. Such reviews aim to answer clinical questions by providing an exhaustive synthesis (textual and statistical) of the current evidence in all published literature. They are fundamental tools in Evidence-based Medicine (EBM) [2]. To produce these reviews data, like potential biases, must be manually extracted from the literature. These extraction tasks are extremely labourous, but could potentially be (semi-)automated using machine learning approaches.

As an example, we consider the risk of bias assessment, in which reviewers assess, e.g., whether study participants and personnel were properly blinded. Assessing risk of bias is a time-consuming task. A single trial may typically take a domain expert ten minutes [3], and a single review may typically include several dozen trials. Making matters worse, due to low rates of reviewer agreement it is regarded as best practice to have each study assessed twice by independent reviewers who later come to a consensus.[4]

[4] From the Old Norse word spá or spæ referring to prophesying (prophecy)

[5] Source code available under GPLv3 at `https://github.com/joelkuiper/spa` [1]; demo available at `http://spa.clinici.co/`

Machine learning methods could provide the machinery to automate such extractions; as they can effectively impose the desired structure onto PDFs. But if such technologies are to be practically useful, we need tools that visualize these model predictions and annotations. Here we describe Spá, which aspires to realize this aim.

Spá is an open-source, web-based tool that can incorporate machine learning to automatically annotate PDF articles. As a practical demonstration of this technology, we have built a machine learning system that automatically annotates PDFs to aid EBM. This tool is unique in that it leverages state-of-the-art machine learning (ML) models applied to full-text articles to assist practitioners of EBM.

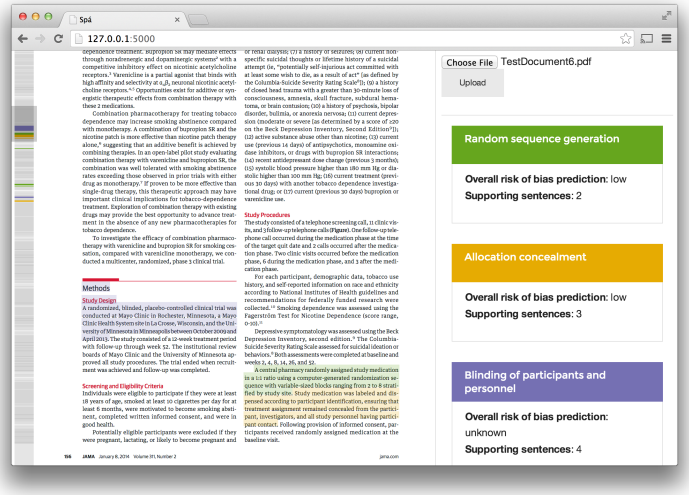

Fig. 1. Screenshot of a PDF with highlighted risk of bias. Here the risk of bias is assessed to be low, for example, and one of the supporting sentences for this assessment describes the randomization procedure (highlighted in green).

While our application of interest is EBM, we emphasize that the visualization tool can be used for any domain in which one wants to annotate PDFs. 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 the risk of biases in published biomedical 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.

2 Automating Evidence-based Medicine

2.1 Machine Learning Approaches

To automatically assess the study risk of bias, we have leveraged the Cochrane Database of Systematic Reviews (CDSR) in lieu of manually annotated data,

which would be expensive to collect. The CDSR contains descriptions and data about clinical trials reported in existing systematic reviews. We match the full-texts of studies to entries in the CDSR, which contains risk of bias assessment; providing document level labels. The CDSR also contains quotations that reviewers indicated as supporting their assessments. We match these strings to substrings in the PDFs to provide sentence-level supervision. This can be viewed as a *distantly supervised* [5,6] approach.

From a ML vantage, we have two tasks for a given article: (1) predict the overall risk of bias for each of the domains of interest, and (2) extract the sentences that support these assessments. For both tasks we leverage standard bag-of-words text encoding and linear-kernel Support Vector Machines (SVMs). Because the risk of bias predictions are correlated (across domains), we take a *multi-task* [7] approach to classification and jointly learn a model for the domains. We accomplish this by way of a feature space construction that includes both shared and domain-specific terms (similar to the domain adaptation approach in [8]). Specifically, we first make sentence level predictions (using one set of trained models), and then insert features representing the tokens (words) in the predicted sentences for exploitation by the document level classifier (further details specified in [9]). Figure 1 shows the system in use.

3 Spá Architecture Overview

Spá relies on Mozilla pdf.js[6] for visualization of the document and text extraction. The results of the text extraction are processed server-side by a variety of processing topologies, as outlined in figure 2. Results are communicated back to the browser and displayed using React components.[7]

For each of the annotations the relevant nodes in the document are highlighted. A custom scrollbar[8] that acts as a 'mini-map' is projected to show where annotations reside within the document. The user can interactively activate and inspect specific results.

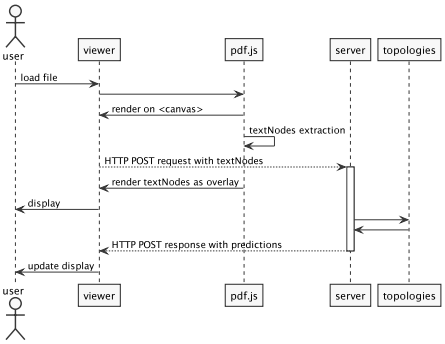

Fig. 2. Sequence diagram of a typical request-response in Spá.

4 Future work

We have presented a web-based tool for visualization of annotations and marginalia for PDF documents. Furthermore, we have demonstrated the use of this

[6] `http://mozilla.github.io/pdf.js`
[7] `http://facebook.github.io/react`
[8] inspired by substance.io

system within the context of EBM by automatically extracting potential risks of bias.

More generally, we believe the tool to be potentially useful for a much wider range of text mining and machine learning applications. To increase the generality of the tool we are developing a pluggable system for processing topologies, allowing developers to quickly plug in new systems for automated PDF annotation. Furthermore, we are working to allow users to save selected annotations, possibly embedded within the document itself, for sharing and off-line use. The vision is to have an extensible system for semi-automated (machine assisted) screening, data extraction and data summarization for EBM, and to allow rapid development of similar systems for other domains.

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. Hartling, L., Bond, K., Vandermeer, B., Seida, J., Dryden, D.M., Rowe, B.H.: Applying the risk of bias tool in a systematic review of combination long-acting beta-agonists and inhaled corticosteroids for persistent asthma. PloS one **6**(2) (January 2011) e17242
4. Hartling, L., Ospina, M., Liang, Y.: Risk of bias versus quality assessment of randomised controlled trials: cross sectional study. BMJ **339** (2009) b4012
5. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL, Association for Computational Linguistics (2009) 1003–1011
6. Nguyen, T., Moschitti, A.: End-to-end relation extraction using distant supervision from external semantic repositories. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2, Association for Computational Linguistics (2011) 277–282
7. Evgeniou, T., Pontil, M.: Regularized multi–task learning. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM (2004) 109–117
8. Daumé III, H.: Frustratingly easy domain adaptation. In: Association for Computatoinal Linguistics (ACL). Volume 1785. (2007)
9. Marshall, I.J., Kuiper, J., Wallace, B.C.: Automatically assessing bias in clinical trials and extracting supporting sentences. In: under review at Knowledge Discovery in Databases (KDD). (2014)