Reviewer's report

Title: ExaCT: automatic extraction of clinical trial characteristics from journal publications

Version: 2 Date: 17 March 2010

Reviewer: Patrick Ruch

Reviewer's report:

Major

Detecting boundaries of textual sections such as bibliographical sections is not trivial; It is less a problem with XML, but in HTML, you can have links at the beginning of the HTML, which refer to any section in the body of the article. What is the percentage of success/failure of your system on such a basic task. Tbahriti et al report that they were able to recover bibliographical sections for about 3200 full-text in a collection of 12500. I guess it means that separating between bibliographical information and body of the article should not be that trivial. The same should apply to section detection. Any idea about the success rate of your system? You working with a limited set of publishers/journals (5)... which clearly questions the scalability of your approach. If you cannot provide estimates regarding the specificity of your architecture, could you at least elaborate on such a major issue?

Not only the journal set is tiny but the sample (132) is really peanut, I would expect a discussion on it.

Minor

I do not understand that sentence: "After that, the remaining sections are split into sentences, and each sentence is annotated with the hierarchy path of the section heading where it originated (e.g. Methods->Patients)." what about subsections, what means "Patient" in the parenthesis, a section like "Population"?

There is more efforts going on the classify rhetorical features (Collier, Lin, Srinivasan...) than PICO features could not you simple map one more or less to the other, at least partially to take advantage of existing and available models?

--


Level of interest: An article whose findings are important to those with closely related research interests
Quality of written English: Acceptable

Statistical review: No, the manuscript does not need to be seen by a statistician.