Author's response to reviews

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

Authors:

Svetlana Kiritchenko (svetlana.kiritchenko@nrc-cnrc.gc.ca)
Berry de Bruijn (berry.debruijn@nrc-cnrc.gc.ca)
Simona Carini (simona.carini@ucsf.edu)
Joel Martin (joel.martin@nrc-cnrc.gc.ca)
Ida Sim (sim@medicine.ucsf.edu)

Version: 3 Date: 4 May 2010

Author's response to reviews: see over
Dear Editor,

Enclosed please find a revised version of our manuscript “ExaCT: automatic extraction of clinical trial characteristics from journal publications”. The manuscript has been modified to address the comments made by the reviewers. The most significant changes have been highlighted in the text of the manuscript. The rest of this letter discusses the reviewers’ comments point-by-point and specifies the smaller changes made to the text.

We would like to thank the reviewers for their insightful critique and valuable remarks and believe that we have adequately addressed all their concerns. We have also copyedited the manuscript to improve the style of written English. The revised manuscript conforms to the journal style.

Sincerely,

For the authors,
Dr. Svetlana Kiritchenko

Reviewer 1 (Ann McKibbon)

Major

We have re-written the abstract’s first two sections “Background” and “Methods” (p. 2) and parts of the Background section (pp. 3-8) to clarify the main message of the manuscript.

Minor

The minor issues have been also addressed throughout the text.

1. The first paragraph of the Background section has been re-written and two references have been added to support the statements.

2. The phrase “journal publications on randomized clinical trials” has been replaced with “journal publications reporting on randomized controlled trials” throughout the text.

3. The mention of the future plans for extending the approach to handle other document types has been removed from the Background section. The future plans are now discussed at the end of the Discussion section.

4. The phrase “information extraction” has been replaced with “IE” throughout the text.

5. Page 5. The sentence about BMJ Clinical Evidence and “Cochrane Reviews” has been changed.

6. The 21 information elements selected for the study are now listed earlier in the Background section (p.4).
7. Page 6. We meant “data analysis that is methodologically rigorous”. The sentence has been modified.

8. All mentions of “i.e.” and “e.g.” have been checked for correct usage.

9. The degrees have been removed from the Author Information section.

10. “US” has been added to National Library of Medicine in the Acknowledgements section.

11. References have been revised for completeness and compliance with the journal style.

12. The importance of the study to various audiences is now stated in paragraphs 1-3 of the Background section (pp. 3-4) and also recapped in the last sentence of the Conclusion section (pp. 26-27).

Reviewer 2 (Patrick Ruch)

Major

1. We completely agree that the section detection task is far from trivial. In the current work, we employ a solution based on the relatively consistent use of HTML tags within an article (see Methods -> System Design -> Pre-processing, pp. 10-11). This method, as estimated on the test set, finds the exact section/subsection boundaries with ~70% precision and ~75% recall. However, it is able to roughly split most of the articles into the Methods, Results, Discussion, and References sections. This level of performance is acceptable for our application, since the section detection step is not crucial for the overall system. It does improve the performance of the classification component by ~5%, but the most success comes from the two-level architecture (the classification and extraction components).

We have added a paragraph on this topic to Methods -> System Design -> Pre-processing, pp.10-11.

2. The sample size of 2772 training tasks and 1050 test tasks (identified in 132 training documents from 22 journals, 50 test documents from 25 journals) was adequate to demonstrate the applicability and efficacy of the system. One pressure that limited the size of the sample was that each (full-text) document was manually curated to locate the description of the 21 clinical trial characteristics. Manual curation entails a considerable amount of work. We would know that a sample size is too small if we could not demonstrate reliable generalization outside of the sample. In the case of this study, the results on the training data (leave-one-out cross-validation) and held-
out test data were found to be consistent with each other. Therefore, we believe that our evaluation gives a fair indication of the system’s performance in real-life settings.

We have added a note on this topic to the Discussion section on p.22.

Minor

1. The sentence has been corrected (see p.11).

2. The rhetorical structure of a document (e.g. Introduction, Methods, Results, Conclusion) is a helpful first step towards locating the description of the PICO elements of a study. In many cases, the PICO elements will be found in the Methods section. So, the rhetorical structure identification helps to narrow down the context for the PICO elements, but is not sufficient for information extraction. In our work, the rhetorical structure of a document is partially captured through the section detection since scientific medical articles often follow the standard practice of section naming: Abstract, Introduction, Methods, Results, Discussion, References.