Reviewer’s report

Title: Fine-grained Information Extraction from German Transthoracic Echocardiography Reports

Version: 2
Date: 26 May 2015

Reviewer: Aurélie Névéol

Reviewer’s report:

**Disclaimer: I see from the editorial system that this manuscript is a revised version of a previous submission. However, the material I had access to does not include previous reviews or point-by-point response to previous reviews from the authors. As a result, I am not able to assess whether any previous comments were addressed suitably. My review is entirely based on the current version of the manuscript.**

This paper reports on the development of a terminology and information extraction system for echocardiography reports in German. The work is conducted on a sizeable corpus of 70k clinical notes. The authors take advantage of redundancy in the clinical documents to extract key concepts for the terminology. The information extraction is rule-based and relies on a few identified document structure types. Overall this work is very interesting, and seems significant. However, the manuscript would benefit from describing the methods, goals and positioning of the work with increased clarity.

Major comments:

- Most of the related work reviewed and discussed addresses NLP and terminology development in English. This work is conducted on German. It would be good have more discussion of other work in clinical NLP in German – I am aware that there are far more efforts addressing English, however, there is a recent paper reporting on terminology and ontology resources for German in the biomedical domain (Schultz et al. 2013) as well as work that proposes software for extracting concepts from clinical text (Hahn et al. 2002) or uses RadLex to extract information from radiology reports (Bretschneider et al. 2013).

- The decision to develop a new terminology is not explained – the assumption is that the contents of the created terminology cannot be found in existing resources. However, it seems from section 2.5 that the terminology was already contained in an existing source in German. Why was this source not used directly? Also, can the terminology and/or existing German standardized resource be mapped to the UMLS? Could it be integrated in the UMLS? Is it made available to the research community in some other way?

- The terminology development process is not clear to me. I find the idea of aggregating redundant content interesting to prioritize the text content to be reviewed by terminology curators. However, it is unclear how many curators
worked on terminology development? If so, was part of the corpus processed by more than one curator, so that the terminology extraction performed on this content could be compared (e.g. by computing inter-annotator agreement)? Was there an evaluation of the terminology contents, other than a domain expert/curator validating it? Also, it is unclear how much of the corpus was used for terminology development? Was the portion of the corpus used to evaluate the information extraction tool separate from the corpus used for terminology development and information extraction tool development? – It should be.

- How was the gold standard for evaluating the information extraction tool produced? It seems that the tool was applied to the corpus, and then the output was validated by a curator. Is that correct? Ideally, more than one curator should be involved for at least part of the process in order to calculate inter-curator agreement. Was the production of this gold standard reference a separate process from terminology development? I am not sure it was, and that could be a severe methodological flaw. This needs to be explained clearly.

- The information extraction process relies on the detection of document structure using rule-based filters. Were the structure filters evaluated?

- There is no mention of an IRB agreement for the study, or whether any steps were taken to ensure patient privacy during the work. Based on the sample report provided, it seems that personal information was either not present in the documents, or removed. Was de-identification of the documents performed? Automatically? It is especially important to comment on that, given the size of the corpus.

Suggested References:


**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.

**Declaration of competing interests:**
I declare that I have no competing interests