Author's response to reviews

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

Authors:

Martin F Toepfer (martin.toepfer@uni-wuerzburg.de)
Hamo Corovic (hamocorovic@googlemail.com)
Georg Fette (georg.fette@uni-wuerzburg.de)
Peter Klügl (pkluegl@uni-wuerzburg.de)
Stefan Störk (Stoerk_S@klinik.uni-wuerzburg.de)
Frank Puppe (frank.puppe@uni-wuerzburg.de)

Version: 3
Date: 31 July 2015

Author's response to reviews: see over
We thank the reviewers for their valuable comments and answer to their questions below.

**Reviewer : J. Garvin**

**G-Q1:** modules are often shared with other NLP developers. Please describe if you will share the terminology with others and how the NLP system could be used and potentially improved by others.

**G-A1:** In order to give insight into the terminology, we have added a supplementary HTML-file with the complete terminology, which should be published as supplement to the paper.

We aim to release the successor of the information extraction workbench under an open-source license. In the meantime, the sources are available on request from the first author.

**G-Q2:** please describe whether or not there is terminology coverage of these terms in SNOMED-CT. If there is, why is this new terminology needed?

**G-A2:** In the background section and in Section 2.5 of the revised manuscript, we explain our decision to develop a new terminology and to map it to Voelker et al.'s work, which is in our opinion the best available conceptualization for German heart echo reports. Germany is not a member of IHTSDO; SNOMED-CT is only available with exemptions.

**G-Q3:** are there plans to map these new terms to SNOMED-CT or to request development of new content so that the new terms can be incorporated into SNOMED-CT?

**G-A3:** Corresponding to the answer given above (G-A2), we do not have plans to address this issue at the moment.

**Reviewer : A. Neveol**

**N-Q2:** [...] would be good [to] have more discussion of other work in clinical NLP in German [...] 

**N-A2:** We address this point in the revised version of the paper and discuss the suggested references as well as additional work.

**N-Q3:** [...] decision to develop a new terminology is not explained [...] 

**N-A3:** We have revised the background section in order to underline why a new terminology has been developed (second paragraph on page 3) and why the above mentioned conceptualization for German heart echo reports from Voelker et al. could not be used directly (page 4, line 9-13).

**N-Q4:** [...] terminology development process is not clear [...] 

**N-A4:** Most of these questions (annotator agreement, etc.) are discussed in Section 3.1 of the manuscript. The revised work contains more details, e.g., about the terminology development process. The terminology has primarily been curated by H. Corovic.

**N-Q5:** How was the gold standard for evaluating the information extraction tool produced?

**N-A5:** After the terminology had been finalized, H. Corovic also annotated test documents. As mentioned in Section 3.1, test documents have also been annotated by another annotator;
agreement has been assessed; two other investigators inspected the manual annotations und decided, that the first annotation was good (e.g. 95.5% agreement on standard layout). Concerning the remaining differences, the investigators decided to prefer the annotation of the first annotator as the reference standard, which happened to disagree in more cases with the system’s results.

**N-Q6: The information extraction process relies on the detection of document structure using rule-based filters. Were the structure filters evaluated?**

N-A6: There is a misconception: the information extraction process is generic and does not rely on structure filters. We only used the knowledge about different document structures for the evaluation. The evaluation results showed that the rare structures still show a high precision, but a lower recall. In future work, we could use the knowledge about different structures for building different information extraction variants, but we did not in this paper.

**N-Q7: 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.**

N-A7: The first version of the manuscript mentioned this aspect in Figure 4. It has not been explained in detail since the de-identification process does not constitute a part of the contribution of this work. The revised manuscript mentions patient privacy aspects in Section 1 as well as Figures 2 and 4. The de-identification concept of the clinical data warehouse was approved by the privacy protection committee of the hospital.

**Summary of changes:**

- revised background section (Section 1), especially:
  - motivate development of a new terminology more clearly, and why Voelker et al.’s guideline is not directly applicable
  - additional and improved related work, especially on German clinical NLP and information extraction
  - mention patient privacy
- Section 2.5
  - Explain preference to German Cardiac Society’s echocardiography report documentation guideline over SNOMED CT
- Section 3.1
  - more detailed description of the terminology development process, e.g., number of terminology curators
- Section 3.3
  - remarks on limitations of our approach and the study
- Figures 2, 4: patient privacy (de-identification)
- Fixed typo in Figure 5
- additional acknowledgement (regarding payment of publication fees)
- modified Table 1 (IE performance comparison), main changes: reflect more related work
- Table 6: changed column header (attribute -> aspect)
• additional minor changes across the manuscript