Reviewer's report

Title: Automated identification of pneumonia in chest radiograph reports in critically ill patients

Version: 1 Date: 29 April 2013

Reviewer: Wendy Chapman

Reviewer's report:

This is a well-written, if a bit brief, manuscript on an NLP application for identifying radiology reports describing pneumonia. I think with a few additions and clarifications, this manuscript will make a nice contribution to the literature.

Major compulsory revisions

Need to cite and describe more related work: The authors cite several similar studies, which is important. I would suggest enriching the citation of these studies with a paragraph describing the methods and results that others have used in the past so that readers get a better feel for the context in which this article comes. Other work that is relevant that the authors do not cite is literature on hedge detection in text, especially in clinical/radiology reports. The Bioscope corpus is particularly relevant. Because uncertainty identification is one of the listed strengths of the manuscript, I think it would be important for the authors to help readers understand the state of the art in this area and know how the uncertainty identification in this application relates (or doesn’t) to other work.

Need to expand the evaluation: The authors rightly try to justify publication of another paper on extracting pneumonia from chest radiology reports by differentiating their manuscript from previous work. The two points of difference they focus on are the ICU setting and their modeling of uncertainty. There are several places in the manuscript where the evaluation of uncertainty is noted as making this paper unique. I think the modeling of uncertainty makes the manuscript a novel contribution to the literature. However, I do not think that the evaluation reflects the focus on uncertainty since it only evaluates the ability to identify positive reports and negative reports, seeming to leave out the 35% of reports that were classified as uncertain. I suggest also reporting statistics on identification of reports with the classification of possible. This type of evaluation also makes sense for practical reasons—one use case of a pneumonia classifier would be to assign reports to three bins for chart review: positive (does not need human review), negative (does not need human review), and possible (needs human review). A paper was recently published on this topic (S Dublin et al. Natural Language Processing to identify pneumonia from radiology reports. Pharmacoepidemiol Drug Saf, 2013 Apr 1. doi: 10.1002/pds.3418). Table 3 could show results for each bin, the numbers of reports in each bin, and the overall performance of the system (across all three bins on validation set). It could also be interesting to look at performance if uncertain classifications were lumped with
negative or were lumped with positive – depending on the application, a user may want to have a more sensitive or a more specific classifier, and having the uncertain classification would allow them to do that, so it would be interesting to see what the performance would be of each. In addition to a discussion of evaluation of the possible category, I also suggest a more extensive discussion of both the processing for uncertain classification (the lexicon – especially the 27 uncertainty profiles--, the patterns that the algorithm is looking for) and the performance (errors, difficulties, etc.). I think the supplementary table needs to be included. A more thorough investigation of uncertainty in the reports would increase the novelty and impact of this manuscript.

Need better description of the system: The authors also note the limitation of using a proprietary system. I agree this is a limitation in several ways. First, the system is not available for others to test; however, many of the systems the authors cite are also closed or unavailable, so this is not a big issue. A large issue to me is that the algorithm is not described at all in the manuscript. It should be possible at least to give a general description of how the system uses the special lexicon created for the pneumonia task to identify pneumonia-related findings. Without a description of the algorithm, it seems especially important for the authors to focus on the lexicon and the queries. I would appreciate a better description of how the three elements are linked to each other. An interesting experiment would be to use the lexicon with an existing open-source algorithm like pyConText followed by the same queries. I do not think this is necessary for publication but would be interested in the outcome (https://code.google.com/p/negex/source/browse/trunk/pyConText/src/pyConText/pycontext.py?r=62).

Minor essential revisions

There were a few points in the manuscript that were not clear to me:

- In the section on I2E queries, the authors discuss classification of ‘other’, ‘non-pneumonia’, ‘low uncertainty pneumonia’, and ‘high uncertainty pneumonia’. It was not clear how these classifications map to the final outcome classifications. In the discussion of uncertainty, I did not see a differentiation of low and high uncertainty.

- Physician interpretation - please add a bit more detail about the physician interpretation - did 2 physicians read and classify each report independently, or did they come to consensus (results section says by consensus but it isn’t clear whether that means they discussed disagreements and came to consensus or read and classified them together)? Was each report classified by 2 physicians or by 1? If reports were classified by 2 physicians independently before coming to consensus, the authors could report agreement statistics.

- It wasn’t clear to me what the “other” category represents

- It would be helpful to show confidence intervals on the outcome measures

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:

No competing interests