Reviewer's report

Title: Improved de-identification of physician notes through integrative modeling of both identifying and non-identifying medical text

Version: 1 Date: 20 December 2012

Reviewer: Joel Martin

Reviewer's report:

General comments
--------------------------

The authors present a new system to address de-identification of medical records and apply it to the I2B2 de-identification dataset. They achieve somewhat competitive results by using a novel source of information: the frequency of words that appear in non PHI sources.

1. Is the question posed by the authors well defined?
The question is well-defined with a clear measure of success.

2. Are the methods appropriate and well described?
The methods are appropriate and fully replicable.

3. Are the data sound?
The experimental data appear to be sound. They are reported differently than the original I2B2 experiments. It is time consuming for the reader to compare.

I believe it is the authors’ responsibility to directly compare with the past results on the data. They make a brief mention early in the paper, but there are no quantitative comparisons. This makes it difficult to know the value of their novel feature or the practical value of their system compared to those that already exist.

4. Does the manuscript adhere to the relevant standards for reporting and data deposition?
Yes.

5. Are the discussion and conclusions well balanced and adequately supported by the data?
In a local sense, the discussion and conclusions are balanced and supported by the data. However, after reading the paper, I am not clear about the true value of frequency as compared to other features. For example, Szarvas et al. (J Am Med Inform Assoc 2007;14:574–80) successfully used frequency from the training set for de-identification. That system appears to have performed better than the current one. What is the incremental advantage, if any?
I am also concerned that some of the advantage of word frequency came because the dataset created artificial words that permuted the order of syllables to create terms like: Girresnet, Diedreo A. Since those new constructions would be highly unlikely in external sources, frequency gives an unnatural advantage. Will the feature be as useful for realistic datasets?

Overall, I am concerned that the paper does not adequately defend the value of the novel feature. They do show that two features (including this one) are sufficient for very high performance. That is valuable and very interesting. However, they do not exclude some obvious uninteresting explanations of that performance, such as the artificial names.

Also, I am not sure what to do with the fact that other deidentification systems perform better than Scrubber on the same dataset. This is especially so because the authors did not provide any easy way to compare their system to the earlier ones. Since I did do the comparison, I know that I would use one of the other systems rather than Scrubber in a practical application.

Maybe Scrubber is better/interesting because it is simpler, has fewer features, or does better on some alternative measure of success. Maybe, but the paper does not make this clear.

6. Are limitations of the work clearly stated?
I mention what I consider to be significant limitations of the empirical comparisons above. The authors do not discuss these.

The authors do mentions some other limitations.

7. Do the authors clearly acknowledge any work upon which they are building, both published and unpublished?
Yes.

8. Do the title and abstract accurately convey what has been found?
The title and abstract are accurate.

9. Is the writing acceptable?
The writing is clear.

Major compulsory revisions:
i. show a quantitative comparative evaluation against previously reported systems that do better than Scrubber. Just quote the previously reported results.
ii. explain clearly why Scrubber or the external frequency feature is still an important contribution
iii. explain the interaction between the special words in the dataset and the use of frequency
iv. explain why Scrubber is an advance over, rather than just different from, other systems that use frequency within the dataset
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.