Reviewer's report

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

Version: 1 Date: 24 December 2012

Reviewer: Oscar Ferrandez

Reviewer's report:

The paper describes an interesting approach that tackles the de-identification problem from a perspective mainly based on the probability distributions of non-PHI/PHI words. A decision tree classifier is built for this purpose, using public medical texts for learning non-PHI words and private physician notes for PHI terms.

- Major Compulsory Revisions

1-PHI types

Authors refer to 8 PHI types (table 1). HIPAA specifies 18 different PHI identifiers, are these 18 HIPAA identifiers covered by the 8 types in table 1? Could the authors provide a mapping between types?

2-Features

I would like to see how a bi- or tri-gram language model would affect the results. Authors commented on calculating term frequencies using two or more words can lead to exponential time complexity; however I do believe there are efficient ways to process this information with reasonable computational cost.

As the feature set is not very big, and the authors are using Weka, apart from the results shown in table 3 for each kind of features, I would like to know the meaningfulness of each feature, for instance, this could be obtained processing information gain, which should be very easy using the Weka framework.

3-i2b2 corpora

Please note that as said in your reference [13], the i2b2 corpora was resynthesized with surrogates that, in most cases, could not be found in dictionaries, making their de-identification difficult for techniques based on dictionaries ("For patients, doctors, locations, and hospitals, we created surrogates by permuting the syllables of existing names from dictionaries such as the U.S. Census Bureau names database […] the generation of surrogates that could be found in dictionaries and did not make any effort to eliminate such surrogates from the data. Still, most of the generated surrogates could not be found in dictionaries, e.g., Valtawnprinceel Community Memorial Hospital and Girresnet, Diedreo A."). This fact makes the term frequency for these identifiers very low and the approach proposed here could be favored by this fact. I would like the authors to address this issue.
4-Patterned features
Is the approach using the regular expression set provided in [15]? Please clarify this when the feature set is described, also I would appreciate few comments about the usefulness of the regexs integrated into the approach and if some tweaking of these regex was performed.

5-Results
In the summary-results section, authors state that the method obtains 99.4% sensitivity for patient name. Where is this number in table 3? With which set of features was this number obtained?
Moreover, I would also like to have the data about specificity for the 9-way classification, i.e. for each PHI type. Conclusions could be made in a more rigorous manner having this information available.

6-Discussion
I want the authors to answer this question for me: in the big picture, why should I use your de-identification approach? What are the advantages of your approach with regard of other de-identification systems? I like the method and idea regarding non-PHI frequency terms extracted from public medical texts, however your approach does not obtain better results than the 2006 i2b2 challenge participants, and it seems that if I train a machine learning classifier or use other available de-identification systems (e.g., MIST) I would get better results (at least on the i2b2 corpora).

Another interesting issue I would like the authors to address is the portability of your models to other work types (i.e., other clinical text apart from discharge summaries). The authors could consider the possibility of using some data from the iDASH repository (http://idash.ucsd.edu/idash-data-collections), such as the MTSamples. At the very least, I would suggest making reasonable comments on the efforts that should be done to adapt the approach to other work types, and some comments on the expected performance/applicability.

- Minor Essential Revisions
- Discretionary Revisions

-Results
1-As the authors used the i2b2 corpus, could they say something about their approach and the top three systems participating in the i2b2 challenge? Maybe, some comments on the differences between your results and the top three i2b2 participants, reasons for such differences. (please take this into the context that your approach was evaluated after the challenge took place).

**Level of interest:** An article of importance in its field

**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