Author's response to reviews

Title: Mining Adverse Drug Reactions from Online Healthcare Forums using Hidden Markov Model

Authors:

Hariprasad Sampathkumar (hariprsd@ku.edu)
Xue-wen Chen (xwchen@wayne.edu)
Bo Luo (bluo@ku.edu)

Version: 3 Date: 12 April 2014

Author's response to reviews: see over
From: Hariprasad Sampathkumar, Xue-wen Chen and Bo Luo

To: Editor

BMC Medical Informatics and Decision Making

RE: Revision of manuscript for 'Research' article (MS: 1098530775116882) titled “Mining Adverse Drug Reactions from Online Healthcare Forums using Hidden Markov Model”

We would like to thank the Associate Editor and the Reviewers for reviewing our manuscript and providing valuable and insightful comments. We have revised our paper in response to address all the comments made by the reviewers. Through this revision we have included additional data from baseline experiments, provided more explanation to clarify any unclear information and in general improved the presentation of the paper. We believe the technical presentation and readability of the paper have significantly improved over the previous version. We provide detailed responses to the each of the individual Reviewer's comments as a follow up.

We thank you again for giving us an opportunity to contribute a part of the BMC Medical Informatics and Decision Making Journal and look forward to hearing from you regarding the review process.

Sincerely,

Hariprasad Sampathkumar
Xue-wen Chen
Bo Luo
Detailed Responses to Reviewer's comments

Comments from the Associate Editor:

Please find attached comments from peer reviewers. We would be grateful if you could address the comments in a revised manuscript and provide a cover letter giving a point-by-point response to the concerns.

Authors' Response: We would like to thank the Associate Editor for coordinating the review process and giving us an opportunity to revise our manuscript. In the revised manuscript, we have improved the paper in responding to the Reviewers' comments. We have included additional data from baseline experiments, provided more explanation to clarify any unclear information and in general improved the presentation of the paper. In the following sections we explain in detail how we have addressed the Reviewers' comments as a part of this revision.

Comments from Reviewer 1:
Reviewer: Hua Xu

Reviewer's report:
This paper proposes an interesting approach that uses hidden markov model to mine adverse drug reactions from online healthcare forums. The topic is important and the evaluation showed promising results. However, there is a lack of details about the proposed method and this reviewer would suggest following revisions:

Major Compulsory Revisions:

1. Please add a baseline method into the study and provide its performance, so that readers can understand the advance of the proposed method, when compared with existing ones. Possible baseline methods could be co-occurrence statistics methods used in Wu et al. [30] or the sliding window approach used in Leaman et al. [28].

Authors' Response: We would like to thank Reviewer 1 for pointing out an important part of the evaluation that we had missed! As suggested by the Reviewer (thanks again!) we have chosen to use a baseline method involving co-occurrence statistics to help compare and contrast the performance of our proposed approach. We have now revised the manuscript to include the following: we include a description of the baseline approach under the Methods section and present its results as a part of the Results and Discussion section.

Methods

Baseline Classifier
In order to compare the performance of the HMM based classifier, a Baseline classifier based on the co-occurrence frequencies of drug names and side-effects was built. Co-occurrence statistics are a very common measure for identifying associations and relationships between words [53]. For all the 760 drug names and 1390 side-effects in the dictionary, a co-occurrence frequency map was constructed based on their occurrence in a forum message. Within a forum message, even though a drug name and a side-effect occur more than one time, their co-occurrence count was still considered to be one, in order
to prevent a single forum post from influencing the co-occurrence frequencies. Similar to the HMM based classifier, the Baseline Classifier is also given the same training data set, from which the drug/side-effect co-occurrence statistics are computed. Given a test set, the Baseline classifier extracts every unique drug/side-effect pair in the given message and flags them to have a positive relationship in case their co-occurrence frequency was computed to be greater than zero. The same metrics used in the evaluation of the HMM-based classifier are also used in case of the Baseline classifier.

Results and Discussions

10-fold Cross-validation on the manually annotated dataset
In order to compare the performance of the classifiers we do a 10-fold cross-validation on the 2000 sample manually annotated training dataset. Table 2 presents the results of a single run of the 10-fold cross-validation for the HMM classifier, while Table 3 presents the corresponding results for the Baseline classifier. Tables 4 and 5 present the mean values of Precision, Recall and F-Score computed across 10 different runs of the 10-fold cross-validation for the HMM and the Baseline classifiers, respectively. In general, it can be seen that the HMM-based classifier performed better with a mean F-Score of 0.76 in comparison to the Baseline classifier which yielded a mean F-Score of 0.575. Figure 10 presents a plot of the F-Score values of the Baseline Classifier against the HMM Classifier across the 10 different runs.

Table 3 Results of single run of a 10-fold cross-validation on the manually annotated data set for Baseline Classifier

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Train set</th>
<th>Test Set</th>
<th>True Positive</th>
<th>False Positive</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Score</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1761</td>
<td>196</td>
<td>27</td>
<td>8</td>
<td>0.771</td>
<td>0.569</td>
<td>0.614</td>
<td>0.827</td>
</tr>
<tr>
<td>2</td>
<td>1761</td>
<td>196</td>
<td>25</td>
<td>15</td>
<td>0.625</td>
<td>0.568</td>
<td>0.595</td>
<td>0.827</td>
</tr>
<tr>
<td>3</td>
<td>1761</td>
<td>196</td>
<td>35</td>
<td>14</td>
<td>0.714</td>
<td>0.614</td>
<td>0.660</td>
<td>0.816</td>
</tr>
<tr>
<td>4</td>
<td>1761</td>
<td>196</td>
<td>27</td>
<td>14</td>
<td>0.659</td>
<td>0.519</td>
<td>0.581</td>
<td>0.801</td>
</tr>
<tr>
<td>5</td>
<td>1761</td>
<td>196</td>
<td>28</td>
<td>9</td>
<td>0.757</td>
<td>0.491</td>
<td>0.596</td>
<td>0.806</td>
</tr>
<tr>
<td>6</td>
<td>1761</td>
<td>196</td>
<td>27</td>
<td>21</td>
<td>0.563</td>
<td>0.614</td>
<td>0.587</td>
<td>0.806</td>
</tr>
<tr>
<td>7</td>
<td>1761</td>
<td>196</td>
<td>23</td>
<td>17</td>
<td>0.575</td>
<td>0.548</td>
<td>0.561</td>
<td>0.816</td>
</tr>
<tr>
<td>8</td>
<td>1761</td>
<td>196</td>
<td>27</td>
<td>17</td>
<td>0.614</td>
<td>0.529</td>
<td>0.568</td>
<td>0.791</td>
</tr>
<tr>
<td>9</td>
<td>1761</td>
<td>196</td>
<td>18</td>
<td>12</td>
<td>0.6</td>
<td>0.409</td>
<td>0.486</td>
<td>0.806</td>
</tr>
<tr>
<td>10</td>
<td>1763</td>
<td>194</td>
<td>28</td>
<td>10</td>
<td>0.737</td>
<td>0.5</td>
<td>0.596</td>
<td>0.804</td>
</tr>
</tbody>
</table>

Table 5 Mean Precision, Recall and F-Score values across 10 runs of 10-fold cross-validation for Baseline

<table>
<thead>
<tr>
<th>Run</th>
<th>Mean Precision</th>
<th>Mean Recall</th>
<th>Mean F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.648</td>
<td>0.519</td>
<td>0.575</td>
</tr>
<tr>
<td>2</td>
<td>0.648</td>
<td>0.527</td>
<td>0.577</td>
</tr>
<tr>
<td>3</td>
<td>0.652</td>
<td>0.519</td>
<td>0.576</td>
</tr>
<tr>
<td>4</td>
<td>0.654</td>
<td>0.523</td>
<td>0.579</td>
</tr>
<tr>
<td>5</td>
<td>0.639</td>
<td>0.522</td>
<td>0.572</td>
</tr>
<tr>
<td>6</td>
<td>0.640</td>
<td>0.525</td>
<td>0.575</td>
</tr>
<tr>
<td>7</td>
<td>0.643</td>
<td>0.513</td>
<td>0.569</td>
</tr>
<tr>
<td>8</td>
<td>0.646</td>
<td>0.519</td>
<td>0.573</td>
</tr>
<tr>
<td>9</td>
<td>0.651</td>
<td>0.523</td>
<td>0.577</td>
</tr>
<tr>
<td>10</td>
<td>0.652</td>
<td>0.522</td>
<td>0.578</td>
</tr>
</tbody>
</table>
2. This reviewer is not clear how exactly the HMM was used in the relation prediction. For example, what is the unit of the sequence, a sentence or entire message? Please add an example (e.g., showing a message with drugs and ADRs).

Authors' Response: We apologize for not clearly stating how the HMM is used for relation prediction. To address this we have enhanced the Information Extraction Module subsection under the Methods section to address this:

If a message contains only a drug name and a side-effect mention, it is not sufficient to denote a positive ADR. There needs to be some form of causal relationship that clearly associates the drug with the side-effect. It is in this regard that the keywords identified by the HMM are used to capture the causal relationship. As a part of the training, the HMM is trained on positive samples where it learns the association between the drugs and side-effects through the presence of keywords and uses this information for relationship prediction on the test data set.

Further in response to the Reviewer's question on whether the unit of sequence is a sentence or entire message, we would like to clarify that it is the entire message. To help alleviate any further confusions we have also enhanced Figure 7 to clearly show how an ADR is extracted from a sample forum message as it traverses through the Text Processing and Information Extraction modules. Below we have included the enhanced figure:
Minor Essential Revisions:

3. The NER module for drug and ADR is weak. The drug and ADR dictionaries seem small and the performance of NER was not reported. Authors need to cite literature on drug name recognition in clinical text (e.g., papers published in JAMIA for the i2b2 2009 clinical NLP challenge) or in biomedical literature (the 2013 BioCreative CHEMNER challenge) and state reasons why comprehensive drug dictionaries were not used (e.g., may introduce false positives when the lexicon file becomes large). Or they can rebuild dictionaries by leveraging drug resources mentioned in those studies.

Authors' Response: Thanks to the Reviewer for pointing out this weakness and helping us present our
proposed solution in a better light. We have updated the text under the Named Entity Recognition subsection to include the following explanation for the smaller dictionary sizes.

When using the lexicon-based methods for performing NER, the choice of vocabulary that is used to create the dictionary entries has a significant impact on the performance of the NER module. So it becomes necessary that the vocabulary of the dictionary reflect the vocabulary of the target corpus to be mined. In this regard with the messages in the online forums being the target corpus in our approach, we would need to construct a dictionary that would reflect the vocabulary observed in such forum messages. Majority of the users of such online forums, do not possess the medical background to use technical terms to identify the drug names and side-effects. The comprehensive drug dictionaries, such as those used as a part of the BioCreative and JAMIA challenges, tend to use technical terms to identify the mentions of drug names, which do not form the vocabulary of the average forum user. So including comprehensive drug name dictionaries may not necessarily improve the performance of the NER module. It is in this regard we construct a custom drug dictionary with only minimal entries that would reflect the vocabulary of an average user.

Thanks to the Reviewer for suggesting additional papers regarding drug name recognition in the literature, we have included these as a part of our citations.


Comments from Reviewer 2:

Reviewer: Samir Abdelrahman

Reviewer's report:

Major Compulsory Revisions:

The article presents a very good well-tested work (with reported acceptable statistical assessments)--Thanks for the authors. However, I do ask the authors to add the following parts:

1- Baseline experiments: I urge the authors to choose any sequence labeling technique, maybe from the literature, and test the experiments using it. It is very useful to access HMM versus the others.

Authors' Response: We would like to thank the Reviewer for the favorable comments and also for pointing out an important part of the evaluation that we had missed! As presented in response to Reviewer 1’s comments, we have chosen to use a baseline method involving co-occurrence statistics to help compare and contrast the performance of our proposed approach. As mentioned above we have revised the manuscript to include the results and discussion on the baseline approach. We hope this would suitably address the comments raised by Reviewer 2.
2- Component-based experiments: it is very interesting for the community to see the affect of each system component in HMM modeling. For example, what if the filtered or extracted plain text module is removed; also, if any of NE dictionary sizes is decreased. Both of the above parts may be augmented in the result tables.

Authors' Response: Those are interesting questions! The plain text extraction sub-module and the text filtering sub-modules are key elements of the overall Text Processing pipeline. They are necessary for removing noise from the forum messages present in the form of HTML tags and for normalizing the words present in the messages into a form that makes them be suitable for further processing. For example, not having the plain text extraction sub-module would result in including HTML tags, entities and other attributes present as a part of the forum message's presentation markup to be considered as a part of the message itself. This would increase the noise in the data set and would affect the prediction accuracy of the Named Entity Recognition module and hence the Relationship Extraction module. Similarly, the filtering sub-modules act on the extracted plain text and normalize it by converting to lower case and removing punctuation and numeric values. Without this normalization, there would be several variations of a word (same word with different cases Eg: 'Lipitor', 'lipitor'; and same word ending with different punctuation, e.g. 'headache,' and 'headache.'). Such variations will in turn affect the prediction accuracy of the NER module and thereby the successful extraction of the Adverse Drug Reaction by the Relationship Extraction module.

By reducing the size of any of the NER dictionaries, we would limit the number of drugs or side-effects identified by the NER module which would in effect reduce the prediction accuracy of the HMM classifier by introducing more False Negatives. In general, we would want to keep the size of the dictionaries to match the vocabulary of the average forum user in order to achieve better coverage and good prediction accuracy.

3- Limitation subsection: I do believe that the authors should list all possible drawbacks in a subsection

Authors' Response: Thanks to the reviewer for pointing out another important subsection that we missed. To address this we have enhanced the 'Results and Discussion' section to include two new subsections that clearly highlight both the advantages and the limitations of our proposed approach. Below we include the text from these two new subsections:

Advantages

The main advantage of the proposed approach is the volume and timeliness of the discovered information. That is, the capability of collecting very large amount of up-to-date information at very low cost. With the source of the data being the online healthcare forums, this approach leverages all the benefits of ‘Big Data’. The online forums which act as a source of ‘Big Data’ are able to provide extremely high volume of raw data that can be used to extract information – discover adverse drug reactions in our approach. With its high volume and diversity, it is able to cover a large number of drugs which are usually not possible to cover in case of clinical trials. While collecting similar information through clinical trials can be very expensive, crawling of data from online forums is almost free, with most of the data being publicly available without any access restrictions when compared to other Literary or Clinical sources. Also with the users constantly providing feedback on the forums, we are able to provide the most up-to-date information on the side-effects of drugs.
Limitations

As with all the benefits leveraged from the ‘Big Data’ source, this approach also inherits some of its drawbacks. One of the major issue with user generated data from online healthcare forums is the amount of noise that could be present in such forum messages. Majority of the members of such forums are average users who don’t necessarily have any medical background, hence, they may provide inaccurate or exaggerated information when it comes to drug side-effects. Using such a source for mining of ADR data may potentially provide false positives. The size of data helps mitigating this problem – repeatedly reported side effects are more likely to be true positives. Also the reports might be biased, as users tend to not make forum posts when there are no side-effects observed on consumption of a drug. Therefore, we present the mined information as early indicators of potential ADRs, and these reports have to be further investigated through rigorous medical and clinical procedures by health authorities to confirm if the drugs involved indeed cause the reported adverse reactions.

Other updates:

In addition to the manuscript changes listed above we have also reduced the size and changed the orientation of Figure 9 that displays the initial state and state transition probabilities of the trained HMM in order to provide a better presentation.