Barriers to NLP Adoption

- We have a long history of research on NLP methods in the clinical domain [1].
- However, the complexity of unstructured clinical text makes analysis a hard problem and its accuracy varies.
- Domain experts may be able to fix problems with the models but they may not be familiar with symbolic and machine learning techniques.

Design Requirements

We have built upon ideas in Visualization, Interactive Machine Learning and Interface Design research.

Our design requirements are summarized as follows:

R1: The tool should make it easier for machine learning non-experts to work with NLP models.
R2: It should incorporate efficient mechanisms for annotation and labeling, and also encourage feedback that is consistent and informative.
R3: The interactive components should support the entire interactive machine learning loop - i.e. a review, feedback and retraining cycle.

Acknowledgments

Our demo uses an example dataset of colonoscopy reports and is based on the work done by Harkema et al. [2]. This research is supported by NIH grant 5R01LM010964.

References


Our Solution: An Interactive Tool for NLP on Clinical Text

Our goal is to close the NLP gap by providing clinical researchers with highly-useful tools that will facilitate the process of reviewing NLP output, identifying errors in model prediction, and providing feedback that can be used to retrain or extend models to make them more effective.

We have developed an interactive web-based tool that facilitates both the review of binary variables extracted from clinical records, and the provision of feedback that can be used to improve the accuracy of NLP models.

User Study and Results

We conducted a formative user study with five clinicians and clinical researchers as participants to gain insight into usability factors of the tool that may be associated with errors or confusion, and to identify opportunities for improvement via re-design or implementation of new functionality.

We used the System Usability Scale [4] consisting of 10 questions on a 5-point Likert scale to help get a global view of subjective assessments of usability. The average SUS score was 70.5 out of 100.

A summary of recommendations inferred from the user study for is given below:

<table>
<thead>
<tr>
<th>Category</th>
<th>Recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Workflow</td>
<td>1. Allow sorting (or filtering) of the documents in the grid based on the prediction probabilities. This would make it easier for the users to prioritize documents for review.</td>
</tr>
<tr>
<td>WordTree</td>
<td>1. Change the layout of the tool to show the WordTree view along with the document view. This would allow the user to quickly access the entire report without having to navigate through the grid when they don’t have their own strategy for selecting documents for review.</td>
</tr>
<tr>
<td>Feedback</td>
<td>1. Provide a feedback mechanism to specify that a text span does not indicate either of the classes. This would allow the user to remove non-informative but possibly misleading features in re-training.</td>
</tr>
<tr>
<td>Re-Training</td>
<td>1. Perform auto-retraining in the background when a sufficient number of feedback items have been provided by the user.</td>
</tr>
<tr>
<td></td>
<td>2. Provide a built-in mechanism to validate and generate a performance report for the current model against a held-out test set.</td>
</tr>
</tbody>
</table>

Future efforts will involve incorporating these recommendations and conducting an empirical evaluation.