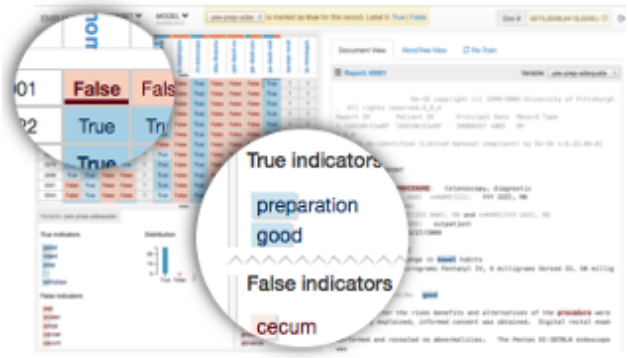
# Bridging the Natural Language Processing Gap: An Interactive Clinical Text Review Tool

**Gaurav Trivedi[1], Phuong Pham[1], Wendy Chapman, PhD[2], Rebecca Hwa, PhD[1], Janyce Wiebe, PhD[1], Harry Hochheiser, PhD[1]**

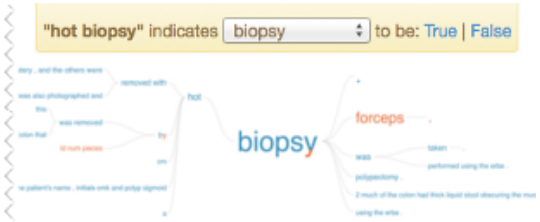
**[1]University of Pittsburgh, Pittsburgh, PA; [2]University of Utah, Salt Lake City, UT**

*Background:* Although Natural Language Processing (NLP) has been widely used in extracting information from clinical text, the gap between NLP techniques and the researchers who want to make use of this data remains large. NLP systems are typically developed by informaticists skilled in symbolic or machine learning techniques unfamiliar to clinical researchers interested in analyzing clinical records. Our goal is to close this gap by providing clinical researchers with highly-usable 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.

*Methods:* We have developed an interactive web-based tool that facilitates both the review of structured values extracted from clinical records, and the provision of feedback that can be used to improve the accuracy of NLP models. Our tool consists of three main views: (a) *The grid view* (Figure 1a) shows boolean variables extracted from the text in columns and individual documents in rows, providing an overview of NLP results; (b) *The WordTree*[1] *view* (Figure 1b) provides the ability to search for and explore word sequence patterns found across the documents in the corpus and provide feedback that will be used to retrain NLP models; (c) *The retrain view* (not shown) shows user-provided feedback, including any potential inconsistencies, and lists changes in variable assignments due to retraining. Our NLP pipeline uses a bag of words feature-set and a support vector machine (SVM) learning model, but it can be extended for use with different models and complement other existing tools as well.




(a) *Grid view* - showing the NLP results. Below it, there is a list of words indicating the predicted value, which are also highlighted in the document text on the right side.




(b) *WordTree view* - All sentences containing the word in the center are shown through a WordTree. Our tree design also shows the class split using color gradients. The top yellow bar is used for providing feedback.

**Figure 1.** Screenshots of two of the main views from the tool. Magnified portions are shown inside the circles.

*Results and Discussion:* This tool visualizes the NLP system's predictions for the values of the variables of interest to the researcher and allows the user to view the evidence for the prediction in the report, view the distribution of the values for each variable, and correct the prediction. The ability to retrain the NLP tool based on user feedback is coupled with feedback to the user about implications of retraining. Future work involves conducting usability and evaluation studies with clinical researchers.

---

[1] Wattenberg, M., Viegas, F.B., "The Word Tree, an Interactive Visual Concordance," *Visualization and Computer Graphics, IEEE Transactions on* , vol.14, no.6, pp.1221,1228, Nov.-Dec. 2008