Smoking Status Detection Across Domains

Michael Tepper, MA[1], Fei Xia, PhD[1,2], Meliha Yetisgen-Yildiz, PhD[2,1]
[1]Department of Linguistics, [2]Biomedical & Health Informatics, School of Medicine, University of Washington, Seattle, WA

Abstract

With the introduction of comprehensive electronic medical records, all aspects of patient history and care can now be captured in both structured and free-text format. As part of a quality improvement (QI) project conducted at University of Washington (UW), we work on automating the abstraction of various types of data elements including smoking status by using natural language processing and machine learning approaches. In this abstract, we present our preliminary results for smoking status detection.

Preliminary Experiments

To determine the feasibility of developing an automated classification approach to identify smoking status, we used physician notes from a cohort of 618 patients who had surgeries in UW Medical Center in 2010. The retrospective review of the reports was approved by the UW Human Subjects Committee of Institutional Review Board. We first analyzed the report types that included smoking-related concepts (e.g., cigarettes, smoker, tobacco, etc.) and identified three report types that frequently contained these concepts, which are Discharge Summaries, Pre-Anesthesia reports, and Pain Management reports. This yielded 334 patients and 847 reports. The UW corpus was randomly selected form this set (one report per patient), yielding 334 total reports. In addition to the UW corpus, we used the training set of the publically available i2b2 corpus annotated for smoking history[1], which contained 398 discharge summaries. We used 389 of the i2b2 discharge summaries, for reasons explained below.

For the UW corpus, we compared the output of our system to a smoking-status determination obtained through manual abstraction[a]. The abstraction used a two category system: *Current Smoker,* for anyone determined to have smoked within the past year, and *Non-Smoker*, for anyone else. The i2b2 corpus used five categories for smoking status. When comparing the output of our system on the i2b2 set against the gold standard, we mapped the i2b2 categories to the UW 2-category system, as follows: *Current Smoker* → *Current Smoker,* {*Past Smoker*, *Non-Smoker*, *Unknown*} → *Non-Smoker*. The i2b2 category *Smoker* means that the patient can be either a current smoker or a past smoker; because it is ambiguous when mapping to the two categories in the UW system, we ignored this category, which resulted in the removal of 9 documents from the i2b2 training set.

To train our system, we first looked for mentions of smoking-related concepts within each report in the training corpus. For each report with smoking mentions, we extracted windows of 15-words around each smoking mention. We represented the content of the combined mention-window text as feature vector, consisting of unigram words, as well as one binary feature to detect a *quit* predicate. We used these feature vectors to train a Support Vector Machine (SVM) classifier. To measure performance, we looked for mentions of smoking-related concepts within each report in the test corpus. Each report without smoking mentions was automatically labeled as *Non-Smoker.* For the remaining reports, the mention windows were extracted and submitted to the classifier. In our initial experiments, we trained on the i2b2 corpus and tested on the UW corpus. For the class *Current Smoker*, the performance was 0.77 precision, 0.69 recall, and 0.73 F1-score. For the class *Non-Smoker*, the performance was 0.94 precision, 0.96 recall, and 0.95 F1-score. The overall performance was 0.91 F1-score.

Conclusion & Future Work

The preliminary results demonstrate that (1) our initial set of features can achieve promising classification performance as measured with respect to manual annotation and (2) data from different institutions can be successfully used to train a system generalizable to other institutions and report types for capturing smoking history. As next steps, we will do a detailed error analysis, enhance the training by including data from the UW corpus, and conduct additional experiments. Since the UW smoking-status is a decision made per patient, we will extend the UW corpus to allow multiple reports per patient and try multi-document classification at the patient level.

Acknowledgements

This work was supported by Agency for Healthcare Research and Quality Grant Number 1 R01 HS 2002501.

References

1. Uzuner O, Golstein I, Luo Y, and Kohane I. Identifying Patient Smoking Status from Medical Discharge Records. J Am Med Inform. 1999; 15(1): 14-24.

[a] The comparison was done automatically. The authors did not have access to the annotations created as part of the QI project.