Technical Brief ■

Using Implicit Information to Identify Smoking Status in Smoke-blind Medical Discharge Summaries

RICHARD WICENTOWSKI, PHD, MATTHEW R. SYDES, MSC

Abstract As part of the 2006 i2b2 NLP Shared Task, we explored two methods for determining the smoking status of patients from their hospital discharge summaries when explicit smoking terms were present and when those same terms were removed. We developed a simple keyword-based classifier to determine smoking status from de-identified hospital discharge summaries. We then developed a Naïve Bayes classifier to determine smoking status from the same records after all smoking-related words had been manually removed (the smoke-blind dataset). The performance of the Naïve Bayes classifier was compared with the performance of three human annotators on a subset of the same training dataset (n = 54) and against the evaluation dataset (n = 104 records). The rule-based classifier was able to accurately extract smoking status from hospital discharge summaries when they contained explicit smoking words. On the smoke-blind dataset, where explicit smoking cues are not available, two Naïve Bayes systems performed less well than the rule-based classifier, but similarly to three expert human annotators.

■ **J Am Med Inform Assoc.** 2008;15:29–31. DOI 10.1197/jamia.M2440.

Introduction

Our study investigates two methods for identifying smoking status from hospital discharge summaries (medical records or records) as part of the 2006 i2b2 NLP Shared Task.[1] The first method uses simple rules to classify discharge summaries based on the presence of smoking-related keywords in the document. The second method uses a Naïve Bayes (NB) classifier trained on word bigrams to determine smoking status in discharge summaries that have no explicit smoking-related keywords in them (smoke-blind discharge summaries). We present results on this smoke-blind dataset and compare it to the performance of human experts on the same dataset.

An unabridged version of this manuscript is available online at the JAMIA website as a data supplement. Tables 1-8, Figures 1-3, and a discussion of the rule-based classifier appear in the online JAMIA supplement to this manuscript found at www.jamia.org.

The Revised Task

Due to our success with the rule-based classifier, we determined that a greater challenge was to identify smoking status in the absence of any explicit cues about smoking status. In the Methods section, we describe in detail the methodology that we developed and applied. Attempting a similar a task, Zeng et al.[2] decided "not to embed decision making logic in [their natural language processing] system: for example, inferring HIV [positive] status from [treatment with] AZT." They noted that "while such logic is very useful, we believe it should be developed and evaluated separately. . . [such] rules may be useful but ideally might be applied in a separate processing step." In this revised task, we consider the feasibility of inferring smoking status in the absence of explicit smoking cues.

Methods

Creating the Smoke-blind Dataset

To pursue a computational approach to determining smoking status in the absence of explicit evidence, we first needed to create a set of documents in which smoking information was not present (the smoke-blind dataset). To do this, we first removed from the training set all of the 252 discharge summaries that were labeled as unknown. The remaining 146 records were hand-edited (RW) to remove overt references to smoking.

The Smoke-blind Systems

To classify the smoke-blind data set, we chose to avoid a keyword-based system. Although we could hypothesize a list of keywords that might be present in the discharge summary of a smoker (e.g., hypertension, coronary artery disease, lung cancer), none of these phrases could be used to definitively classify a record. Rather, such phrases provide partial evidence of smoking. In this way, multiple phrases present in the same document can serve as accumulating evidence, increasing the confidence of a classification.

Another reason to avoid the keyword-based method is that some keywords more strongly indicate smoking (lung cancer) than others (hypertension). It is unlikely that a human expert could manually devise a complete list of possible keywords and weight each of these keywords appropriately.

For these two reasons, we wanted a classifier that could learn these phrases and weights from training data. We chose to use a NB classifier, trained on word bigrams found in the smoke-blind data. An NB classifier chooses the label

Affiliations of the authors: Swarthmore College (RW), Swarthmore, PA; MRC Clinical Trials Unit (MRS), London, England.

Correspondence: Richard Wicentowski, Swarthmore College, Computer Science Department, 500 College Avenue, Swarthmore, PA 19081; e-mail: <mailto:richardw@cs.swarthmore.edu>.

Received for review: 3/16/07; accepted for publication: 10/3/07.

that maximizes the similarity between a record, R, and a class label, C_j, where the similarity is defined as: $\mathrm{Sim}(R,C_j) = \mathrm{P}(R,C_j) = \mathrm{P}(C_j)\mathrm{P}(R|C_j)$

The *a priori* probability of the class labels C_j was assumed to be uniform because we were not expecting the evaluation data to have the same underlying distribution as the training data. The conditional probability $\mathrm{P}(R \mid C_j)$ was based on a bigram language model using modified Kneser-Ney discounting.[4,5]

The classifier was used to build two systems. The first system (NB System 1) was trained on the smoke-blind dataset with labels provided as part of the shared task training set. This training set included 80 smoking and 66 nonsmoking records.

The second system (NB System 2) used an expanded training set by supplementing the smoke-blind dataset with the 43 additional records that were part of the shared task's official test set. We automatically labeled these additional records using our rule-based classifier, knowing that this should be very accurate in providing the true answer, and then made these additional records smoke-blind using the previously described procedure (RW). The combination of these additional records and the original smoke-blind dataset formed a larger training set for this second system with 104 smoking and 83 nonsmoking records.

NB1 System 1 and NB System 2 were evaluated using leave-one-out cross-validation; leave-one-out cross-validation maximizes the size of the training set records while ensuring that the system is not trained on the individual record that is being classified.

These classifiers were trained and evaluated using coarse-grained labels only, folding Past Smoker and Current Smoker into the existing label Smoker. This was necessary because, after removing all evidence of smoking from the patient summaries, it would have been extremely difficult (if not impossible) to recover the temporal information needed to distinguish between a current and a past smoker.

Expert Annotation

Because all explicit smoking cues were removed, it was possible that this smoke-blind dataset would not contain enough information, even for human experts, to confidently predict the label of many records. Therefore, to test the effectiveness of the NB method trained and evaluated on the smoke-blind data, we recruited three human annotators with expert medical knowledge: a statistician experienced in oncology clinical trials (A1), an oncology certified nurse (A2), and an oncology research fellow (A3).

We expected the annotation to be time consuming, so we provided the annotators with only a subset of the 146 smoke-blind summaries: a total of 54 summaries, composed of 34 smokers and 20 nonsmokers.

These three annotators were asked to make educated guesses about smoking status based on their knowledge of health and medicine and their common sense. We provided guidelines worded closely to those used by the task organizers, noting that all direct evidence of tobacco smoking status had been removed and that absence of information about smoking status was not an indication of a nonsmoker. As was done with the NB task, these annotators were asked to provide only coarse-grained smoking status: Smoker, Nonsmoker, and Unknown, omitting Current Smoker and Past Smoker. It is important to remember that the smoke-blind dataset excluded all summaries labeled as Unknown by the shared task organizers. Therefore, the annotators were not attempting to predict when a record had an Unknown label attached to it; rather, annotators were allowed to provide the label Unknown when they could not determine the smoking status of the patient described in the discharge summary.

We evaluated the performance of each annotator individually, and we obtained a combined answer (Â) by taking a simple plurality of the three annotators' assessments. We considered Unknown a nonvote, and returned the label Missing when there was no plurality, or when all three annotators chose Unknown (Figure 2).

Analysis

We assessed the performance of the rule-based system, both NB systems, and our human annotators using standard methodology from the fields of natural language processing and medical statistics to calculate recall (sensitivity), precision (positive predictive value), specificity, and F-measure.

We submitted the maximum-permitted three entries to the i2b2 Shared Task.[1,6] The first entry labeled the test dataset of 104 records using the rule-based classifier. The second entry used NB System 1, and the third entry used NB System 2. The performance results of these entries are discussed below in the Results section.

Results

The Revised Task

Table 4 shows the performance of the plurality result (Â) of human experts using coarse-grained labels on the 54 smoke-blind discharge summaries that they annotated. Precision in classifying Smokers was 92%, but for Nonsmokers, precision was 46%.

Table 5 shows the performance of the individual human experts and their plurality result on the same 54 smoke-blind summaries. Overall, the Â classifications achieved 77% precision at 56% recall. Table 6 shows the performance of the NB classifier at levels of recall that match the human annotators.

Figure 3 shows a plot of precision against recall for the NB classifiers (NB System 1 and NB System 2) with the standard and extended training sets. This is shown for the 54 smoke-blind records that were also assessed by the human annotators. The results of the two NB classifiers are broadly similar to each other and to the humans, as shown in the graph. By eliminating low-confidence guesses, any classifier can achieve higher precision but at the expense of lower recall.

On the evaluation data, Table 7 shows confusion matrices for the two NB systems, and Table 8 shows the per-label and overall system performance for both systems.

Discussion

The NB Classifier and the Smoke-blind Dataset

We investigated approaches to extracting smoking status when the smoking terms used in the rule-based method were removed from the hospital discharge summaries. We found that a simple NB approach yielded reasonable levels of accuracy within the constraints of this task, even given a training set of limited size.

It is natural to ask how well humans could determine smoking status from such smoke-blind records. The human annotations are important because they serve as a plausible upper limit for the performance we should expect with a statistical model. We used a purposive sample of three annotators with expert health/medical knowledge to provide a comparison. With our small set of annotators, we have shown that the simple NB approach provides results not too dissimilar from expert human annotators, both individually and combined ($\hat{A}$).

We had originally intended to develop a computational approach that used medical keywords, as also suggested by Zeng et al.,[2] to identify the patients as either Smokers or Nonsmokers in the smoke-blind hospital discharge summaries. To this end, we had asked our annotators to note verbatim the keyword cues that they had used to ascertain smoking status. The rationale for such an approach is that there are a number of diseases or conditions for which smoking is a recognized risk factor and that are more prevalent among smokers than nonsmokers, e.g., emphysema and lung cancer. Similarly, there are social habits that may be expected to correlate reasonably with smoking e.g., regularly drinking alcohol or smoking substances other than tobacco. In theory, one could derive a list of such keywords and base a probability of a given patient smoking on the presence of these keywords. However, we found this was not practicable, at least in this context, for a number of reasons.

First, the list of potential keywords is not exhaustive and the training set was unlikely to be representative of all future medical records; furthermore, there may be as yet unknown or unrecognized conditions that predict smoking well. Indeed, the medical literature is not entirely clear on for what, exactly, smoking is a risk factor. This would lead to underprediction of Smoking.

Second, smoking may be a risk factor for a given condition, but it may not be the main risk factor, i.e., there are fairly prevalent conditions where smokers have a higher risk but where many nonsmokers also have the condition. For example, smokers have a higher risk of having a stroke (cerebrovascular accident, CVA) but nonsmokers also experience CVAs. Using CVA as a keyword trigger for predicting smoking status would lead to false-positive predictions of Smoking.

Thirdly, although developing a list of keyword cues that positively indicate smoking may be potentially feasible, it is unclear whether or not we could develop a sufficient list of keywords that contraindicate smoking (or that predict nonsmoking), especially in the context of hospital records. We note that the best annotator at predicting nonsmoking (A2) did take the most sophisticated approach to this. In a postannotation interview, A2 stated that classification of Nonsmoker often came from social cues, e.g., obese people, very elderly people, and pregnant women were seen as less likely to smoke.

Finally, it would be difficult to distinguish between Current Smokers and Past Smokers using this method. Although it is thought that the risk of some adverse health conditions decreases when smoking is stopped, the risk may persist for other conditions.

Conclusions

A simple rule-based classifier can be used to accurately extract smoking status from hospital discharge summaries when they contain explicit smoking words. A simple NB model trained on word bigrams performs less well when these smoking cues are not available, but similarly well to expert human annotators.

References ■

Note: Reference 3 is cited in the online data supplement to this article at jamia.org.

1. Uzuner Ö, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records. J Am Med Inform Assoc 2008;15:14–24.
2. Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazaurs R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Med Inform Decis Mak 2006;6:30.
3. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform 2001;34:301–10.
4. Chen SF, Goodman J. An Empirical Study of Smoothing Techniques for Language Modeling. Boston, MA: Harvard University, 1998.
5. Stolcke A. SRILM—An Extensible Language Modeling Toolkit. Proceedings of the International Conference on Spoken Language Processing, Denver, Colorado, September 16–20, 2002. 2002;2:901–4.
6. Wicentowski R, Sydes MR. Identifying smoking status from implicit information in medical discharge summaries. i2b2 Workshop on Natural Language Processing Challenges for Clinical Records. Proc AMIA Annu Fall Symp 2006.