# Identifying Smoking Status From Implicit Information in Medical Discharge Summaries

**Richard Wicentowski**[1], **Matthew R. Sydes**[2]
[1] Swarthmore College, Computer Science Department, Swarthmore, PA, USA
[2] Medical Research Council Clinical Trials Unit, London, UK

**Abstract**

*Human annotators and natural language applications are able to identify smoking status from discharge summaries with high accuracy when explicit evidence regarding their smoking status is present in the summary. We explore the possibility of identifying the smoking status from discharge summaries when these smoking terms have been removed. We present results using a Naïve Bayes classifier on a smoke-blind set of discharge summaries and compare this to the performance of human annotators on the same dataset.*

## INTRODUCTION

The goal of this study was to identify the smoking status of patients from their hospital discharge studies. This was done as part of the 2006 i2b2 NLP Shared Task. The organizers of the shared task provided discharge summaries for patients whose smoking status had been previously determined by human annotators. These annotators were blind to the true smoking status of the patients and were asked to take into account:

1. The explicit mention of smoking status in the patient's discharge summary.
2. Any "general knowledge of medicine and common sense" which would determine the smoking status.

Annotators were asked to label each patient as either a "Non-Smoker", "Current Smoker", "Past Smoker" or simply "Smoker" if temporal information was not present. If the annotator was unable to determine the smoking status, the label "Unknown" was assigned.

Our visual inspection of the dataset revealed that all records could be labeled based solely on the inclusion (or exclusion) of explicit mention of the patient's smoking status: there were no instances where the annotator would have needed to rely on their general knowledge or common sense to provide a label.

This observation motivated our construction of a simple rule-based classifier that determined the smoking status of discharge summaries based solely on these explicitly presented smoking words. This classifier proved to be far more successful than we had anticipated.

A greater challenge was to identify smoking status in the absence of explicit cues. Attempting a similar a task, Zeng et al (2006) 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."[1]

Our study investigates identifying smoking status from discharge summaries which have no explicit mention of smoking status in them ("smoke-blind" discharge summaries). We present results on this dataset using a Naïve Bayes classifier and compare it to the performance of human experts on the same dataset.

## METHODS

### The rule-based system

The first classifier we built was a rule-based classifier which attempts to match each clause[1] in a hospital discharge summary against a list of 7 hand-crafted rules. The classification was

[1]Clauses were formed by splitting sentences on the words "and" and "but".

determined by matching each clause in the record in turn against an ordered list of rules. If no rule was matched in the entire record, the label "Unknown" was assigned.

The rules make use of three classes of words described below.

NOT: `not`, `no`, `never*`, `deni*`, `negative`[2]

FORMER: `former`, `quit`, `stopped`, `discontinued`, `ago`

SMOKE: `smok*`, `tob`, `tobacco`, `cigarette*`, `cig`

The notation "`smok*`" indicates any word which begins "`smok`", such as "`smoking`", and "`smoker`". This helps to identify morphological variants.

The seven rules, listed below, specify the label of the record if the explicit smoking cue is present in a single clause. For example, "Label this record as Non-Smoker if the word `non-smoker` appears".

1. Non-Smoker: `non-smoker`
2. Non-Smoker: NOT appears before SMOKE.
3. Past Smoker: `ex-smoker` or `smoked`
4. Past Smoker: FORMER and SMOKE appear.
5. Current Smoker: `smoke` or `cigarette` appear.
6. Current Smoker: `packs` before `year`, `week`, `month` or `day`
7. Past Smoker: `tob` or `tobacco`

It is notable that each of the above rules contains either a word from the SMOKE class or a phrase such as `packs per day` (rule 6). Thus, the rule-based classifier makes decisions based *solely* on the presence of explicit cues.

## Creating the "smoke-blind" dataset

In order to pursue a computational approach to determining smoking status in the absence of explicit smoking words, we needed to create a set of documents where such information was not present (the "smoke-blind" dataset). To do this, we first removed from the training set all of the discharge summaries which were labeled as Unknown. The remaining 146 records were hand-edited[3] to remove overt references to smoking.

For consistency, we chose to remove all the words and phrases that corresponded to the cues we searched for in the rule-based classifier: all SMOKE class words, and phrases of the form `packs` per time-period. We were careful to reword all sentences (and renumber all bulleted lists) where smoking references were removed. Many instances were straightforward, but some required careful editing. As a simple example, the sentence "She does not smoke or drink." (record #868) was rewritten as "She does not drink."

## The "smoke-blind" systems

The second classifier we built was a Naïve Bayes classifier trained on the "smoke-blind" data described above. A Naïve Bayes classifier chooses the label which maximizes the similarity between a record, $R$, and a class label, $C_j$, where the similarity is defined as:

$$Sim(R, C_j) = P(R, C_j) = P(C_j)P(R|C_j) \quad (1)$$

The *a priori* probability of the class labels $C_j$ was assumed to be uniform, and $P(R|C_j)$ was based on a bigram language model using modified Kneser-Ney discounting[3].

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

The second system ("NB System 2") included additional records that were part of the competition's official test set. We automatically labeled these new records using our rule-based classifier, then made the records "smoke-blind" using the previously described procedure. The combination of these new records and the original "smoke-blind" dataset formed a larger training set for this second system with 104 smoking and 83 non-smoking records.

Both systems were evaluated using leave-one-out cross-validation, ensuring that neither system was ever trained on the data that was being classified, yet maximizing training size. These classifiers used coarse-grained labels only, folding "Past Smoker" and "Current Smoker" into the existing label "Smoker".

## Expert annotation

We recruited three human annotators with expert medical knowledge: a statistician experienced

[2]This set of negation words was inspired by [2].

[3]All editing performed by co-author R.W.

| | Fine Grained | | | | | Coarse Grained | | |
|---|---|---|---|---|---|---|---|---|
| | Current Smoker | Past Smoker | Smoker | Non Smoker | Unknown | Smoker | Non Smoker | Unknown |
| # Records | 35 | 36 | 9 | 66 | 252 | 80 | 66 | 252 |
| Sensitivity | 77.1% | 77.8% | 0.0% | 97.0% | 99.2% | 97.0% | 97.5% | 99.2% |
| Specificity | 96.4% | 96.4% | n/a | 99.4% | 99.3% | 99.4% | 99.1% | 99.3% |
| Precision | 67.5% | 68.3% | n/a | 97.0% | 99.6% | 97.0% | 96.3% | 99.6% |
| F-measure | 72.0% | 72.7% | n/a | 97.0% | 99.4% | 97.0% | 96.9% | 99.4% |

Table 1: Performance of rule-based classifier measured using fine-grained and coarse-grained labels on the original 398 discharge summaries in the training data.

in oncology clinical trials (A1), an Oncology Certified Nurse (A2), and an oncology research fellow (A3). The annotators were asked to label a subset of 54 smoke-blind discharge summaries[4] which included 34 smokers and 20 non-smokers.

Annotators were asked to make (educated) guesses about smoking status based on their knowledge of medicine and common sense. We provided guidelines[4] 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 non-smoker.

These annotators used only coarse-grained smoking status: Smoker, Non-Smoker and Unknown, omitting Current Smoker and Past Smoker.

We obtained a combined answer (Â) by taking a simple plurality of the three annotators. We considered "Unknown" a non-vote, returning the label "Missing" when there was no plurality, or when all three annotators chose "Unknown".

### Analysis

We assessed the performance of the rule-based system, Naïve Bayes models and our human annotators using standard methodology from NLP and medical statistics fields to calculate recall (sensitivity), precision (positive predictive value), specificity and F-measure.

We submitted three entries to the i2b2 Shared Task. The first entry labeled the test dataset of 104 records using the rule-based classifier. The second entry used Naïve Bayes (NB) System 1, and the third entry used NB System 2. The performance results of these entries are not known at the time of writing.

[4]See `http://nlp.cs.swarthmore.edu/i2b2/`

| Plurality Classification | True Label | |
|---|---|---|
| | Non-Smoker | Smoker |
| Non-Smoker | **6** | 2 |
| Smoker | 7 | **24** |
| Missing | 7 | 8 |

(2a) Confusion Matrix

| | True Label | |
|---|---|---|
| | Non-Smoker | Smoker |
| # of Summaries | 20 | 34 |
| Precision | 75.0% | 77.4% |
| Recall | 42.9% | 73.8% |
| F-Measure | 55.6% | 76.9% |

(2b) Performance of Plurality Classifications

Table 2: The plurality result of human experts (Â) using coarse-grained labels on 54 smoke-blind discharge summaries, shown separately for Smokers and Non-Smokers.

## RESULTS

Table 1 shows the performance of rule-based classifier measured using fine-grained and coarse-grained labels on the original discharge summaries in the training data. We note that the rule-based classifier never guesses Smoker, so all 9 of the records labeled Smoker were mislabeled as either Current Smoker or Past Smoker. Excellent results are observed for Unknown and Non-Smoker, though there was a fair amount of confusion between Past and Current Smokers using the fine-grained labels.

Table 2 shows the performance of the plurality result (Â) of human experts using coarse-grained labels on 54 smoke-blind discharge summaries.

| | Precision | Recall | F-Measure |
|---|---|---|---|
| A2 | 73.9% | 63.0% | 68.0% |
| NB Sys. 1 | 66.7% | 66.7% | 66.7% |
| NB Sys. 2 | 64.8% | 64.8% | 64.8% |
| Â | 76.9% | 55.6% | 64.5% |
| A1 | 72.5% | 53.7% | 61.7% |
| A3 | 100.% | 11.1% | 20.0% |

Table 3: Performance of human experts (A1, A2, A3), plurality result (Â), and Naïve Bayes sytems, using coarse-grained labels on 54 smoke-blind discharge summaries, sorted by F-Measure.

| | | Precision | | |
|---|---|---|---|---|
| Annonator | Recall | Human | NB 1 | NB 2 |
| A3 | 11.1% | **100.%** | 85.7% | **100.%** |
| A1 | 53.7% | 72.5% | 72.5% | **74.4%** |
| Â | 55.6% | **76.9%** | 71.4% | 73.2% |
| A2 | 63.0% | **73.9%** | 66.7% | 64.8% |

Table 4: Precision of the Naïve Bayes systems compared to the human annotators at the annotator's level of recall. NB achieves higher precision by eliminating low-confidence classifications.

Precision classifying Smokers and Non-Smokers was quite similar, but Recall for the Non-Smokers was poor.

Table 3 shows the performance of the individual human experts and their plurality result on the same 54 smoke-blind summaries.

Figure 1 shows a plot of precision against recall for the Naïve Bayes classifiers 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 Naïve Bayes classifiers are broadly similar to each other and to the humans. By eliminating low-confidence guesses, the classifier can achieve higher precision at the expense of recall.

## DISCUSSION

We have shown that it is possible to accurately extract smoking status from hospital discharge summaries using a rule-based classification system which focuses on simple words and phrases. This approach had very high rates of precision and recall on the coarse-grained approach, but was less successful with the fine-grained approach. This was partly due to conflicting information in some records. For example, one record (#685 in the test data) says "no tobacco" on one line, followed by "Smoked 3 packs per day x 17 years" on the next. This was classified as a Non-Smoker by the rule-based system, although they are clearly a Past Smoker.

As a greater challenge, 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 Naïve Bayes approach yielded reasonable levels of accuracy within the constraints of this task.

In producing the smoke-blind dataset, we chose not to remove information about medications which might be clear indications of smoking status, such as Nicoderm patches. We felt that (a) we did not wish to include a potentially very long list of contemporary proprietary names and (b) by expanding the list of smoking-related words we worried about sliding down a slippery slope: moving perhaps through "cigarette lighters" (common to smokers, but not exclusive), through "boxes of matches" to "whitening toothpaste" for which smokers may have a greater need.

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 would expect with a statistical model. We used a purposive sample of 3 annotators with expert medical knowledge to provide a comparison. Given further time and resources we would have preferred a larger and somehow more representative set of annotators with a larger training set to annotate. This would have given us a fairer and more reliable estimate of human performance with which to informally compare our statistical models. In summary, we have shown that the simple Naïve Bayes approach provides results not dissimilar to our expert human annotators, both on individual scores and overall.

We made efforts to standardize the methodology of the human annotators during the assessment period. We provided the first annotator with only rough guidelines for labeling the summaries. For the final two annotators, we formalized the methodology used by the first annotator by providing more explicit guidelines. Our annotators also provided a confidence rating for each of

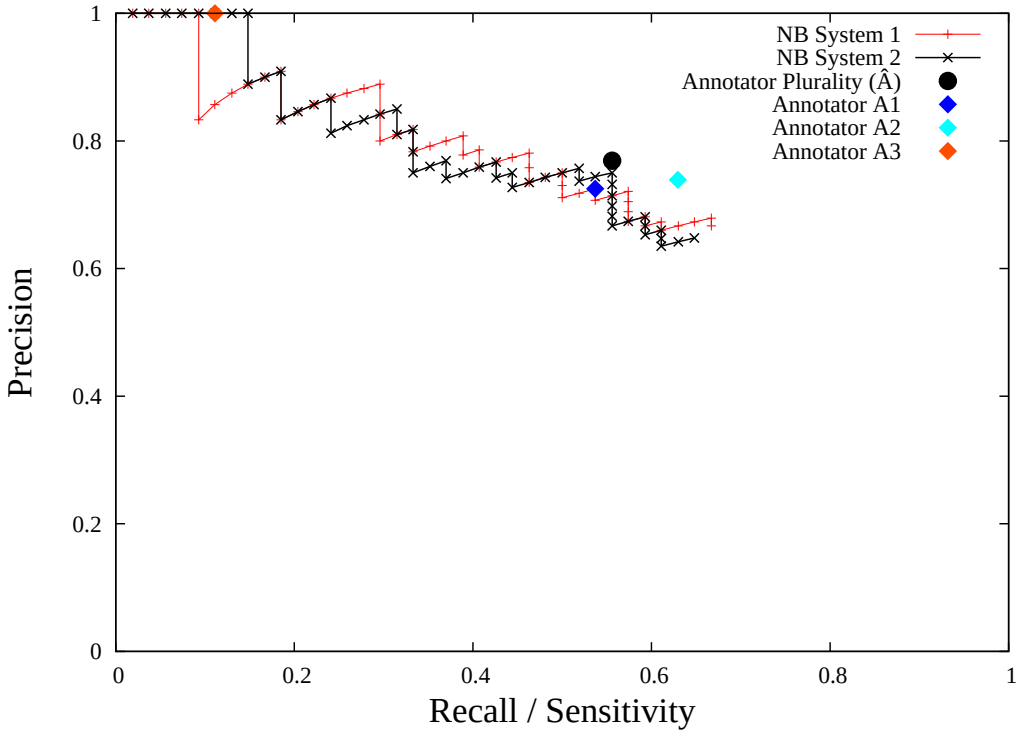


Figure 1: Comparison of the two Naïve Bayes (NB) classifier systems against the expert human annotators on 54 smoke-blind discharge summaries

their their labels but we were unable to make good use of these scores in this situation since we felt that the measures of confidence had not been uniformly interpreted by the annotators. However, the guidelines will be well developed for future annotators.

With further annotators, we would envision three methods for arriving at a group answer from our annotators. First, we could continue to use a simple plurality vote, as we do here, disregarding confidences. Second, we could use a weighted vote scheme, asking annotators to provide a quantification of their confidence in the label for each record. The overall human score would likely be less confident than the individual scores of the rating humans due to regression to the mean. Finally, we would like to explore having a group of annotators discuss each summary and reach a consensus (or large majority) decision, rather than rating in isolation.

We had originally intended to investigate a computational approach which used medical keywords, as also suggested by Zeng et al [1] to identify patients as either Smokers or Non-Smokers in the smoke-blind discharge summaries. To this end, we had asked annotators to note verbatim the keyword clues that they had used to estimate 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 which are more prevalent among smokers than non-smokers, e.g. emphysema and lung cancer. Similarly, there are social habits which may be expected to correlate with smoking.

In theory, one could derive a list of such keywords and base a probability of a given patient smoking on these presence of these keywords. We found this was not practicable for a number of reasons. Firstly, 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. Indeed, the medical literature is not entirely clear on what smoking is a risk factor for exactly. This would lead to under-prediction of smoking status.

Secondly, 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 non-smokers also have the condition. For example, many smokers experience stroke (cerebrovascular accident, CVA) but many CVAs are in non-smokers. This would lead to many false positive predictions of smoking status.

Thirdly, while keywords for smoking are potentially feasible, it is harder still to develop a sufficient list of keywords that counsel against smoking (or that predict non-smoking), especially in the context of hospital records. We note that the best annotator at predicting non-smoking (A2) did take the most sophisticated approach to this, e.g. obese people were seen as less likely to smoke, very elderly people may not have smoked (because they have lived to be old) and pregnant women commonly do not smoke.

We also note one annotator (A1) was more explicit in trying to strictly follow a knowledge-based keyword approach (with awareness of its limitations) but that the responses from this annotator were better than the keyword approach alone would be based on preliminary evidence. The interpretation is that, despite an explicit policy of assessing on keywords, the annotator was implicitly supplementing the keyword approach with additional information.

We note some practical and conceptual limitations in our ability to perform the i2b2 Shared Task. The most difficulties arose with the records where the correct task answer was "Unknown". Firstly, a large proportion of patients in the training dataset were Unknown and this greatly reduced the size of the record set that could be used for training. A complete labeling would have provided 398 training records, rather than the 146 we had to work with.

Moreover, the Unknown category is an artificial construct. Philosophically, one must ask: what is "truth" in the context of these records? The truth in the Shared Task is that of the annotators chosen by the task organizers; the task requires us to predict their truth. More fundamental is the underlying truth: the truth of the patient. The patient's smoking status must be known to the patient, even if it is labeled Unknown in the shared task data set. Therefore, it would be more truthful and more clinically relevant to model the patient's truth rather than that of the label set.

We addressed these issues by working only with records where smoking status was not "Unknown" in the training set. This provided a smaller but better defined record set and one in which calculations of specificity could be naïvely but consistently expressed.

We also found a similar conceptual struggle when dealing with the fine-grained approach with regards to the label "Smoker". The underlying truth of the patient must involve temporal information which may or not have been included in the discharge summary.

## 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 Naïve Bayes model trained on bigrams performs less well when these smoking cues are not available, but similarly well to expert human annotators.


## Acknowledgments

The authors wish to thank Marjorie W. Lieberman, RN, OCN, General Clinical Research Center, Georgetown University, Washington, D.C., USA and Dr. Martha Perisoglou, Department of Oncology, University College Hospital, London, UK for generously donating their time to help annotate discharge summaries. (The remaining annotator was co-author M.S.)



## Address for Correspondence

Richard Wicentowski, Swarthmore College, Computer Science Department, 500 College Avenue, Swarthmore, PA 19081, USA, richardw@cs.swarthmore.edu

Matthew R. Sydes, Medical Research Council Clinical Trials Unit, 222 Euston Road, London, NW1 2DA, UK, ms@ctu.mrc.ac.uk


## References


[1] 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 Medical Informatics and Decision Making. 2006;6(30).

[2] Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. Journal of Biomedical Informatics. 2001;34:301–310.

[3] Chen SF, Goodman J. An Empirical Study of Smoothing Techniques for Language Modeling. Harvard University; 1998.