Technical Brief ■

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

RICHARD WICENTOWSKI, MATTHEW R. SYDES

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 to 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.

1 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 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.

1.1 The Datasets

The organizers of the 2006 i2b2 NLP Shared Task provided copies of hospital discharge summaries for 502 patients. Of these, 398 summaries were provided as training data, and 104 summaries formed the test set. These records had been de-identified by the shared task organizers so that no individual patient was identifiable. An undisclosed number of human reference annotators provided the smoking status of the patients from these records. These reference annotators were blind to the true smoking status of the patients and were asked to ascertain the smoking status by taking into account:

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

Reference annotators were asked to label each patient as a "Non-Smoker", a "Current Smoker", a "Past Smoker", or simply as a "Smoker" if temporal information was not present or definitive. If the reference annotator was unable to determine the smoking status, the label "Unknown" was to be assigned. These five labels formed the fine-grained label set. The responses provided by the reference annotators were taken as the "true" smoking status of each patient for the purposes of the shared task; because the dataset was de-identified, the patients could not be consulted to verify this information.

1.2 The Initial Task

Our visual inspection of the provided training dataset revealed that the smoking status of all patients' records could be determined based solely on the inclusion (or exclusion) of explicit mention of the patient's smoking status. Despite the instructions provided to the reference annotators, there were no instances where the reference annotator should have needed "to rely on their general knowledge or common sense" to provide a label.

This observation motivated the construction of a simple rule-based classifier (see Section 2.1) that accurately determined the smoking status of discharge summaries based solely on these explicitly presented smoking words. This simple classifier proved to be highly successful, motivating extensions to the original task.

1.3 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 Sections 2.2 through 2.4, 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.

2 Methods

2.1 The Rule-Based Classifier

The first classifier we built was a rule-based classifier that attempted to match each clause in a hospital discharge summary against a list of 7 rules. The classification was determined by matching each clause in the hospital discharge summary (record) in turn against an ordered list of rules. In other words, the classification of the record is determined by the first matching clause. If no clause was matched in the entire record, the label "Unknown" was assigned.

The rules make use of four classes of words described below:

1. **NOT class** [3]**:** not, no, never*, deni*, negative
2. **FORMER class:** former, quit, stopped, discontinued, ago
3. **SMOKE class:** smok*, tob, tobacco, cigarette, cig
4. **TIME class:** year, month, week, day

The "*" notation is used to denote all matching morphological variants. For example, "smok*" includes words such as "smoke", "smokes", "smoked", "smoking", "smoker" and "smokers".

Each of the seven rules listed below consist of a label and a keyword trigger. The keyword triggers were created manually based on visual inspection of the provided training data. A discharge summary is classified with the specified label (e.g. "Non-Smoker") if the trigger (e.g. "**NOT** appears before **SMOKE**") is present in a single clause (see Figure 1). For example, Rule 1 below is read: "Label this record as Non-Smoker if the keyword `non-smoker` appears in this clause".

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

It is important to note 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* evidence of smoking status.

```
for each clause c in the summary d:
  for each rule r in the ruleset:
    if c contains r's trigger:
      classification = r's label
if no clause matched:
  classification = Unknown
```

Figure 1. Pseudo-code for the rule-based algorithm

2.2 Creating the "Smoke-Blind" Dataset

In order to pursue a computational approach to determining smoking status in the *absence* of explicit evidence, we first needed to create a set of documents where 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 which were labeled as Unknown. The remaining 146 records were hand-edited (RW) 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 described in Section 2.1. Therefore, all **SMOKE** class words, and phrases of the form packs per **TIME** were removed. We were careful to reword all sentences (and renumber all bulleted lists) where these smoking references were removed in order to retain correct grammar. Many instances were straightforward; for example, the sentence "She does not smoke or drink." (record #868) was rewritten as "She does not drink.". However, some required more complex, careful editing; for example, hospital discharge summary #626 in the training set contained the following: "She is a resident of Barnes-vantsver Community Hospital, where she was smoking in the smoking room, when her clothing caught on fire, necessitating admission to the Fairm of Ijordcompmac Hospital." For the purposes of removing all explicit smoking terms, we could have minimally removed "she was smoking"; however, the resulting sentence, indicating that she was in the smoking room, would have been a very strong indicator that she was a smoker. Alternatively, we could have removed the phrase "where she was smoking in the smoking room". In this case, the resulting sentence, "She is a resident of Barnes-vantsver Community Hospital when her clothing caught on fire, necessitating admission to the Fairm of Ijordcompmac Hospital", would have been free of explicit smoking terms, but the tense of the main verb would have been incorrect. Therefore, the final change we made was replacing the second word of the sentence, "is", with the word "was".

2.3 The "Smoke-Blind" Systems

In order to classify the smoke-blind data set, we chose to avoid another keyword-based system. Though 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 which could learn these phrases and weights from training data. We chose to use a Naïve Bayes (NB) classifier, trained on word bigrams found in the "smoke-blind" data described in Section 2.2. 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:

$$\mathrm{Sim}(R,C_j) = \mathrm{P}(R,C_j) = \mathrm{P}(C_j)\mathrm{P}(R \mid C_j)$$

The *a priori* probability of the class labels C_j was assumed to be uniform since 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 non-smoking 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 non-smoking 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 used 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.

2.4 Expert Annotation

Since 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 Naïve Bayes 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. We created this subset by first determining the number of records that we wanted annotated (55), and then determining the underlying distribution of the records (35 smokers, 20 non-smokers). We then selected each smoking and non-smoking records at random.

Upon receipt of the annotated summaries, two annotators pointed out that we had failed to remove all of the explicit smoking terms from a particular discharge summary. We decided to omit this discharge summary from our dataset, yielding a total of 54 summaries, comprised of 34 smokers and 20 non-smokers.

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 non-smoker.

As was done with the Naïve Bayes task described in Section 2.3, these annotators were asked to provide only coarse-grained smoking status: Smoker, Non-Smoker 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 *non-vote* (see Figure 2, row 3), and returned the label "Missing" when there was no plurality, or when all three annotators chose "Unknown" (see Figure 2).

A1	A2	A3	Â
Smoker	Non-Smoker	Smoker	Smoker
Smoker	Non-Smoker	Unknown	Missing
Smoker	Unknown	Unknown	Smoker
Unknown	Unknown	Unknown	Missing

Figure 2. Sample calculation of the plurality answer Â

Table 1 ■ Confusion matrices from the rule-based classifier.

Truth	Classification				
	C	P	S	N	U
C	27	8	0	0	0
P	6	28	0	1	1
S	7	2	0	0	0
N	0	2	0	64	0
U	0	1	0	1	250

(*1a*) Fine-grained / Training

Truth	Classification				
	C	P	S	N	U
C	5	6	0	0	0
P	3	7	0	1	0
S	2	0	0	1	0
N	0	1	0	15	0
U	0	0	0	0	63

(*1b*) Fine-grained / Evaluation

Truth	Classification		
	S	N	U
S	78	1	1
N	2	64	0
U	1	1	250

(*1c*) Coarse / Training

Truth	Classification		
	S	N	U
S	23	2	0
N	1	15	0
U	0	0	63

(*1d*) Coarse / Evaluation

The abbreviations C, P, S, N, and U refer to the 5 labels: Current smoker, Past smoker, Smoker, Non-smoker, and Unknown. Matrices are shown for the training data (*1a,c*), the evaluation data (*1b,d*), and for both fine-grained (*1a,b*) and coarse-grained (*1c,d*).

Table 2 ■ Performance of the rule-based classifier measured on fine-grained and coarse-grained labels.

	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%	97.5%	97.0%	99.2%
Specificity	96.4%	96.4%	n/a	99.4%	99.3%	99.1%	99.4%	99.3%
Precision	67.5%	68.3%	0.0%	97.0%	99.6%	96.3%	97.0%	99.6%
F-Measure	0.720	0.727	0.000	0.970	0.994	0.969	0.970	0.994

(*2a*) On the original 398 discharge summaries in the training data

# Records	11	11	3	16	63	25	16	63
Sensitivity	45.5%	63.6%	0.0%	93.8%	100.0%	92.0%	93.8%	100.0%
Specificity	94.6%	92.5%	n/a	97.7%	100.0%	98.7%	97.7%	100.0%
Precision	50.0%	50.0%	0.0%	88.2%	100.0%	95.8%	88.2%	100.0%
F-Measure	0.476	0.560	0.000	0.909	1.000	0.939	0.909	1.000

(*2b*) On the 104 discharge summaries in the evaluation data

2.4 Analysis

We assessed the performance of the rule-based system, both Naïve Bayes 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 (Section 2.1). The second entry used Naïve Bayes (NB) System 1, and the third entry used NB System 2. The performance results of these entries are discussed in Section 3.

3 Results

3.1 The rule-based classifier

Table 1 shows the confusion matrices for the rule-based classifier. We note that when considering fine-grained labels, the rule-based classifier has no rules that label a discharge summary as a Smoker. Therefore, all 9 of the training records and all 3 of the evaluation records with a "true" label of Smoker were mislabeled as either Current Smoker or Past Smoker. Excellent results are observed for Unknown and Non-Smoker, though there was an element of confusion between Past and Current Smokers using the fine-grained labels. A caveat is in order regarding the results on the training data: the labels and triggers were manually created from visual inspection of the training data, so high precision when applying the rule-based classifier to the training data is to be expected.

Table 2 shows the performance of rule-based classifier on a per-label basis, measured using fine-grained and coarse-grained labels on both the original discharge summaries in the training data and on the evaluation data. Table 3 shows the aggregate system performance using micro-averaged (weighted) F-measure.

Table 3 ■ Performance of the rule-based classifier using micro-averaged F-measure.

	Fine-Grained	Coarse-Grained
Training Data	0.927	0.985
Evaluation Data	0.865	0.971

3.2 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 Non-Smokers, 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 Naïve Bayes classifier at levels of recall that match the human annotators.

Table 4 ■ Plurality result of human experts (Â) on 54 discharge summaries.

Truth	Classification		
	Smoker	Non-Smoker	Missing
Smoker	24	2	7
Non-Smoker	7	6	7

(*4a*) Confusion Matrix

	True Label	
	Smoker	Non-Smoker
# of Summaries	34	20
Precision	92.3%	46.2%
Recall	70.6%	30.0%
F-Measure	.800	.364

(*4b*) Performance of Plurality Classifications

F-Measure	.639

(*4c*) "System" Performance of Â

Table 5 ■ Smoke-blind comparison.

	Precision	Recall	F-Measure
A2	73.9%	63.0%	.680
NB Sys. 1	66.7%	66.7%	.667
NB Sys. 2	64.8%	64.8%	.648
Â	76.9%	55.6%	.645
A1	72.5%	53.7%	.617
A3	100.0%	11.1%	.200

Human experts (A1-A3), plurality result (Â) and the two Naïve Bayes systems on 54 smoke-blind summaries (sorted by F-Measure)

Table 6 ■ Precision of the NB systems compared at similar levels of recall.

		Precision		
Annotator	Recall	Human	NB1	NB2
A3	11.1%	100.0%	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%

The Naïve Bayes classifiers achieve higher precision through the elimination of low-confidence classifications

Figure 3 shows a plot of precision against recall for the Naïve Bayes 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 Naïve Bayes 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 Naïve Bayes systems, and Table 8 shows the per-label and overall system performance for both systems.

Table 7 ■ Naïve Bayes Confusion Matrices

Truth	Classification		
	S	N	U
S	13	12	0
N	6	10	0
U	0	0	63

(*7a*) NB System 1

Truth	Classification		
	S	N	U
S	14	11	0
N	7	9	0
U	0	0	63

(*7b*) NB System 2

Confusion matrices for the two NB Systems submitted to the i2b2 Shared Task evaluated on smoke-blind data.

Table 8 ■ Naïve Bayes Performance

	Naïve Bayes System 1		
	S	N	U
# Records	25	16	63
Sensitivity	52.0%	62.5%	100.0%
Specificity	92.4%	86.4%	100.0%
Precision	68.4%	45.5%	100.0%
F-Measure	.591	.526	1.00
Overall	.829		

	Naïve Bayes System 2		
	S	N	U
# Records	25	16	63
Sensitivity	56.0%	56.2%	100.0%
Specificity	91.1%	87.5%	100.0%
Precision	66.7%	45.0%	100.0%
F-Measure	.609	.500	1.00
Overall	.829		

Performance of the two Naïve Bayes systems for the 104 discharge summaries in the evaluation data.

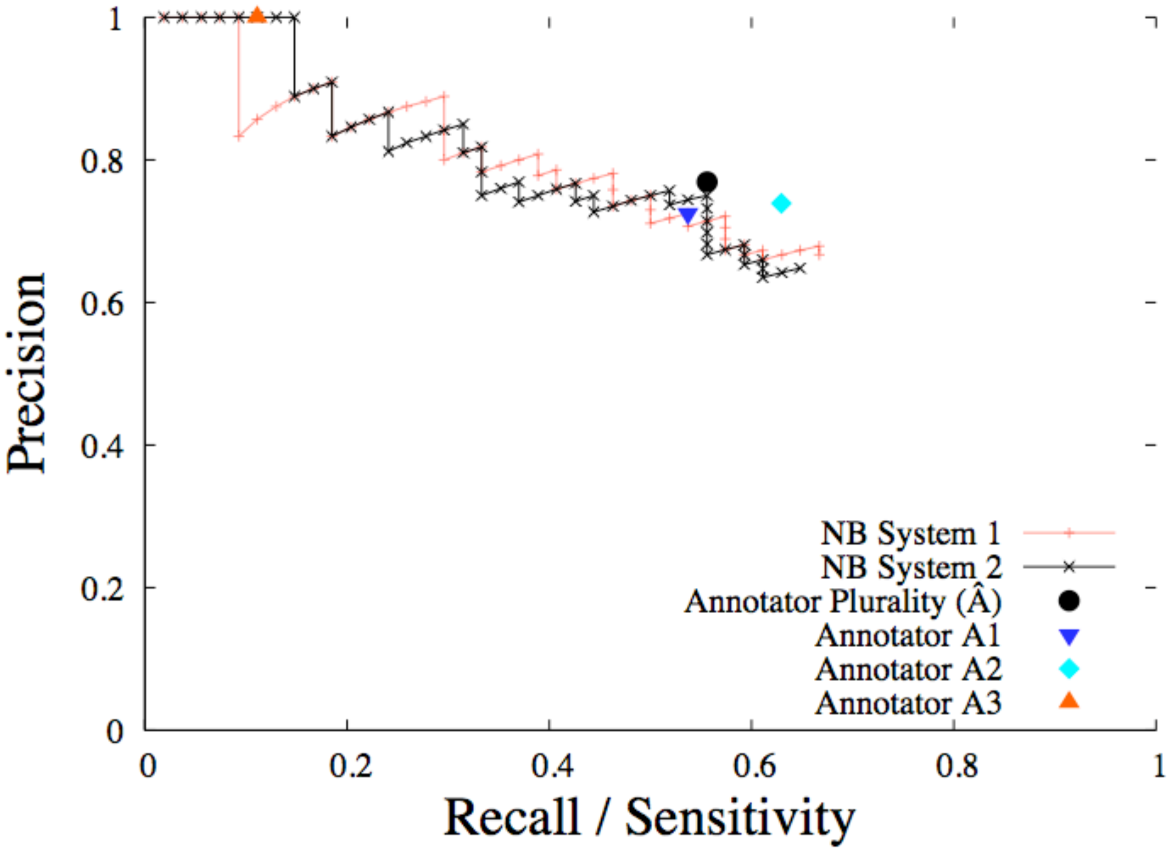

Figure 3. Comparison of the two Naïve Bayes (NB) classifier systems against the expert human annotators on 54 smoke-blind discharge summaries.

4 Discussion

4.1 The rule-based classifier

We have shown that it is possible to accurately extract smoking status from hospital discharge summaries using a rule-based classification system that uses simple keywords and phrases as triggers for rules. Using coarse-grained labels, this approach had very high rates of precision and recall, though it was less sensitive to fine-grained labels. The poorer performance with fine-grained labels compared with coarse-grained labels may be partly due to the conflicting information presented in some discharge summaries (records). For example, one record (#685 in the test data) says "no tobacco" on one clause, followed by "Smoked 3 packs per day x 17 years" in the next clause. This was classified as a Non-Smoker by the rule-based system because the first matching clause, "no tobacco", triggered Rule #2 and the patient was labeled as a Non-Smoker. However, taking both clauses into account, a human would deduce that "no tobacco" indicated that the patient doesn't currently smoke and that "Smoked 3 packs per day x 17 years" indicated the patient's past usage. This reading is consistent with this record's reference annotation, Past Smoker.

It is likely that there are further terms that could have been added to our word list, such as "*n't", "ex-" and "nicotine" which may have further improved accuracy. Recall that the initial list of words was formed by visually inspecting the training data. If indicative keywords were present in the evaluation data but not in the training data, the rule-based system could not have been effective. We also note that manually created keyword triggers are often insensitive to common typographic errors (e.g. 'tobaco' [sic]) and low frequency acronyms and abbreviations (e.g. 'cig').

4.2 The Naïve Bayes Classifier and the Smoke-Blind Dataset

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, even given a training set of limited size.

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, especially one where we would not know all of the elements 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 but which every member of the population may reasonably need (or so manufacturers would have us believe).

This 'slippery slope' can be illustrated by returning to record #626 (see Section 2.2). Recall that we had edited the sentence to read: "She is a resident of Barnes-vantsver Community Hospital when her clothing caught on fire, necessitating admission to the Fairm of Ijordcompmac Hospital." If we had been using stricter guidelines about what to exclude, we may have been conflicted about whether or not the phrase "when her clothing caught on fire" was a positive indicator of smoking, since it is highly likely that smoking would have been the cause. However, our strict editing guidelines required that we leave that phrase intact.

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 3 annotators with expert health/medical knowledge to provide a comparison. Given additional 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, more representative and more reliable estimate of human performance with which to informally contrast our statistical models. With our small set of annotators, we have shown that the simple Naïve Bayes approach provides results not too dissimilar from expert human annotators, both individually and combined (Â).

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 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, such guidelines should be further developed for future annotators.

With a larger group of human annotators, we would envision three methods for arriving at a group answer from their assessments. 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 could have a group of annotators discuss each summary and reach a consensus (or large majority) decision, rather than rating in isolation; this may provide a better quality group answer.

We had originally intended to develop a computational approach which used medical keywords, as also suggested by Zeng et al [2], in order to identify the patients as either Smokers or Non-Smokers 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 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 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 under-prediction 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 non-smokers also have the condition. For example, smokers have a higher risk of having a stroke (cerebrovascular accident, CVA) but non-smokers also experience CVAs. Using CVA as a keyword trigger for predicting smoking status would lead to false positive predictions of Smoking.

Thirdly, while developing a list of keyword cues which 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 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. In a post-annotation interview, A2 stated that classification of Non-Smoker often came from social cues, e.g. obese people (who were seen as less likely to smoke), very elderly people (who were presumed not to smoke because they have lived to be old) and pregnant women (who are routinely counseled against smoking during pregnancy).

In a post-annotation interview with A1, we found that this annotator was more explicit in trying to strictly follow a knowledge-based keyword approach (with an awareness of the limitations). In other words, A1 was intentionally looking for particular keywords that could be provided to us as a basis for a keyword-based classification system. Interestingly, a preliminary study indicated that a classifier based on the keywords provided by A1 *underperformed* the actual classifications provided by A1. Our interpretation is that, despite an explicit policy of assessing on keywords, the annotator was implicitly supplementing the keyword approach with additional information.

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

4.3 Reference Annotations in the Shared Task Dataset

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 label provided by the i2b2 reference annotators was "Unknown". First, nearly two thirds of patients in the training dataset were labeled by the reference annotators as Unknown and this greatly reduced the size of the record set that could be used for training our Naïve Bayes classifier. 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 reference annotators chosen by the task organizers; the task requires us to predict the truth of the reference annotators, not the underlying truth. Yet, the patient's smoking status must be known to the patient, even if it is labeled as a “true” Unknown in the shared task data set.

We also found a similar conceptual struggle when dealing with the "Smoker" label in the fine-grained label set. The underlying truth of the patient must involve temporal information, which may or not have been included in the discharge summary. In other words, the patient knew whether they were a “Past Smoker” or “Current Smoker”, even if this information was not reflected in the reference annotation.

Therefore, it would be more truthful, and more clinically relevant, to model the patient's truth rather than that of the labeled data set. We tried to address 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.

4.4 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 word bigrams performs less well when these smoking cues are not available, but similarly well to expert human annotators.

Acknowledgements ■
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 smoke-blind discharge summaries. (The remaining annotator was co-author M.S.)

References ■

1. Uzuner, Ö., Goldstein I., Luo Y., Kohane I. “Identifying Patient Smoking Status from Medical Discharge Records”. J Am Med Inform Assoc. 2008; 15(1).
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 Medical Informatics and Decision Making. 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. Journal of Biomedical Informatics. 2001;34:301–310.
4. Chen SF, Goodman J. An Empirical Study of Smoothing Techniques for Language Modeling. Harvard University; 1998.
5. Stolcke, A. SRILM -- An Extensible Language Modeling Toolkit. Proceedings of the International Conference on Spoken Language Processing. 2002;2:901-904.
6. R. Wicentowski and M. Sydes. ``Identifying Smoking Status From Implicit Information in Medical Discharge Summaries''. In The i2b2 Workshop on Natural Language Processing Challenges for Clinical Records. Proceedings of the Fall Symposium of the American Medical Informatics Association (AMIA), 2006.