Use of Semantic Features to Classify Patient Smoking Status

Patrick J. McCormick, MEng[1], Noémie Elhadad, PhD[2], Peter D. Stetson, MD, MA[2,3]
[1]College of Physicians & Surgeons, Columbia University, New York, NY
[2]Department of Biomedical Informatics, Columbia University, New York, NY
[3]Department of Medicine, Columbia University Medical Center, New York, NY

ABSTRACT

The recent i2b2 NLP Challenge smoking classification task offers a rare chance to compare different natural language processing techniques on actual clinical data. We compare the performance of a classifier which relies on semantic features generated by an unmodified version of MedLEE, a clinical NLP engine, to one using lexical features. We also compare the performance of supervised classifiers to rule-based symbolic classifiers. Our baseline supervised classifier with lexical features yields a microaveraged F-measure of 0.81. Our rule-based classifier using MedLEE semantic features is superior, with an F-measure of 0.83. Our supervised classifier trained with semantic MedLEE features is competitive with the top-performing smoking classifier in the i2b2 NLP Challenge, with microaveraged precision of 0.90, recall of 0.89, and F-measure of 0.89.

INTRODUCTION

One of the challenges of designing methods for natural language processing of medical narratives is that it is difficult to compare approaches on a centralized data set, as data sets differ from one institution to another. Also, it is difficult for institutions to exchange medical data, even after deidentification, due to privacy and liability concerns. The recent smoking classification challenge by i2b2 (Informatics for Integrating Biology to the Bedside) has made such comparison possible[1].

In 2006, i2b2 announced an open smoking classification task using discharge summaries. Source data originated from hospitals within the Partners HealthCare system, and covered outpatient, emergency room, and inpatient domains. After deidentification, a team of pulmonologists evaluated the smoking status of each discharge summary according to detailed criteria. Every patient was classified as "smoker", "non-smoker", or "unknown". Among the smokers, if temporal hints were present in the discharge summary they were further classified as "past smoker" or "current smoker." Summaries without temporal hints remained classified as "smoker". The interannotator agreement based on explicit textual information was 0.84, as measured by Cohen's kappa[2].

The i2b2 team issued an open challenge for teams to use whatever tools at their disposal to build automated smoking classifiers. Teams developed classifier approaches with a training set of annotated data. Once development was complete, each team evaluated their work against the same test set. The results were published in a 2008 issue of JAMIA[1].

In this paper, we describe a late-entry to the i2b2 smoking challenge. Our goal is to investigate the impact of semantic features extracted from clinical notes for the task of classifying smoking status. We compare the performance of three classifiers: a symbolic, rule-based classifier that relies on semantic features, a supervised classifier which relies on lexical features only, and a supervised classifier which relies on semantic features.

Our paper is organized as follows. We first review the past approaches employed by the different teams who entered the i2b2 challenge. We then describe our three classifiers and the feature selection process. After reporting our results, we discuss the impact of different classification strategies and choice of features.

Past Approaches

Automated text classification has been widely investigated in the natural language processing and machine learning communities[3]. Most of the smoking determination systems described in the i2b2 challenge used traditional lexical techniques. Only a few attempted to derive semantic features from the text and perform classification based on those features.

Clark[4] presented a system based on the production version of the Nuance Communications medical extraction engine. This engine could normalize medical expressions and use document structure to infer meaning. Smoking references were identified as problems with a status category that indicated if smoking was asserted or denied. In the absence of smoking references, the document was marked "unknown." In addition to this engine, a supplemental corpus of over 4,000 documents was

used for training. A Support Vector Machine (SVM) classifier was used for classification. Both instance-level and document-level classification were tested, with document-level classification performing better. This system performed best overall in the challenge with a microaveraged F-measure of 0.90.

Cohen[5] used a multilayer approach that began with "hot-spot" passage isolation using smoking-related word fragments. Only the words in these passages were used for classification. Feature generation was performed using error-correcting output codes which provided binary predictions that were presented to an SVM, which was weighted based on class rarity. This approach performed second-best, with a microaveraged F-measure of 0.89.

Aramaki[6] performed information extraction followed by lexical classification. A "smoking status sentence" is defined as any sentence with one of a set of keywords: "nicotine, smoker, smoke, smoking, tobacco, cigarette". Documents with no identified sentences were marked "unknown." Classification was performed using a combination of Okapi-BM25 similarity and a K-nearest-neighbor classifier. The system placed third in the challenge with a microaveraged F-measure of 0.88. Errors were due to rare expressions and long sentences containing irrelevant words.

Wicentowski[7] primarily used a rule-based lexical approach which used keywords to identify relevant phrases within documents. This group also created a second approach which trained a Naïve-Bayes classifier on training data bigrams stripped of all smoking mentions. The rule-based approach achieved a microaveraged F-measure of 0.86.

Szarvas[8] used word chunks to identify relevant passages, with "cigar, smoke, tobacco" being most frequent. Their multi-stage processor began with a preprocessor to mark "unknown" documents, followed by a feature extractor. This team experimented with several classifiers including Artificial Neural Networks, K-nearest-neighbor, C4.5 decision tree, AdaBoost, and Support Vector Machines. After instance classification, a majority voting model determined the document classification. In the challenge, this team's method had a microaveraged F-measure of 0.85.

Savova[9] used a sentence-level classifier with three layers. The first layer applied the NLM Lexical Variant Generation library to normalize words, and present the results to a bag-of-words SVM classifier that eliminated "unknown" documents. The second layer used negation detection with an anchor word list created from the top 10 SVM features. If negation was found the document was labeled "nonsmoker." The final layer extracted temporal features and used a linear SVM to classify current versus past smokers. This approach achieved a microaveraged F-measure of 0.84.

One interesting characteristic of the i2b2 training set is that "Unknown" is the most frequent label. Clark noted that among their private data set of over 4,000 documents, over 80% had no mentions of smoking according to their natural language engine[4]. Cohen found that "hot-spots" of informative text could be identified by looking for a list of five word fragments[5]. Cohen also excluded data from the training set which produced a "zero vector" of features, in order to improve the quality of instances. These individual techniques combine to reduce the amount of non-smoking-related data presented to the classifier.

Clark compared using each smoking mention as an instance versus an entire document's worth of mentions as an instance, and found that document-level classification produced very slightly superior results. We follow this general strategy, although our specific heuristic to determine the overall label of a document differs. Our heuristic is described in more detail in the Methods section.

Our research hypothesis is that semantic features are helpful in identifying smoking status. The best-performing classifier relies on semantic features extracted from text. Along with many of the i2b2 teams profiled above it relies on support vector machines for the classification algorithm. We chose to test our hypothesis with a different class of learning algorithms, namely BoosTexter[10], in order to validate that the gain in performance comes indeed from the use of semantic features rather than the use of SVMs. BoosTexter presents the additional advantage that it is specifically implemented to handle text-based features.

METHODS

Experimental Design

To follow the i2b2 Challenge as closely as possible, all development work was performed using only the training data (n=398). Once classifiers were complete, results were generated with the test data (n=104). Test data were not manually inspected at any point. Each classifier output an XML document which was evaluated with the i2b2 reporting script.

Classifier Architecture

We experimented with several classifiers. We present here the overall architecture of all the classifiers.

In the selection step, we filter out the "unknown" instances at the document level. Each report is evaluated against the lexical pattern `"smok|cig|tob|pack[^e]|nico"`. Documents that do not match it are labeled "unknown" and are not processed further.

In the classification step, we choose one of several possible strategies for instance selection, feature selection, and classification algorithm. Three types of classifiers were used: lexical supervised, semantic supervised and semantic symbolic (i.e., rule-based).

Instance Selection

We investigated document-level and sentence-level instances in order to identify performance differences. We hypothesized that since some documents contain several mentions of smoking that more instances would be available for training at the sentence level.

When sentence-level instances were used, a heuristic merged all classifications for a given document into a single judgment. This heuristic chose the most frequent classification with the precedence: *{non-smoker, past smoker, current smoker, smoker, unknown}*.

Feature Selection

We experimented with both lexical and semantic features. For each type of instance, we also experimented with filtered and unfiltered features. For lexical features, the filter only included the sentence matching the pattern plus one sentence before and after. For semantic features, the filter only included features where the value of "problem" was smoking-related.

To extract the lexical features, each instance was stripped of formatting and reduced to a word stream. Word order was preserved.

We used the MedLEE natural language processor[11] to identify semantic features related to smoking. MedLEE was originally designed to process radiology reports, and was later extended for various tasks including encoding clinical documents with Unified Medical Language System (UMLS) codes[12] and extracting phenotypic data from biomedical abstracts[13]. MedLEE was applied to the entire set of training data without any modification to generate a structured output with semantic information about each discharge summary. MedLEE rendered each document as an XML file, in which each semantic feature is expressed as an XML node. A sample feature is shown in Figure 1.

```
<problem v="tobacco">
  <behavior v="smokes"/>
  <date v="19850000">
  <reltime v="in"/></date>
  <parsemode v="mode1"/>
  <sectname v="social history"/>
  <sid idref="s24"/>
  <status v="end"/>
</problem>
```

Figure 1. Semantic feature to represent the sentence "She quit smoking tobacco in 1985."

While MedLEE structures can contain a substantial amount of semantic data, we found three aspects useful for smoking classification: *{problem, certainty, status}*. "Problem" corresponded to the smoking-related action or object, "certainty" contained negation information, and "status" contained temporal information. We experimented with expressing each MedLEE node as a word stream containing all semantic data versus a triple of *{problem, certainty, status}*. If no "status" tag was present, but there was other semantic data (such as "date") indicating an event in the past, a status value of "previous" was inserted.

Classifier algorithms

We created a small rule-based XQuery[14] classifier to classify a given document. A portion is shown in Figure 2, demonstrating both semantic filtering and classification.

```
let $smoking_probs :=
    problem[matches(@v,
    "^(smokes|tobacco|cigarette|
      cigar|non-smoker|nicotine)$")]

let $certainty :=
    $smoking_probs/certainty/@v

if ($smoking_probs/@v="non-smoker" or
    (some $c in $certainty satisfies
    matches($c,"^(no|negative)$")))
    then "NON-SMOKER" else ...
```

Figure 2. Partial XQuery expression using semantic features to determine if instance is "non-smoker".

Our primary supervised classifier was BoosTexter. We trained with the n-grams parameter up to n=4 and with 40 rounds of boosting. For completeness we also tested the AdaBoost/J48 and SVM classifiers supplied with the Waikato Environment for Knowledge Analysis (Weka) platform[15].

RESULTS

Following the i2b2 methodology, we generated statistics both unweighted (macroaveraged) and weighted (microaveraged) by the number of documents per class. The test set (n=104) had 11

documents each in "current" and "past smoker" classes, 3 "smoker" documents, and 16 "non-smoker" documents. The remaining documents were in the "unknown" class.

We present in this section the results of three classifiers, each of which produced the best result in its category: lexical feature (text) supervised, semantic (MedLEE) supervised, and semantic rule-based.

Our best lexical classifier used sentence-level classification. The best rule-based classifier used document-level classification. The best MedLEE-based classifier used sentence-level classification with filtered semantic features expressed as MedLEE triples and the BoosTexter algorithm. Results from these systems are shown in Table 1 and Table 2.

Run	Macroaveraged		
	Precision	Recall	F-measure
MedLEE trained	0.84	0.73	0.75
Rule-based	0.62	0.58	0.58
Text trained	0.55	0.57	0.55

Table 1. Macroaveraged results.

Run	Microaveraged		
	Precision	Recall	F-measure
MedLEE trained	0.90	0.89	0.89
Rule-based	0.85	0.82	0.83
Text trained	0.81	0.82	0.81

Table 2. Microaveraged results.

We tested using lexical and semantic features in combination. Approaches which combined lexical and semantic features did better than lexical-only but worse than semantic-only.

Using the Weka classifiers instead of BoosTexter resulted in similar, but not superior results.

DISCUSSION

The prevalence of smoking in US adults as of 2005 is 20.6%[16]. Every clinician will have some smokers, past or present, in their patient pool. It is clearly important that clinical systems use best available methods to identify patients who smoke.

In the absence of pre-structured records, natural-language processing software is required to determine a patient's smoking status from the free text. This can done as an isolated analysis, or as a "structured narrative" system that attempts to identify medically relevant semantics as medical observations are added to an EMR[17].

Previous approaches have focused on either lexical techniques or combinations of semantic and medical

knowledge. Our goal in this paper was to determine the power of semantic features alone. In addition, we wanted to test the use of different classification algorithms from the previous approaches. Our results indicate that semantic features are more powerful than lexical features to classify smoking status, independently of the learning algorithm used.

The best run in the i2b2 Challenge (by Clark) had a microaveraged precision, recall, and F-measure of 0.90. This classifier also used a clinical NLP engine for feature identification. Our MedLEE supervised classifier places second among the i2b2 entries on both micro-F1 and macro-F1 measures. Table 3 shows a few of the rules (called "weak hypotheses" by BoosTexter) learned by our semantic supervised classifier. Note that most useful hypotheses are trigrams, that is, a combination of the problem, certainty and status as extracted by MedLEE (the notation "unk" means that no "certainty" or "status" field was available for this particular instance.) The first learned hypothesis, for instance, "tobacco#no#unk" means that if MedLEE interpreted the sentence as saying that there is a problem "tobacco" mentioned, but with a negation, and without any particular certainty, then the classifier labels the instance as "non-smoker." By contrast, the second hypothesis means that if MedLEE parsed the use of "cigarette" with "moderate" certainty, then the classifier should label this instance as "smoker."

tobacco#no#unk	Smokes#no#now
cigarette#moderate#unk	Smokes#high
smokes#unk#history	Non_smoker

Table 3. Sample learned rules from our semantic supervised classifier.

The rule-based approach was in the middle compared to other i2b2 entries, most of which utilized machine learning. The fact that the rule-based classifier held its own against supervised classifiers is an important confirmation that semantic features derived from MedLEE are very effective at predicting smoking status, even when not part of a training process.

Document-instance versus sentence-instance results were comparable. However, the test set of 41 non-unknown documents only had 5 with more than one instance, and none with more than 3 instances, so there was little difference in the data set of each approach.

Our method contains some elements specific to smoking and other elements that can be applied to any clinical classification problem. The filtering of documents for relevant content uses a task-specific lexical pattern. The filtering of features for smoking-related information uses task-specific lexical and

semantic patterns. The generation of lexical and semantic features is not bound to smoking, but can be applied to other clinical problems such as the upcoming obesity and co-morbidity classification challenge from i2b2. Specific to semantic features, the negation and temporal attributes expressed in the "certainty" and "status" fields apply to many clinical problems.

CONCLUSIONS

Automated classification of a patient's smoking status is clinically useful and technically feasible. Because we rely on a general-purpose clinically driven text processor (MedLEE) and a general-purpose text classifier (BoosTexter), our method is not specific to the smoking status, and can be used for other clinical classification tasks.

The classification results demonstrate that semantic feature identification and filtering provide measurable advantages to purely lexical approaches. When semantic features are used as the sole data for automated text classifiers, high precision and recall can be achieved in classification tasks.

ACKNOWLEDGMENTS

REFERENCES

1. Uzuner O, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records. J Am Med Inform Assoc 2008;15(1):14-24.
2. Cohen J. A coefficient of agreement for nominal scales *Educ Psychol Meas* 1960;20:37-46.
3. Sebastiani F. Machine learning in automated text categorization. ACM Computing Surveys 2002;34:1-17.
4. Clark C, Good K, Jezierny L, Macpherson M, Wilson B, Chajewska U. Identifying smokers with a medical extraction system. J Am Med Inform Assoc 2008;15(1):36 –39.
5. Cohen AM. Five-way smoking status classification using text hot-spot identification and error-correcting output codes. J Am Med Inform Assoc 2008;15(1):32–35.
6. Aramaki E, Imai T, Miyo K, Ohe K. Patient Status Classification by Using Rule based Sentence Extraction and BM25 kNN-based Classifier. i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data. 2006.
7. Wicentowski R, Sydes MR. Using implicit information to identify smoking status in smoke-blind medical discharge summaries. J Am Med Inform Assoc 2008;15:29 –31.
8. Szarvas G, Farkas R, Iván S, Kocsor A, Busa Fekete R. Automatic Extraction of Semantic Content from Medical Discharge Records. i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data. 2006.
9. Savova GK, Ogren PV, Duffy PH, Buntrock JD, Chute CG. Mayo Clinic NLP system for patient smoking status identification. J Am Med Inform Assoc 2008;15(1):25– 28.
10. Schapire RE, Singer Y. BoosTexter: a boosting-based system for text categorization. Machine Learning 2000;39(2/3):135-168.
11. Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural language text processor for clinical radiology. J Am Med Inform Assoc 1994;1(2):161-174.
12. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc 2004;11(5):392-402.
13. Chen L, Friedman C. Extracting phenotypic information from the literature via natural language processing. Medinfo 2004;11(Pt2):758-762.
14. XQuery 1.0: an XML query language. W3C Recommendation. 23 January 2007. Available from http://www.w3.org/TR/xquery/
15. Witten IH, Frank E. Data Mining: Practical machine learning tools and techniques. 2nd ed. Morgan Kaufmann, San Francisco. 2005.
16. State-specific prevalence of current cigarette smoking among adults and secondhand smoke rules and policies in homes and workplaces United States, 2005. MMWR Morb Mortal Wkly Rep 2006.;55(42):1148-1151.
17. Johnson SB, Bakken S, Dine D, Hyun S, Mendonça E, Morrison F, Bright T, Van Vleck T, Wrenn J, Stetson PD. An Electronic Health Record Based on Structured Narrative. J Am Med Inform Assoc 2008;15(1):54-64.