Focus on Medical Record Identification of Smoking Status JAMIA

Viewpoint Paper ■

Identifying Patient Smoking Status from Medical Discharge Records

ÖZLEM UZUNER, PHD, IRA GOLDSTEIN, MBA, YUAN LUO, MS, ISAAC KOHANE, MD, PHD

Abstract The authors organized a Natural Language Processing (NLP) challenge on automatically determining the smoking status of patients from information found in their discharge records. This challenge was issued as a part of the i2b2 (Informatics for Integrating Biology to the Bedside) project, to survey, facilitate, and examine studies in medical language understanding for clinical narratives. This article describes the smoking challenge, details the data and the annotation process, explains the evaluation metrics, discusses the characteristics of the systems developed for the challenge, presents an analysis of the results of received system runs, draws conclusions about the state of the art, and identifies directions for future research. A total of 11 teams participated in the smoking challenge. Each team submitted up to three system runs, providing a total of 23 submissions. The submitted system runs were evaluated with microaveraged and macroaveraged precision, recall, and F-measure. The systems submitted to the smoking challenge represented a variety of machine learning and rule-based algorithms. Despite the differences in their approaches to smoking status identification, many of these systems provided good results. There were 12 system runs with microaveraged F-measures above 0.84. Analysis of the results highlighted the fact that discharge summaries express smoking status using a limited number of textual features (e.g., "smok", "tobac", "cigar", Social History, etc.). Many of the effective smoking status identifiers benefit from these features.

■ **J Am Med Inform Assoc.** 2008;15:14–24. DOI 10.1197/jamia.M2408.

Introduction

Clinical narrative records contain much useful information. However, most clinical narratives are in the form of fragmented English free text, showing the characteristics of a clinical sublanguage. This makes their linguistic processing, search, and retrieval challenging.[1] Traditional natural language processing (NLP) tools are not designed for the fragmented free text found in narrative clinical records; therefore, they do not perform well on this type of data.[2] Limited access to clinical records has been a barrier to the widespread development of medical language processing (MLP) technologies. In the absence of a standardized, publicly available ground truth that encourages the development of MLP systems and allows their head-to-head comparison, successful MLP efforts have been limited, e.g., MedLEE[3] and Symtxt.[4] A few MLP systems have been developed,[5] and such efforts have successfully shown the usefulness of MLP in clinical settings.[6-8]

To improve the availability of clinical records and to contribute to the advancement of the state of the art in MLP, within the i2b2 (Informatics for Integrating Biology to the Bedside) project, the authors de-identified and released a set of clinical records from Partners HealthCare. These records provided the basis for the development of ground truth for two challenge questions:

1. Automatic de-identification of clinical data, i.e., de-identification challenge.
2. Automatic evaluation of the smoking status of patients based on medical records, i.e., smoking challenge.

Representative teams from the MLP community participated in the two challenges and met at a workshop organized by the authors to discuss the results of the challenges. The workshop was co-sponsored by the American Medical Informatics Association and met in conjunction with its Fall Symposium in November 2006. This article provides an overview of the smoking challenge

Affiliations of the authors: University at Albany, State University of New York (OU, IG, YL), Albany, NY, Massachusetts Institute of Technology (OU), Boston, MA, Children's Hospital and Harvard Medical School (IK), Boston, MA.

Supported in part by the National Institutes of Health through research grant 1 RO1 EB001659 from the National Institute of Biomedical Imaging and Bioengineering and through the National Institutes of Health Roadmap for Medical Research, grant U54LM008748.

The authors thank all participating teams for their contributions to the challenge, AMIA for their support in the organization of the workshop that accompanied the challenge, and Stephen DeLong and the anonymous reviewers of JAMIA for their feedback on this research article.

Correspondence: Özlem Uzuner, PhD, University at Albany, SUNY, Draper 114A, 135 Western Avenue, Albany, NY 12222; e-mail: <mailto:ouzuner@albany.edu>.

Received for review: 02/21/07; accepted for publication: 06/30/07.

Table 1 ■ Annotator Training Samples

No.	Sample Sentences	Smoking Status (based on text)
1	She is a past smoker, but quit two years ago when she was found to have right upper lobe nodule, which was resected and found to be positive for TB granuloma, for which she was treated with antibiotics for nine months.	Past Smoker
2	She quit smoking four months ago.	Current Smoker
3	Depression, anxiety, chronic obstructive pulmonary disease/asthma, history of tobacco abuse, chronic headaches, atypical chest pain with 6/97 Dobutamine MIBI revealing no ischemia and a history of tuberculosis exposure.	Smoker
4	No tobacco.	Non-Smoker
5	Most recently, she developed dyspnea two days prior to admission, trigger was felt to be marijuana smoke in the building where she lives where there are many drug dealers.	Unknown

and the findings of the workshop. An overview of the de-identification challenge can be found in Uzuner et al.[9]

Related Work

The smoking challenge continues the tradition of attempting to identify the state of the art in automatic language processing. Outside of the medical domain, there have been many efforts in this direction. Most of these efforts have been led by Message Understanding Conferences (MUC)[10] and the National Institute of Standards and Technology (NIST).[11] MUC organized shared tasks on named entity recognition. NIST organized a series of Text Retrieval Evaluation Conferences (TREC) on various domains including blogs and legal documents; they also organized a series of shared tasks on topic detection and tracking, speaker recognition, language recognition, spoken document retrieval, machine translation, and entity extraction. In the biomedical domain, three such prominent efforts were BioCreAtIvE[12] for information extraction, ImageCLEF[13,14] for image retrieval, and TREC Genomics[15] for question answering and information retrieval.

Keeping the goals of TREC,[16] MUC,[17] BioCreAtIvE,[18] etc., in mind for the smoking challenge, we created a collection of actual medical discharge records. We invited the development of systems that can predict the smoking status of patients based on the narratives in these medical discharge records. We limited the scope of this task to understanding only the explicitly reported smoking information. In other words, information that implicitly reveals the smoking status was excluded from this study. Our smoking challenge continued the work on the application of classification techniques to the medical domain,[19-22] and extended the MLP studies on medical discharge records.[8,23-29] Information on the smoking status of patients is important for many health studies, e.g., studies on asthma; however, before this challenge, the only system for the automatic evaluation of the smoking status of patients from their records was the HITEx system.[30]

Smoking Challenge Data

The data for the smoking challenge consisted exclusively of discharge summaries from Partners HealthCare. We preprocessed these records so that they were de-identified, tokenized, broken into sentences, converted into XML format, and separated into training and test sets. Institutional review boards of Partners HealthCare, Massachusetts Institute of Technology, and the State University of New York at Albany approved the challenge and the data preparation process.

The data for the challenge were annotated by pulmonologists. The pulmonologists were asked to classify patient records into five possible smoking status categories. For the purposes of this challenge, we defined these categories as follows:

1. A Past Smoker is a patient whose discharge summary asserts explicitly that the patient was a smoker one year or more ago but who has not smoked for at least one year. The assertion "past smoker" without any temporal qualifications means Past Smoker unless there is text that says that the patient stopped smoking less than one year ago.
2. A Current Smoker is a patient whose discharge summary asserts explicitly that the patient was a smoker within the past year. The assertion "current smoker" without any temporal qualifications means Current Smoker unless there is text that says that the patient stopped smoking more than a year ago.
3. A Smoker is a patient who is either a Current or a Past Smoker but whose medical record does not provide enough information to classify the patient as either.
4. A Non-Smoker's discharge summary indicates that they never smoked.
5. An Unknown is a patient whose discharge summary does not mention anything about smoking. Indecision between Current Smoker and Past Smoker does not belong to this category.

Second-hand smokers are considered Non-Smokers for the purposes of this study, unless there is evidence in their record that they actively smoked. Similarly, as we are only concerned with tobacco, marijuana smoking should not affect the patients' smoking status.

Annotations

In addition to being provided with the above definitions, the annotators were trained on 55 sample sentences[30] (see Table 1 for a subset) and 10 sample records.

Two pulmonologists annotated each record with the smoking status of patients based strictly on the explicitly stated smoking-related facts in the records. These annotations constitute the textual judgments of the annotators. The same two pulmonologists also marked the smoking status of the patients using their medical intuitions on all information in the records. These annotations constitute the intuitive judgments of the annotators. In all, 928 records were annotated. The interannotator agreement on the textual judgments on these records, as measured by Cohen's kappa (κ),[31,32] was 0.84; observed agreement on textual judgments was 0.93; specific agreement per category on textual judgments ranged from 0.4 to 0.98 (Table 2; also see the Methods section for definitions of Cohen's

Table 2 ▪ Observed and Specific Agreement

Agreement	Textual Judgment	Intuitive Judgment
Observed	0.93	0.73
Specific (Past Smoker)	0.85	0.56
Specific (Current Smoker)	0.72	0.44
Specific (Smoker)	0.40	0.30
Specific (Non-Smoker)	0.95	0.60
Specific (Unknown)	0.98	0.84

kappa, observed agreement, and specific agreement). The interannotator agreement on the intuitive judgments was 0.45; observed agreement on intuitive judgments was 0.73; specific agreement per category on intuitive judgments ranged from 0.3 to 0.84 (Table 2).

Guidelines for κ are subject to interpretation and depend on parameters such as the task and categories involved.[33] However, κ of 0.8 is widely used as the threshold for strong agreement.[31-35] On our data, we observed strong agreement only on the textual judgments. We further observed that the intra-annotator agreement between a given doctor's intuitive and textual judgments varied from 0.62 to 0.99. This indicates that the reliance of the intuitive judgments on the explicit textual information varies considerably from doctor to doctor. Given these observations, we limited the challenge task to the identification of the smoking status based on information that is explicitly mentioned in the records.

To generate the ground truth, we resolved the disagreements (on textual judgments) between the annotators by obtaining judgments from two other pulmonologists. We omitted the records that the annotators disagreed on from the challenge, unless a majority vote could identify a clear textual judgment for them. In all, 63 records were omitted from the challenge for lack of a clear textual judgment. In addition, annotation results showed heavy bias in the data for Unknown records. This is the least interesting category for the purposes of the smoking challenge as Unknown records do not contain any smoking-related information. To focus the smoking challenge less on the Unknown category and more on the other four categories, we omitted a portion of the Unknown records (363 records) from the challenge.

A total of 502 de-identified medical discharge records were used for the smoking challenge. Table 3 shows the distribution of annotated records into training and test sets, and into Past Smoker, Current Smoker, Smoker, Non-Smoker, and Unknown categories. The training and test sets show similar distribution of records into the five smoking categories; however, these distributions are far from uniform. This reflects the realities of real-world data; our records were drawn at random from the Partners' database, in which some smoking categories are better represented than others. In our test set, the smallest smoking category is Smokers, with only three records. The training and test data can be obtained from i2b2.org.

Table 3 ▪ Smoking Status Training and Test Data Distribution

	Training Data	Test Data
Smoking Status	Frequency (%)	Frequency (%)
Past Smoker	36 (9)	11 (11)
Current Smoker	35 (9)	11 (11)
Smoker	9 (2)	3 (3)
Non-Smoker	66 (17)	16 (15)
Unknown	252 (63)	63 (61)
Total	398	104

Methods

We evaluated system performances using microaveraged and macroaveraged precision, recall, and F-measure, as well as Cohen's kappa.

Precision, Recall, and F-Measure

Precision, recall, and F-measure are performance metrics frequently used in NLP.[36,37] These metrics are easily derived from a binary confusion matrix.

In a binary decision problem, a classifier labels entities as either positive or negative (where positive and negative represent two generic categories) and produces a confusion matrix. This matrix contains four entities: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Given such a matrix, precision is the percentage of entities classified correctly to be in a given category in relation to the total number of entities classified for the given category (Equation 1). Recall is the percentage of entities classified correctly in a given category in relation to the actual number of items in the given category (Equation 2). F-measure is the harmonic mean of precision and recall (Equation 3). β enables F-measure to favor either precision or recall. We give equal weight to precision and recall by setting $\beta = 1$.

$$\text{Precision} \quad P = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Recall} \quad R = \frac{TP}{TP + FN} \tag{2}$$

$$\text{F-measure} \quad F = \frac{(1 + \beta^2) \times P \times R}{(\beta^2 \times P) + R} \tag{3}$$

We computed precision, recall, and F-measure on each individual smoking category. We computed the microaverages and macroaverages of each of these metrics to evaluate the overall system performance. Microaverages give equal weight to each document and provide a measure of performance on each individual record; however, they are dominated by those categories with the greatest number of documents. Macroaverages give equal weight to each category, including rare ones, and as a result, discount the performance on better-populated categories.[36,37] Given the complementary strengths of microaverages and macroaverages, we reported results in terms of both. Equation 4 and Equation 5 show the formulae for microaveraged and macroaveraged F-measure respectively. Microaveraged and macroaveraged precision and recall can be obtained similarly.

$$\text{Microaveraged F-measure} \quad F(micro) = \sum_{i=1}^{M} \frac{F_i(TP_i + FN_i)}{(TP + FP)} \tag{4}$$

where M is the number of categories.

Table 4 ▪ Microaverages and Macroaverages for Precision, Recall, and F-Measure, Sorted by Microaveraged F-Measure

	Macroaveraged			Microaveraged		
Group Run	Precision	Recall	F-Measure	Precision	Recall	F-Measure
Clark_3	0.81	0.73	0.76	0.90	0.90	0.90
Cohen_2	0.64	0.67	0.65	0.88	0.89	0.89
Aramaki_1	0.64	0.67	0.65	0.88	0.89	0.88
Cohen_1	0.64	0.65	0.64	0.88	0.88	0.88
Clark_2	0.76	0.69	0.72	0.87	0.88	0.88
Cohen_3	0.62	0.62	0.62	0.87	0.88	0.87
Wicentowski_1	0.58	0.61	0.59	0.85	0.87	0.86
Szarvas_2	0.59	0.60	0.59	0.85	0.87	0.85
Clark_1	0.69	0.65	0.66	0.86	0.87	0.85
Szarvas_3	0.56	0.58	0.57	0.84	0.86	0.84
Savova_1	0.62	0.60	0.60	0.84	0.86	0.84
Szarvas_1	0.56	0.58	0.57	0.83	0.86	0.84
Sheffer_1	0.59	0.59	0.58	0.83	0.86	0.84
Savova_2	0.56	0.57	0.56	0.81	0.84	0.82
Savova_3	0.55	0.55	0.55	0.80	0.83	0.81
Pedersen_1	0.55	0.56	0.54	0.82	0.82	0.81
Guillen_1	0.45	0.51	0.44	0.77	0.79	0.76
Carrero_1	0.52	0.47	0.48	0.74	0.77	0.75
Carrero_2	0.44	0.43	0.41	0.71	0.71	0.70
Rekdal_1	0.68	0.45	0.47	0.77	0.74	0.67
Pedersen_3	0.23	0.35	0.27	0.53	0.68	0.60
Pedersen_2	0.23	0.36	0.28	0.53	0.69	0.59
Carrero_3	0.26	0.31	0.27	0.54	0.63	0.57

Macroaveraged F-measure

$$F(macro) = \frac{\sum_{i=1}^{M} F_i}{M} \tag{5}$$

where M is the number of categories.

Cohen's Kappa and Agreement

Cohen's kappa (κ) (Equation 6) is a measure of agreement[31] between pairs of annotators who classify items into a set number of mutually exclusive categories. κ depends on observed agreement (A_o in Equation 7) and the agreement expected due to chance (A_e in Equation 8). A κ value of 0.8 is widely used as the threshold for strong agreement,[31-35] whereas a κ of 0 indicates that the observed agreement is due to chance.[33] We used κ as a measure of inter-annotator and intra-annotator agreement (see Annotations section) as a measure of agreement between two automatic systems and as a measure of agreement between a system and the ground truth (see Results and Discussion). Equation 6 through Equation 8 collectively describe κ between an automatic system and the ground truth. κ for inter-annotator and intra-annotator agreement and for inter-system agreement can be computed analogously.

Cohen's Kappa

$$\kappa = \frac{A_o - A_e}{1 - A_e} \tag{6}$$

Observed Agreement

$$A_o = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

Expected Agreement Due To Chance

$$A_e = \frac{(TP + FP) \times (TP + FN) + (TN + FP) \times (TN + FN)}{TP + TN + FP + FN} \tag{8}$$

According to Hripcsak and Rothschild,[38] there exists a correspondence between F-measure and κ. However, κ provides clearer insights into the relative strengths of the systems (see Intersystem Agreement section). We evaluated systems using F-measure but compared them using both F-measure and κ.

Specific agreement (A_{sp})[33] measures the degree of agreement (Equation 9) on each category and is not adjusted by chance. We used specific agreement to get a sense of the level of agreement between annotators without taking chance into consideration.

Specific Agreement

$$A_{sp} = \frac{2TP}{2TP + FP + FN} \tag{9}$$

Significance Tests

We tested the significance of the differences of the systems using a randomization technique that is frequently utilized in NLP.[39] The null hypothesis is that the absolute value of the difference in performances, e.g., F-measures, of two systems is approximately equal to zero. The randomization technique does not assume a particular distribution of the differences. Instead, it empirically generates the distribution. Given two actual systems, it randomly shuffles (at each iteration, we simulated a coin flip to decide whether the answers should be swapped) their responses to the records in the test set N times (e.g., N = 9,999), and thus creates N pairs of pseudosystems. It counts the number of times that the difference between the performances of pairs of pseudosystems is greater than the difference between the two actual systems' performances. Let this count be equal to n and compute $s = (n + 1)/(N + 1)$. If s is greater than a predetermined cutoff α, then the difference of the performances of the two actual systems can be explained by chance; otherwise, the difference is significant at level α. Following MUC's example, we set α to 0.1.

Table 5 ▪ Significance Tests on Microaveraged and Macroaveraged F-Measures (Sorted by Microaveraged F-Measure)

	Cohen_2	Aramaki_1	Cohen_1	Clark_2	Cohen_3	Wicentowski_1	Szarvas_2	Clark_1	Szarvas_3	Savova_1	Szarvas_1
Clark_3	+*	+*	+*	+*	+*	+*	+	+	+	+*	+
Cohen_2		+	+*	+	+*	+*	+*	+	+*	+*	+*
Aramaki_1			+*	+	+*	+*	+*	+	+*	+*	+*
Cohen_1				+	+*	+*	+*	+	+*	+*	+*
Clark_2					+*	+*	+*	+*	+*	+*	+*
Cohen_3						+*	+*	+	+*	+*	+*
Wicentowski_1							+*	+	+*	+	+*
Szarvas_2								+	+*	+	+*
Clark_1									+*	+*	+*
Szarvas_3										+	+*
Savova_1											+*
Szarvas_1											
Sheffer_1											
Savova_2											
Savova_3											
Pedersen_1											
Guillen_1											
Carrero_1											
Carrero_2											
Rekdal_1											
Pedersen_3											
Pedersen_2											

Only the upper diagonal is marked.
*Pairs not significantly different in macroaveraged F-measure.
+Pairs not significantly different in microaveraged F-measure.

Smoking Challenge Submissions

A total of 11 teams participated in the smoking challenge. The training data for the challenge were released in July 2006, and the test data were released for only three days in September 2006. Each team was permitted to submit up to three system runs on the test data. A total of 23 runs were received (this count includes only one of the three runs of Wicentowski and Sydes; their other two runs were evaluated separately from the rest (see Wicentowski below) and are therefore not included in this count). In this section, we describe each team's submissions.

Aramaki

Aramaki et al.[40] presented a two-step classifier; this classifier first extracted sentences that relate to the smoking status of patients and then applied Okapi-BM25 and k-nearest-neighbors (kNN) to the extracted information. The extraction step took advantage of the observation that only a few sentences in a patient's record referred to the smoking status and that these sentences could be easily identified by their key words (e.g., smoking, tobacco). If more than one sentence in a record contained smoking information, then only the last sentence was extracted. If no such sentences were found, then the record was classified as Unknown. To predict the smoking status of a record from the test set, each sentence extracted from this record was compared with sentences from the training set. The sum of the similarity measures (Okapi-BM25) between each extracted sentence and the k most similar sentences (kNN) in the training set determined the smoking status of the extracted sentence.

Carrero

Carrero et al.[41] engineered various attributes for text classification. They experimented with Naïve Bayes, Support Vector Machines (SVMs), C4.5 Decision Tree, and AdaBoost classifiers. They found that the best performance came from the use of AdaBoost and that bigrams and trigrams helped classification more than unigrams.

Clark

Clark et al.[42] also benefited from lack of explicit smoking information in the Unknown documents and filtered these documents out prior to further classification. They presented two different approaches to processing the remaining documents. In the first approach, they classified documents based on phrase-level references to smoking using SVMs. In the second approach, they first classified each explicit phrase-level reference to smoking using additive logistic regression and then applied heuristics to derive document-level categories from judgments of phrase-level references. Where multiple references to smoking appeared, the smoking category for the collection reflected the dominant category in the set. Given the data driven nature of their classifiers, Clark et al. enriched the i2b2 data set with 1,200 additional records and their smoking categories. They hypothesized that a model based on the resulting combined data set would perform better than models trained on each of the data sets alone. This hypothesis is consistent with a popular conclusion: the error rate of a learning system is generally reduced as the sample size increases.[43]

Clark et al.'s smoking status evaluation system was developed with Weka[44] and benefited from the Nuance Medical Extraction system. This system identified document structure (e.g., sections, headings, lists, etc.), medical entities, and the status of these entities (e.g., smoking-related medication, such as Zyban and Nicoderm). It thus strengthened the

Table 5 ▪ (continued)

Sheffer_1	Savova_2	Savova_3	Pedersen_1	Guillen_1	Carrero_1	Carrero_2	Rekdal_1	Pedersen_3	Pedersen_2	Carrero_3
+										
+*	*									
+*										
+*	+*	*								
+*	+*	+*								
+*	+*	*	*							
+*	+*	+*	+*							
+*	+*	+*	+*							
+*	+*	+*	+*							
+	+*	+*	+*		*		*			
+*	*	*	+*							
+	+*	+*	+*		*		*			
	+*	+*	+*				*			
		+*	+*	+	*		*			
			+*	+*	+*		*			
				+*	+*		*			
					+	+*	+			
							+*			
							+			
										+
									+	+*
										+*

feature set consisting of unigrams and bigrams of explicit references to smoking.

Clark et al. found that classifying the documents using SVMs directly on the phrase-level references to smoking performed better than the two-step classification approach using logistic regression (see Clark_1 and Clark_3 in Table 4 and Table 5) (in this article, system runs are identified by the last name of the first author and submission number; for example, Clark_1 refers to the first system run submitted by Clark et al.). The context and span of explicit phrase-level references contributed to their interpretation. For example, the phrase "she does not currently smoke" in the Social History section could refer both to a Past Smoker and to a Non-Smoker, depending on context. However, such context and span information was difficult to identify.[42]

Cohen

Cohen[45] approached smoking status evaluation as a word-level task and applied a four-step process. The first step marked smoking-related passages. In Cohen's case, a smoking-related passage was a window of ±100 characters surrounding specific string-based features (e.g., "smok", "cig", and "tobac"). The second step tokenized the passages with the StandardAnalyzer from the Apache Lucene[46] library. The third step identified documents without any specific string-based features and filtered them out. The last step took the remaining documents and classified them using several linear SVMs. Cohen created variations of his system by adding postprocessing rules.

Guillen

Guillen[47] presented a rule-based system consisting of contextual rules for identifying the smoking status of patients. This system utilized both lexical and syntactic features, such as verbs, negations of verbs, adverbs, and tenses of verbs.

Pedersen

Pedersen[48] presented three approaches, one supervised and two unsupervised, for identifying the smoking status of patients (a description of Pedersen's smoking challenge systems appears as a JAMIA on-line supplement to the current article, and can be viewed at www.jamia.org). For his supervised approach, he experimented with several learning algorithms from the Weka toolkit[44] and found "[t]he J48 decision tree learner [to be] the most accurate . . . when evaluated using 10-fold cross validation on the training data."[48] Furthermore, this classifier gave its best performance when trained using only the following words and their presence/absence as features: cigarette, drinks, quit, smoke, smoked, smoker, smokes, smoking, tobacco. For his unsupervised approaches, Pedersen experimented with Latent Semantic Analysis and SenseClusters' second-order representation. His features were bigrams, with up to five intervening words, where one of the two words in the bigram began with the string "smok". These two systems tended to assign most of the test data to the Unknown category. Pedersen found that neither of his unsupervised approaches performed as well as his supervised approach. Table 5 shows that Pedersen's unsupervised approaches differed significantly from his supervised approach, in both microaveraged and macroaveraged F-measures, at $\alpha = 0.1$.

Rekdal

Rekdal[49] applied the Argus Medical Language Processor to the challenge. Argus is a system for medical text processing in Norwegian and handles records typed by physicians. For the smoking challenge, this system was adapted to English medical discharge summaries.

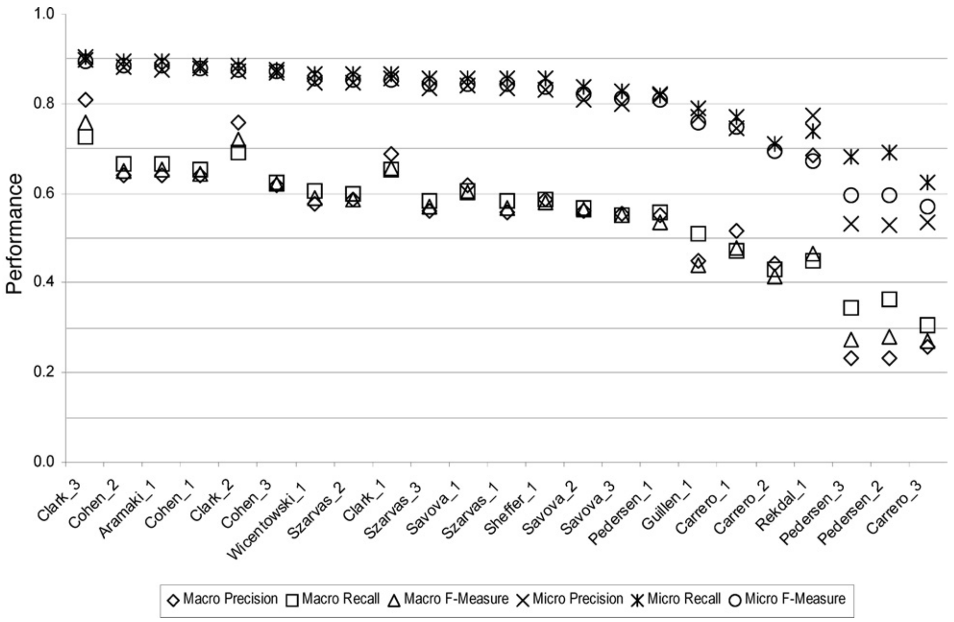

Figure 1. Results from Table 4 sorted by microaveraged F-measure.

Savova

Savova et al.[50] recast the smoking challenge as a sentence classification task. Their system marked the smoking status category of each sentence in a record and applied higher-level rules to classify the document itself. The sentence-level judgments of the training set were derived manually. Most of the text processing was handled through the components of the Unstructured Information Management Architecture (UIMA) system of IBM (http://uima-framework.sourceforge.net/) and Weka.

Sentence classification was achieved in three stages and by using lexical features with an SVM. In the first stage, the Unknown category was filtered out, creating a subcorpus of sentences that excluded this category. In the second stage, the Non-Smoker category was identified through the study of negations. For this, Chapman's Negex,[24] with minor modifications, was used. In the third stage, the Smoker, Current, and Past Smoker categories were identified by studying a subcorpus of sentences related to Current and Past Smokers. This was cast as a temporal resolution problem and was implemented using an SVM. Unlike the first stage, in this stage the tokens were not normalized so as to retain tense information of verbs. Temporal resolution features, such as the tense of the verbs and words such as day and ago, helped differentiate between Current and Past Smokers. Finally, the document-level categories were assigned. The document categories were ordered from highest to lowest priority as follows: Current Smoker, Past Smoker, Smoker, Non-Smoker, and Unknown. Each document was assigned the label of the highest priority category for which it could provide evidence. For example, a single sentence classifying Current Smoker was adequate to mark a document of any length as Current Smoker.

Sheffer

Sheffer et al.[51] adapted and applied the LifeCode expert system (A-Life Medical, San Diego, CA) to the smoking challenge. LifeCode is designed for extracting demographic and clinical information from free text, and already contains a patient smoking status module. This system consists of four components: document segmenter, lexical analyzer, phrase parser, and concept matcher. The "segmenter delimits and categorizes the content of a record based on the meanings of the section headings."[51] The lexical analyzer morphologically processes strings to match them with concepts in a database. The phrase parser syntactically parses the text bottom-up and is highly tolerant to idiosyncratic language. The concept matcher uses vector analysis to assign concept labels to each phrase. The set of extracted concepts is then refined using logic.

LifeCode provides diagnosis codes based on portions of full patient charts. This system examines multiple documents for a given patient and assigns a certainty score to indicate the confidence associated with the codes assigned to each document. To handle the differences of discharge summaries from patient charts, Sheffer et al. adapted LifeCode to the smoking challenge. During this process, they took into consideration the differences in the labels of the i2b2 smoking challenge and the system's existing smoking tags. Finally, they manipulated the system to return a single specific categorization, rather than a set of codes and confidences, for each document. The investigators reported that they found the temporal differentiation of smoking categories to be a significant challenge.

Szarvas

Szarvas et al.[52] took advantage of the short explicit references to smoking status. They examined keyword-based kNN, SVM, Artificial Neural Networks, and AdaBoost+C4.5 classifiers using Weka.[44] They combined the results of these classifiers and labeled records using a voting scheme. In addition to keywords, they used part-of-speech tags, negations, and information about verbs.

Wicentowski

Wicentowski and Sydes[53] used a rule-based system to predict the smoking status based on explicit mentions of smoking. After noting that the rule-based system performed fairly well, they explored methods for extracting information about the smoking status of patients when the records were purged of explicit smoking-related terms. For this, they removed all explicit mentions of smoking from the records and created a "smoke-blind" data set that included only the nonsmoking-related text. They trained two Naïve Bayes classifiers on the smoke-blind data. They showed that human annotators achieved around 75% precision, recall, and F-measure on the smoke-blind data. The two smoke-blind systems approached the performance of human annotators with F-measures of 0.69 on the smoke-blind test data.

Results and Discussion

Table 4 and Figure 1 show the precision, recall, and F-measure, both macroaveraged and microaveraged, for each of the system runs submitted to the smoking challenge. Table 5 shows the results of the significance tests on microaveraged and macroaveraged F-measures of the systems; this table reveals that the differences in the microaveraged F-measures of the top 12 systems are not significant at $\alpha = 0.1$; 7 of these 12 systems are not significantly different from each other in their macroaveraged F-measures at this α.

Table 4 shows that systems that used similar machine learning and/or rule-based algorithms, e.g., the systems that used SVMs, did not all necessarily perform similarly. This implies that other factors such as the engineering details and the features used with the algorithm also contributed to the final outcome.

Table 4 also shows that a majority of the top overall performances (microaveraged F-measure) came from systems that took advantage of the lack of explicit smoking information in the Unknown records and filtered those documents out prior to further processing or classification (see the results of Clark et al., Cohen, and Aramaki et al.). These systems gave comparable microaveraged F-measures to those that were built on explicit references to smoking status (Wicentowski and Sydes, Szarvas et al.). Many of the best performing systems assumed that the smoking classification was Unknown unless some sentence in the document showed it to be otherwise. Leveraging this characteristic resulted in an F-measure of 1.0 in the Unknown category for Cohen, Wicentowski and Sydes, Szarvas et al., and Savova et al.

Table 6 shows the precision, recall, and F-measure of each system run on each of the five smoking status categories. In general, the systems successfully handled the Unknown and Non-Smoker categories; had some difficulty identifying Current and Past Smokers; and, with the exception of the systems of Clark et al. (their systems correctly classified one of the three records), all failed in recognizing Smokers.

Intersystem Agreement

In addition to measuring system performance on the overall test set and on individual categories in the test set, we analyzed the performance and agreement of systems on individual data points, i.e., records, in the test set. For this, we used κ. Table 7 shows the level of κ agreement of each system with the ground truth as well as the level of κ agreement of pairs of systems. Most notable is the high level of agreement between systems submitted by the same group, but only for some of the groups. This high level of agreement is not surprising because these groups submitted runs from variations of the same basic system. The systems of Savova et al. differed in their feature selection; Cohen added postprocessing rules to submissions Cohen_1 and

Table 6 ▪ Precision, Recall, and F-Measure for All Five Categories, in Alphabetical Order

	Unknown			Non-Smoker			Past Smoker			Smoker			Current Smoker		
Group Run	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
Aramaki_1	0.98	1	0.99	0.93	0.88	0.90	0.62	0.73	0.67	0	0	0	0.67	0.73	0.70
Carrero_1	0.88	0.97	0.92	0.60	0.75	0.67	0.67	0.36	0.47	0	0	0	0.43	0.27	0.33
Carrero_2	0.90	0.89	0.90	0.48	0.81	0.60	0.33	0.27	0.30	0	0	0	0.50	0.18	0.27
Carrero_3	0.76	0.87	0.81	0.33	0.56	0.42	0.20	0.09	0.13	0	0	0	0	0	0
Clark_1	0.93	1	0.96	1	0.94	0.97	0.60	0.27	0.38	0.25	0.33	0.29	0.67	0.73	0.70
Clark_2	0.93	1	0.96	1	0.94	0.97	0.56	0.45	0.50	0.50	0.33	0.40	0.80	0.73	0.76
Clark_3	0.93	1	0.96	0.94	0.94	0.94	0.86	0.55	0.67	0.50	0.33	0.40	0.82	0.82	0.82
Cohen_1	1	1	1	0.87	0.81	0.84	0.80	0.73	0.76	0	0	0	0.53	0.73	0.62
Cohen_2	1	1	1	0.88	0.88	0.88	0.80	0.73	0.76	0	0	0	0.53	0.73	0.62
Cohen_3	1	1	1	0.88	0.94	0.91	0.70	0.64	0.67	0	0	0	0.50	0.55	0.52
Guillen_1	0.98	0.98	0.98	0.90	0.56	0.69	0	0	0	0	0	0	0.35	1	0.52
Pedersen_1	0.98	0.98	0.98	0.77	0.63	0.69	0.57	0.36	0.44	0	0	0	0.43	0.82	0.56
Pedersen_2	0.81	1	0.89	0	0	0	0	0	0	0	0	0	0.36	0.82	0.50
Pedersen_3	0.82	1	0.90	0	0	0	0	0	0	0	0	0	0.35	0.73	0.47
Rekdal_1	0.72	1	0.84	1	0.06	0.12	1	0.55	0.71	0	0	0	0.70	0.64	0.67
Savova_1	0.95	1	0.98	0.75	0.75	0.75	0.86	0.55	0.67	0	0	0	0.53	0.73	0.62
Savova_2	0.95	1	0.98	0.75	0.75	0.75	0.60	0.55	0.57	0	0	0	0.50	0.55	0.52
Savova_3	0.94	1	0.97	0.75	0.75	0.75	0.67	0.55	0.60	0	0	0	0.42	0.45	0.43
Sheffer_1	0.97	1	0.98	0.75	0.94	0.83	0.57	0.36	0.44	0	0	0	0.64	0.64	0.64
Szarvas_1	1	1	1	0.80	1	0.89	0.55	0.55	0.55	0	0	0	0.44	0.36	0.40
Szarvas_2	1	1	1	0.80	1	0.89	0.63	0.45	0.53	0	0	0	0.50	0.55	0.52
Szarvas_3	1	1	1	0.80	1	0.89	0.56	0.45	0.50	0	0	0	0.45	0.45	0.45
Wicentowski_1	1	1	1	0.88	0.94	0.91	0.50	0.64	0.56	0	0	0	0.50	0.45	0.48

Table 7 ▪ Agreement (Kappa) Between Pairs of Systems and the Ground Truth

	Ground Truth	Aramaki_1	Carrero_1	Carrero_2	Carrero_3	Clark_1	Clark_2	Clark_3	Cohen_1	Cohen_2	Cohen_3
Aramaki_1	*0.82*										
Carrero_1	0.58	0.63									
Carrero_2	0.51	0.50	0.75								
Carrero_3	0.30	0.33	0.43	0.37							
Clark_1	0.76	0.71	0.56	0.47	0.32						
Clark_2	0.80	0.74	0.53	0.45	0.32	*0.89*					
Clark_3	*0.83*	0.76	0.54	0.49	0.33	**0.91**	**0.93**				
Cohen_1	*0.80*	0.78	0.55	0.51	0.29	0.74	0.74	0.74			
Cohen_2	*0.82*	*0.80*	0.58	0.51	0.30	0.76	0.76	0.76	**0.95**		
Cohen_3	0.79	0.77	0.58	0.52	0.30	0.76	0.74	0.74	**0.92**	*0.87*	
Guillen_1	0.64	0.60	0.50	0.39	0.22	0.64	0.61	0.62	0.68	0.68	0.64
Pedersen_1	0.69	0.70	0.67	0.56	0.35	0.66	0.66	0.66	0.75	0.77	0.70
Pedersen_2	0.41	0.43	0.33	0.22	0.12	0.40	0.39	0.38	0.47	0.47	0.40
Pedersen_3	0.39	0.44	0.39	0.24	0.17	0.37	0.35	0.37	0.46	0.46	0.41
Rekdal_1	0.45	0.44	0.25	0.14	0.10	0.34	0.38	0.43	0.43	0.41	0.39
Savova_1	0.75	0.76	0.59	0.49	0.29	0.66	0.66	0.73	0.73	0.73	0.76
Savova_2	0.71	0.76	0.59	0.49	0.29	0.65	0.65	0.68	0.70	0.69	0.75
Savova_3	0.69	0.74	0.58	0.51	0.30	0.61	0.61	0.68	0.71	0.69	0.74
Sheffer_1	0.75	0.73	0.57	0.58	0.36	0.75	0.72	0.74	0.75	0.76	0.80
Szarvas_1	0.75	0.77	0.62	0.53	0.36	0.73	0.74	0.76	0.70	0.72	0.75
Szarvas_2	0.77	0.78	0.61	0.53	0.34	0.71	0.72	0.78	0.72	0.73	0.75
Szarvas_3	0.75	0.77	0.62	0.53	0.34	0.73	0.71	0.76	0.70	0.72	0.75
Wicentowski_1	0.77	*0.80*	0.57	0.50	0.33	0.71	0.74	0.74	0.75	0.74	0.75

Agreement between 0.8 and 0.9 is in italic type, and agreement >0.9 is in boldface type.

Cohen_3; Clark_2 and Clark_3 differed only in their training set; and Szarvas_2 used a voting scheme that included the classifiers from Szarvas_1 and Szarvas_3.

Table 7 also shows that, in general, the highest-ranking systems showed a reasonable level of agreement with each other. For example, comparing each of the systems of Clark et al. with each of Cohen's systems showed κ ranging between 0.74 and 0.76. Some systems showed strong agreement with each other despite the differences in their approach to smoking status identification. For example, Wicentowski and Sydes' rule-based system showed κ of 0.82 to 0.87 when compared with Szarvas et al.'s SVMs. Similarly, the systems of Aramaki et al. had a κ of 0.8 with that of Wicentowski and Sydes, yet the former classified the records at the sentence level whereas the latter was at the word level. The high level of agreement between different approaches to the smoking challenge, compounded by the F-measures of these approaches on the ground truth, indicates the richness of the set of potential solutions to smoking status evaluation.

On the other hand, the disagreements among the systems reveal their relative strengths. For example, Clark_3 and Cohen_2 disagreed with each other on 13 records (κ = 0.76). These systems showed relative strengths in different categories: Clark_3 in Non-Smoker and Current Smoker, and Cohen_2 in Unknown and Past Smoker. There were only three records that neither system marked correctly. The state of the art in smoking status evaluation may be improved by combining the strengths of such complementary systems.

Ground Truth

Several of the teams noted disagreement with the annotation of a few of the documents in the challenge data. As mentioned in the Annotations section, our medical records were annotated by pulmonologists. Although they provide a valid ground truth, human judgments can include errors. However, given the medical expertise of the annotators and the agreement on the labels of these disputed records among the annotators, we let the annotations stand as is. Nevertheless, if the plain English reading of the disputed records is correct, for all of the disputed documents, a majority of the systems found the correct classification (and disagreed with the potentially erroneous judgments of human annotators). Nonetheless, we note that "the opinions of experts should be tempered by an attempt to measure the 'weight of the evidence' that the experts interpret."[54] Even in the best of cases, the ground truth only approximates reality.

Implications for Future Research

Based on our findings on the smoking challenge, we plan to continue our investigation in two general directions. First, we are especially encouraged by the complementary strengths of the top-ranked systems and would like to investigate ways of combining these approaches to improve the overall results. Second, we find Wicentowski and Sydes' investigation with "smoke-blind" data to be intriguing. We hypothesize that a "smoke-blind" system may help provide insights into the doctors' intuitive judgments on the smoking status of patients.

Conclusion

In this report, we described the i2b2 smoking challenge, the data and the data preparation process, the evaluation metrics, and each of the submissions. We presented the evaluation and analysis of the submitted system runs, and provided a synthesis of the implications of our findings for future research.

Table 7 ■ (continued)

Guillen_1	Pedersen_1	Pedersen_2	Pedersen_3	Rekdal_1	Savova_1	Savova_2	Savova_3	Sheffer_1	Szarvas_1	Szarvas_2	Szarvas_3
0.67											
0.61	0.55										
0.56	0.54	*0.85*									
0.38	0.36	0.48	0.46								
0.64	0.66	0.45	0.49	0.42							
0.59	0.64	0.44	0.48	0.38	**0.95**						
0.59	0.62	0.41	0.45	0.37	**0.95**	**0.93**					
0.63	0.73	0.43	0.39	0.35	0.71	0.72	0.67				
0.61	0.72	0.38	0.38	0.40	0.75	0.76	0.71	*0.83*			
0.65	0.74	0.42	0.43	0.41	0.80	0.78	0.76	*0.81*	**0.95**		
0.64	0.72	0.41	0.41	0.39	0.78	0.80	0.74	*0.83*	**0.97**	**0.98**	
0.62	0.69	0.41	0.39	0.41	0.75	0.80	0.76	0.76	*0.87*	*0.82*	*0.83*

For this challenge, we built a document collection derived from actual medical discharge records, studied this collection in reference to a real-world medical classification task, and examined system performances. We showed that when asked to make a decision on the smoking status of patients based on the explicitly stated information in medical discharge summaries, human annotators agreed with each other more than 80% of the time.

The systems that participated in the smoking challenge represented various approaches from supervised and unsupervised classifiers to handcrafted rules. Despite the differences in their approaches to smoking status identification, many of these systems produced good results. In particular, there were 12 system runs with microaveraged F-measures above 0.84. A majority of these 12 systems took advantage of the idiosyncrasies of the challenge data, e.g., lack of references to smoking in records marked Unknown, and/or made use of explicit references to smoking. Collectively, the systems in the smoking challenge showed that discharge summaries express smoking status using a limited number of key textual features (e.g., "smok", "tobac", "cigar", Social History, etc.). Many of the effective smoking status identifiers benefit from these features.

References ■

1. Chang JT, Altman RB. Promises of text processing: natural language processing meets AI. Drug Discov Today 2002;7:992–3.
2. Lovis C, Baud RH. Fast exact string pattern matching algorithms adapted to the characteristics of the medical language. J Am Med Inform Assoc 2000;7:378–91.
3. 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;1:161–74.
4. Haug PJ, Koehler S, Lau LM, Wang P, Rocha R, Huff SM. Experience with a mixed semantic/syntactic parser. Proc Annu Symp Comput Appl Med Care 1995;19:284–8.
5. Goryachev S, Sordo M, Zeng QT. A suite of natural language processing tools developed for the i2b2 project. AMIA Annu Symp Proc 2006:931.
6. Bates W, Evans RS, Murff H, Stetson PD, Pizziferri L, Hripcsak, G. Detecting adverse events using information technology. J Am Med Inform Assoc 2003;10:115–28.
7. Westbrook JI, Coiera EW, Gosling AS. Do online information retrieval systems help experienced clinicians answer clinical questions? J Am Med Inform Assoc 2005;12:315–21.
8. Melton GB, Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. J Am Med Inform Assoc 2005;12:448–57.
9. Uzuner Ö, Luo Y, Szolovits P. Evaluating the state of the art in automatic de-identification. J Am Med Inform Assoc 2007;14:550–63.
10. Grishman R, Sundheim B. Message Understanding Conference 6: a brief history. 16th Conference on Computational Linguistics; 1996. Copenhagen, Denmark: Association for Computational Linguistics, 1996, pp 466–71.
11. NIST. 2005. Available at: http://www.nist.gov/speech/tests/. Accessed February 15, 2007.
12. Krallinger M. BioCreAtIvE. 2006. Available at: http://biocreative.sourceforge.net/. Accessed February 15, 2007.
13. Hersh WR, Muller H, Jensen JR, Yang J, Gorman PN, Ruch P. Advancing biomedical image retrieval: development and analysis of a test collection. J Am Med Inform Assoc 2006;13:488–96.
14. Braschler M, Peters C. Cross language evaluation forum: objectives, results, achievements. Information Retrieval 2004;7:7–31.
15. Hersh W, Bhupatiraju RT, Corley S. Enhancing access to the Bibliome: the TREC Genomics track. MedInfo 2004;11:773–7.
16. Sparck Jones K. Reflections on TREC. Info Process Manage 1995;31:291–314.
17. Hirschman L. The evolution of evaluation: lessons from the message understanding conferences. Comput Speech Lang 1998;12:281–305.
18. Hirschman L, Yeh A, Blaschke C, Valencia A. Overview of BioCreAtIvE: critical assessment of information extraction for biology. BMC Bioinform 2005;6:S1.

19. Chapman WW, Christensen LM, Wagner MM, et al. Classifying free text triage chief complaints into syndromic categories with natural language processing. Artif Intell Med 2005;33:31–40.
20. Chapman WW, Dowling JN, Wagner MM. Classification of emergency department chief complaints into 7 syndromes: a retrospective analysis of 527,228 patients. Ann Emerg Med 2005;46:445–55.
21. Chute CG. Clinical classification and terminology: some history and current observations. J Am Med Inform Assoc 2000;7:298–303.
22. Huang Y, Lowe HJ. A grammar based classification of negations in clinical radiology reports. Proc AMIA Annu Fall Symp 2005:988.
23. Sibanda T, He T, Szolovits P, Uzuner Ö. Syntactically informed semantic category recognizer for discharge summaries. Proc AMIA Annu Fall Symp 2006:714–8.
24. 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.
25. Hellesø R. Information handling in the nursing discharge note. J Clin Nurs 2006;15:11–21.
26. Hripcsak G, Zhou L, Parsons S, Das AK, Johnson SB. Modeling electronic discharge summaries as a simple temporal constraint satisfaction problem. J Am Med Inform Assoc 2005;12:55–63.
27. Kukafka R, Bales ME, Burkhardt A, Friedman C. Human and automated coding of rehabilitation discharge summaries according to the international classification of functioning, disability, and health. J Am Med Inform Assoc 2006;13:508–15.
28. Liu H, Friedman C. CliniViewer: a tool for viewing electronic medical records based on natural language processing and XML. MedInfo 2004;11:639–43.
29. Zhou L, Melton GB, Parsons S, Hripcsak G. A temporal constraint structure for extracting temporal information from clinical narrative. J Biomed Inform 2006;39:424–39.
30. Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus 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.
31. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20:37–46.
32. What is Kappa. November 25, 2000. Available at: http://www.musc.edu/dc/icrebm/kappa.html. Accessed February 13, 2007.
33. Hripcsak G, Heitjan DF. Measuring agreement in medical informatics reliability studies. J Biomed Inform 2002;35:99–110.
34. Krippendorff K. Content Analysis: An Introduction to Its Methodology. 2nd ed. Thousand Oaks, California: Sage, 2004.
35. Congalton RG. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sensing Environ 1991;37:35–46.
36. Yang Y, Liu X. A re-examination of text categorization methods. In: Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, California. New York: ACM Press, 1999, pp 42–49.
37. Salton G, McGill MJ. Introduction to Modern Information Retrieval. New York: McGraw Hill, 1983.
38. Hripcsak G, Rothschild AS. Agreement, the F measure, and reliability in information retrieval. J Am Med Inform Assoc 2005;12:296–8.
39. Chinchor N. The Statistical Significance of the MUC 4 Results. Fourth Message Understanding Conference (MUC 4), McLean, Virginia. Morristown, NJ: Association for Computational Linguistics, 1992.
40. 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.
41. Carrero FM, Gómez Hidalgo JM, Puertas E, Maña M, Mata J. Quick Prototyping of High Performance Text Classifiers. i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, 2006.
42. 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;15:36–9.
43. Xu H, Markatou M, Dimova R, Liu H, Friedman C. Machine learning and word sense disambiguation in the biomedical domain: design and evaluation issues. BMC Bioinformatics 2006;7:334–50.
44. Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Techniques. 2nd ed. San Francisco: Morgan Kaufmann, 2005.
45. Cohen AM. Five-way smoking status classification using text hot-spot identification and error-correcting output codes. J Am Med Inform Assoc 2008;15:32–5.
46. Gospodnetic O, Hatcher E. Lucene in Action. Greenwich, Conn: Manning Publications, 2005.
47. Guillen R. Automated De-identification and Categorization of Medical Records. i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data. 2006.
48. Pedersen T. Determining Smoker Status using Supervised and Unsupervised Learning with Lexical Features. i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data. 2006. Available as JAMIA on-line data supplement to the current article, at www.jamia.org.
49. Rekdal M. Identifying Smoking Status Using Argus MLP. i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data. 2006.
50. 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;15:25–8.
51. Heinze DT, Morsch ML, Potter BC, Sheffer RE. A-Life Medical I2B2 NLP smoking challenge system architecture and methodology. J Am Med Inform Assoc 2008;15:40–3.
52. 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.
53. 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.
54. Miller RA. Reference standards in evaluating system performance. J Am Med Inform Assoc 2002;9:87–8.