Pneumonia identification using statistical feature selection

Cosmin Adrian Bejan,[1] Fei Xia,[1,2] Lucy Vanderwende,[1,3] Mark M Wurfel,[4] Meliha Yetisgen-Yildiz[1,2]

► An additional appendix is published online only. To view this file please visit the journal online (www.jamia.org/content/early/recent).

[1]Department of Biomedical and Health Informatics, School of Medicine, University of Washington, Seattle, Washington, USA
[2]Department of Linguistics, University of Washington, Seattle, Washington, USA
[3]Microsoft Research, Redmond, Washington, USA
[4]Division of Pulmonary and Critical Care Medicine, School of Medicine, University of Washington, Seattle, Washington, USA

Correspondence to
Dr Cosmin Adrian Bejan, Department of Biomedical and Health Informatics, School of Medicine, University of Washington, 1959 NE Pacific Street, HSB I-264, Box 357240, Seattle, WA 98195-7240, USA; bejan@u.washington.edu

Received 5 December 2011
Accepted 22 March 2012

ABSTRACT

Objective This paper describes a natural language processing system for the task of pneumonia identification. Based on the information extracted from the narrative reports associated with a patient, the task is to identify whether or not the patient is positive for pneumonia.

Design A binary classifier was employed to identify pneumonia from a dataset of multiple types of clinical notes created for 426 patients during their stay in the intensive care unit. For this purpose, three types of features were considered: (1) word n-grams, (2) Unified Medical Language System (UMLS) concepts, and (3) assertion values associated with pneumonia expressions. System performance was greatly increased by a feature selection approach which uses statistical significance testing to rank features based on their association with the two categories of pneumonia identification.

Results Besides testing our system on the entire cohort of 426 patients (unrestricted dataset), we also used a smaller subset of 236 patients (restricted dataset). The performance of the system was compared with the results of a baseline previously proposed for these two datasets. The best results achieved by the system (85.71 and 81.67 F1-measure) are significantly better than the baseline results (50.70 and 49.10 F1-measure) on the restricted and unrestricted datasets, respectively.

Conclusion Using a statistical feature selection approach that allows the feature extractor to consider only the most informative features from the feature space significantly improves the performance over a baseline that uses all the features from the same feature space. Extracting the assertion value for pneumonia expressions further improves the system performance.

INTRODUCTION

The availability of comprehensive electronic medical records that include narrative reports provides an opportunity for natural language processing (NLP) technologies to play a major role in clinical research. One of the main advantages of employing these technologies is the automatic extraction of relevant clinical information to identify critical illness phenotypes and to facilitate clinical and translational studies of large cohorts of critically ill patients. Using NLP technologies in clinical research also has the advantage of solving problems which require the processing of a large number of narrative reports in real time. In contrast, this would not be feasible for the traditional approaches that use chart abstractions or bedside data acquisition since they are expensive, labor intensive, and involve many subjective assessments.

As part of a clinical research study, we designed an NLP system to automatically identify intensive care unit (ICU) patients with *pneumonia*. More exactly, using the narrative reports associated with each patient, the task is to analyze the information from these reports, and, based on this analysis, to classify the patient as positive or negative for pneumonia. To solve this task, our system relies on a supervised machine learning framework that selects only the most relevant features extracted from the ICU reports. Selecting the most relevant features resulted in a significant improvement in the performance of the learning algorithm. In addition, discarding the irrelevant and redundant features reduced the dimensionality of the original feature space, and, as a result, decreased the computational cost of the classification task.

For evaluation, we used the same collection of narrative reports as Yetisgen-Yildiz *et al* used in their work.[1] Specifically, the set consists of various types of reports created by physicians during the patients' ICU stay. While most previous work used only radiology reports to identify patients with pneumonia, we believe that a full set of physician daily notes is a rich source of clinical information for accurately identifying complex illness phenotypes such as pneumonia. In contrast to the narrow scope of information provided by radiology reports, physician daily notes include text detailing the patient narrative, physiologic, imaging, and laboratory data, and, most importantly, the physician's interpretation of these data.

Although this study focuses only on pneumonia identification, our main research goal is to build automated NLP systems that are able to detect multiple critical illness phenotypes and to model their progression from ICU data.

BACKGROUND AND RELATED WORK

Over the last years, several NLP systems have demonstrated their utility in a variety of healthcare applications.[2 3] For instance, many hospitals are currently using NLP systems for pneumonia surveillance, because this type of application is resource intensive and, at the same time, requires real-time assessments. In this direction, automated methods for identifying different types of pneumonia have been widely studied, although all focus exclusively on radiology reports. As one of the earliest examples, Fiszman *et al* tested an NLP tool called SymText to identify acute bacterial pneumonia-related concepts in chest x-ray reports and compared its performance against human

annotations.[4] Their results indicated that the performance of SymText was comparable to that of the physician. The same research group used the concepts identified by SymText as features for automatically identifying chest x-ray reports that supported pneumonia.[5–8] Their results showed that a machine learning framework based on Bayesian networks performs as well as expert constructed systems and the manual annotations of a physician and a lay person.

Mendonca *et al* and Haas *et al* investigated the feasibility of using NLP approaches in identifying healthcare associated pneumonia in neonates from chest x-ray reports.[9 10] Their NLP approach involved two components: the MedLEE system and rules that access the MedLEE output. The rules were manually constructed by a medical expert to identify chest x-ray reports indicating the presence of pneumonia.

Elkin *et al* also applied NLP approaches to identify pneumonia cases in narrative radiology reports.[11] Their system encoded the radiology reports with SNOMED CT ontology and subsequently applied a set of manually constructed rules to the SNOMED CT annotations to identify the radiological findings and diagnoses related to pneumonia.

The previous studies outlined above have focused primarily on the identification of pneumonia cases from radiologist chest x-ray reports. While radiologic changes within the lung are a necessary condition for a diagnosis of pneumonia, there are data within other domains such as the disease presentation narrative, physiologic measures, and laboratory abnormalities that could add significant accuracy and depth to the identification of pneumonia cases.[12 13] Because chest x-ray abnormalities comprise only part of the pneumonia definition, any system for pneumonia identification which incorporates only chest x-ray information will therefore lead to significant phenotypic misclassification. Thus, there remains an unmet need to accurately capture the clinical components of the pneumonia phenotype.

Other clinical studies have used NLP methodologies for processing a mixture of clinical note types. For instance, Meyestre and Haug used various types of clinical documents such as radiology reports, diagnostic procedure reports, surgery reports, discharge summaries, and others, to identify 80 different medical problems.[14] For this purpose, they extracted the medical problems from the clinical documents with MetaMap[15] and assigned to each problem an assertion value using Negex.[16] Meyestre and Haug reported a significant increase in recall when using MetaMap with a restricted subset of Unified Medical Language System (UMLS) concepts. In another study, the MediClass system[17] was employed to enable the assessment of smoking-cessation care delivery.[18] The role of MediClass was to evaluate a recommended treatment model proposed by a group of clinicians and tobacco-cessation experts, which involves five steps, the 5A's: (i) ask about smoking status; (ii) advise patients to quit smoking; (iii) assess a patient's willingness to quit; (iv) assist the patient's quitting efforts; and (v) arrange follow-up. Using various forms of clinical data, which include progress notes, patient instructions, medication data, referrals, visit reasons, and other smoking-related data, the MediClass system was able to achieve results similar to that of a trained human coder.

METHOD

In this section, we describe our approach to pneumonia identification. First, we introduce the dataset used for this study, and then we present the system architecture as well as the features employed for this task.

Table 1 Corpus statistics by the frequency of report types and the number of distinct patients who had the report type

Report type	Reports	Patients
Admit note	481	280
ICU daily progress note	2526	388
Acute care daily progress note	1357	203
Interim summary	164	115
Transfer/transition note	243	175
Transfer summary	18	18
Cardiology daily progress note	133	17
Discharge summary	391	350

Dataset

The dataset consists of narrative reports for 426 patients. Initially, the annotations of this dataset were performed for another study of ICU subjects which was described previously.[19] A research study nurse with 6 years of experience manually annotated a patient as *positive* if the patient had pneumonia within the first 48 h of ICU admission and as *negative* if the patient did not have pneumonia or the pneumonia was detected after the first 48 h of ICU admission. As a result, 66 patients were identified as cases positive for pneumonia and the remaining 360 patients as negative cases. Moreover, because subjects in this dataset were admitted to the ICU from the emergency department as well as from other hospitals, cases of pneumonia included both community acquired pneumonia (ie, pneumonia acquired outside of the hospital settings) and hospital acquired pneumonia (ie, pneumonia acquired after admission to hospital). Overall, our dataset includes a total of 5313 reports, each report having one of the eight report types: admit note, ICU daily progress note, acute care daily progress note, transfer/transition note, transfer summary, cardiology daily progress note, and discharge summary. The total number of reports per patient ranged widely due to the high variability in the ICU length of stay: median 8, IQR 5–13, minimum 1, maximum 198.

Table 1 shows the distribution of reports and patients among the eight different report types from the dataset. The second column of the table gives the number of reports for each report type, while the third column shows the number of distinct patients who had the report type in the dataset. As can be seen from the table, not all patients have all report types. For example, only 280 (65%) patients have admit notes; the remaining 146 patients who have no admit notes are likely to have been transferred to the ICU from other medical units. There were 350 (82%) patients with discharge summaries. Out of the 426 patients, only a subset of 236 (55%) patients had both admit notes and discharge summaries. We will later use this subset of patients as the restricted dataset.

It is also worth noting that, of the admit notes collected in this corpus, over 75% derive from the first day of hospital admission. Furthermore, ICU progress notes were consistently generated throughout the first 96 h of admission. These data show that admit notes will generally document the pre-hospital situation and early hospital stay, while ICU progress notes will be the predominant daily text source following the day of admission. Discharge notes largely arose after 96 h since admission.

System architecture

The main components of our system architecture are depicted in figure 1. As illustrated, we designed the architecture on top of a supervised machine learning framework, where features

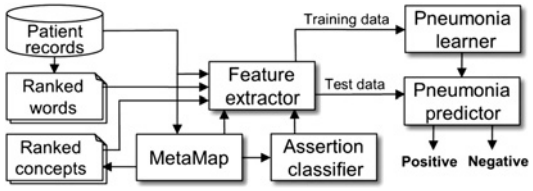

Figure 1 System architecture for pneumonia identification.

associated with each data instance (ie, patient) are automatically extracted to be used by a binary classifier. For classification, we employed LIBLINEAR,[20] an implementation of support vector machines. In this framework, we represented each patient by all of the corresponding reports. To extract features for each patient, in the preprocessing phase, we first tokenized the reports with SPLAT[21] and filtered out the punctuation tokens.

Unlike a conventional learning framework, however, we also implemented a methodology to select only the most informative features for pneumonia identification. Specifically, this methodology uses statistical significance tests to measure the association strength between each feature from the training set and the two categories of this task (ie, positive and negative pneumonia). As a result, the features will be ranked based on those values such that those with a strong association with the two categories will be on top. Finally, only the most relevant features that are within a specific threshold will be selected for training. This methodology is also called *statistical feature selection*.

Statistical feature selection

Feature selection algorithms have been successfully applied in text categorization in order to improve the classification accuracy. By significantly reducing the dimensionality of the feature space, they also improve the efficiency of the classifiers and provide a better understanding of the data.[22–24] Since our task is very similar to text categorization, using a feature selection approach is recommended. Indeed, representing each patient by a large document containing all the patient's reports, we can cast the pneumonia identification problem as the problem of categorizing these large documents into positive or negative pneumonia categories.

Before learning a model for pneumonia identification, we first built lists of ranked features from the training set. For this purpose, we considered as features all possible uni-grams and bi-grams of words and UMLS concepts. We identified the UMLS concepts in our dataset by using MetaMap. In order to rank the set of features associated with a feature type as illustrated in figure 1 (eg, the ranked list of word bi-grams), we constructed a contingency table for each feature from the set and used statistical hypothesis testing to determine whether there is an association between the feature and the two categories of our problem. Specifically, we computed the χ^2 and t statistics,[25] which generate two different orderings for each feature set. The reason for generating two different orderings is based on the fact that the t test assumes that the event of a feature being associated with a specific category is sampled from a normal distribution whereas the χ^2 test does not use this assumption.

Table 2 lists the top 10 uni-grams, bi-grams, and UMLS concepts ranked by these statistical tests. As can be observed, many of these features are closely linked to the known causes (eg, *influenza*) and clinical signs and symptoms (eg, *sputum*, *coughs*, *decreased breath*) of pneumonia. However, this table may also list features that are not directly related to the diagnostic

Table 2 The top 10 most informative uni-grams, bi-grams, and Unified Medical Language System (UMLS) concepts for pneumonia identification according to χ^2 and t statistics

Uni-gram	Bi-gram	UMLS concept
χ^2 Statistic		
col	sputum cx	Microbial culture of sputum
tan	sputum culture	Sputum
coughs	h1n1 positive	Fluorescence Units
pneumo	acquired pneumonia	Structure of middle lobe of lung
sputum	positive h1n1	Influenza preparation
consolidation	pneumonia continue	H1N1
stacking	continue lpv	Influenza virus vaccine
cart	coarse mechanical	Novel H1N1 influenza
proning	continue oseltamivir	Oseltamivir
coccyx	treatment pneumonia	Infiltration
t Statistic		
sputum	sputum cx	Microbial culture of sputum
suctioning	sputum culture	Sputum
h1n1	continue lpv	Consolidation
ventilatory	h1n1 influenza	Infiltration
consolidation	acquired pneumonia	Influenza preparation
secretions	bacterial pneumonia	Influenza virus vaccine
lpv	continue oseltamivir	Pneumonia
coughing	decreased breath	Fluorescence Units
flu	positive h1n1	Influenza
tachypneic	urine sputum	Decreased breath sounds

criteria for pneumonia since they may well indicate latent risk factors for pneumonia or simply capture a predominant association with one of the two categories (eg, *urine sputum, coccyx*). Report writing is often formulaic, so there can be a preponderance of evidence for the order of, for example, *urine* and *sputum* as we see in *'urine, sputum and blood remain negative,' 'blood, urine, sputum cultures have not grown pathogenic organisms to date,' 'urine, sputum, blood cultures pending,'* etc.

Once all features are ranked and their corresponding threshold values are established, the feature extractor is now able to build a feature vector for each patient. Specifically, given a fixed subset of relevant features determined by selecting the top features from the ranked lists of features up to a threshold value, the feature extractor considers in the representation of a patient's feature vector only the features from the subset of relevant features that are also found in the patient's reports. Therefore, the size of the feature space will be equal to the size of the relevant features subset, whereas the length of each feature vector will be at most this value.

Because the χ^2 and t significance tests will generate two different feature rankings for every feature set, the two feature subsets extracted from these rankings at any given threshold value will also be different. Therefore, an interesting experiment will be to choose relevant features from the union of the two feature subsets (χ^2+t) in order to see the contribution of both statistical tests to the overall system performance. To determine whether this is a feasible experiment, we first studied how many features each pair of feature subsets will have in common. As an example, the χ^2 and t bi-gram subsets listed in table 2, at a common threshold value of 10, have 60% of features in common. For a more elaborate overview, figure 2 shows the percentage of words and UMLS concepts shared by the two statistics for all possible threshold values. As can be seen from this figure, the lists of words and UMLS concepts considered have approximately 58 000 and 20 000 features, respectively.

In our experiments, we also measured the association between a feature and the two categories using the Fisher exact test,

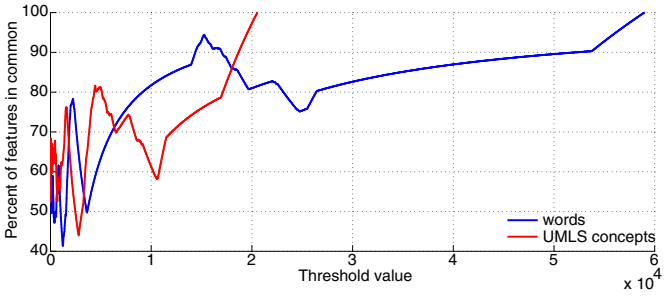

Figure 2 Words and Unified Medical Language System (UMLS) concepts shared in common by the χ^2 and *t* lists. The horizontal axis specifies the threshold values that determine the size of the feature subsets extracted from the χ^2 and *t* lists. The vertical axis shows the percentage of features these subsets have in common.

pointwise mutual information, and the Dice coefficient; however, these measures did not perform as well as the χ^2 and *t* statistical tests.

Assertion of pneumonia expressions

There are cases where the ICU reports include an explicit statement that the patient has, is suspected of, or does not have pneumonia. To account for these cases, in addition to the features selected using statistical tests, we implemented a binary feature, called *the assert feature*, which assigns to each patient a label corresponding to positive or negative pneumonia. The assignment of this label is based on the assertion values associated with pneumonia expressions and their related words (eg, pneumonitis) found in the patient's reports.

We associated an assertion value with a pneumonia expression by training a maximum entropy classifier on the dataset provided by the 2010 i2b2/VA challenge for assertion classification.[26] The purpose of this task is to classify the assertion value of a medical concept expressed in a free-text report as *present*, *absent*, *possible*, *conditional*, *hypothetical*, or *associated with someone else*. Using simple lexical features that explore the surrounding context of a medical concept in text as well as features extracted with the NegEx and ConText[27] tools, our classifier achieves a satisfactory 92.79 micro-averaged F1-measure.

The extraction of pneumonia expressions from our dataset was performed by first parsing the reports with MetaMap and then selecting only the identified medical phrases that have the same identifier as pneumonia (CUI:C0032285) in the UMLS Metathesaurus. For a more complete set, we also ran simple regular expressions to identify the word *pna*, an abbreviation often used by physicians in clinical reports for pneumonia but which is not yet tagged as a pneumonia concept in the UMLS Metathesaurus. After we ran the assertion classifier for all the pneumonia concepts of a patient, we counted how many times each of the six assertion values were identified, and then mapped the most frequent value to one of the two categories of pneumonia identification. We found that a good mapping of the assertion values is (*present*) → positive pneumonia, and (*absent, possible, conditional, hypothetical, associated with someone else*) → negative pneumonia. For those binary features corresponding to patients with no pneumonia concepts identified in their reports (223 out of 426), we assigned a default value of negative pneumonia.

RESULTS

Since we employed the same dataset as used by Yetisgen-Yildiz *et al*, we considered that work as the baseline for our system. This system consists of a supervised learning framework, where the feature vector corresponding to a patient is represented as a 'bag of words' built from the patient's reports.[1] Yetisgen-Yildiz *et al* experimented with different representations of the patient data but concluded that using all of the report types for feature generation gave the best performance, and consequently, we use only the representation comprised of all report types. As features, they considered various combinations of word n-grams, UMLS concepts, and their corresponding semantic types. Besides performing experiments on the entire cohort of 426 patients (the unrestricted dataset), they also considered a smaller subset of 236 patients restricted to those with both an admit note and a discharge summary (the restricted dataset). Using a fivefold cross-validation scheme, their system achieved the best results of 50.7 and 49.1 F1-measure on the restricted and unrestricted dataset, respectively, when the entire set of word uni-grams was considered. For an accurate comparison, in our experiments we used the same folds of the two datasets as Yetisgen-Yildiz *et al*.

Experimenting with word n-grams

In a first set of experiments, we studied how the performance of our system evolves for various threshold values on the χ^2, *t*, and χ^2+*t* ranked word lists. Figure 3 shows the results of these experiments. The plots at the top correspond to the results found for the restricted dataset, while the plots at the bottom show the results for the unrestricted dataset. The results are computed using the F1-measure, which represents the harmonic mean of precision and recall. For each experiment, we considered 27 different values that capture the threshold variation for selecting from a range of 10—40 000 significant word n-grams. For instance, if the feature extractor selects the first 30 features of the *t* word lists built from the training set of the restricted corpus, our system will achieve 63.25 F1-measure (threshold=30 in the top left plot for the *t* experiment). For a clear understanding of our experimental setup, we would like to emphasize that by selecting the first 30 features from the *t* word lists, we actually consider the union of the first 30 word uni-grams and the first 30 word bi-grams, for a total feature space of at most 60 features. We consider a similar setup when including the UMLS concept lists. We also consider different threshold values for the UMLS concept lists since the sizes of these lists are considerably smaller than the sizes of word n-gram lists.

As can be observed in figure 3, our system shows some fluctuation in performance across the range of threshold values. This fluctuation implies that the order imposed by the two statistical tests does not guarantee a perfect order for selecting features for pneumonia identification. Therefore, noisy features do exist at the top of the lists ranked by these statistical tests that have a negative impact on our system performance; conversely, there are relevant features toward the bottom of the ranked lists which can improve the performance of our system. However, two clear observations from the plots drawn in figure 3 indicate that the statistical tests group most of the relevant features toward the beginning of their corresponding lists. First, most of the results obtained in all the experiments significantly outperform the baseline results. For instance, the best results on the restricted and unrestricted datasets were achieved by the *t* experiment (80.95 F1-measure when considering the first 10 000 features) and by the χ^2+*t* experiment (75.63 F1-measure when considering the first 800 features), respectively. And second, the results corresponding to all the experiments considered reached a plateau for the last threshold values. In fact, the results of the last threshold values are very close to the baseline results since, in these cases, the feature extractor selects almost the entire feature set.

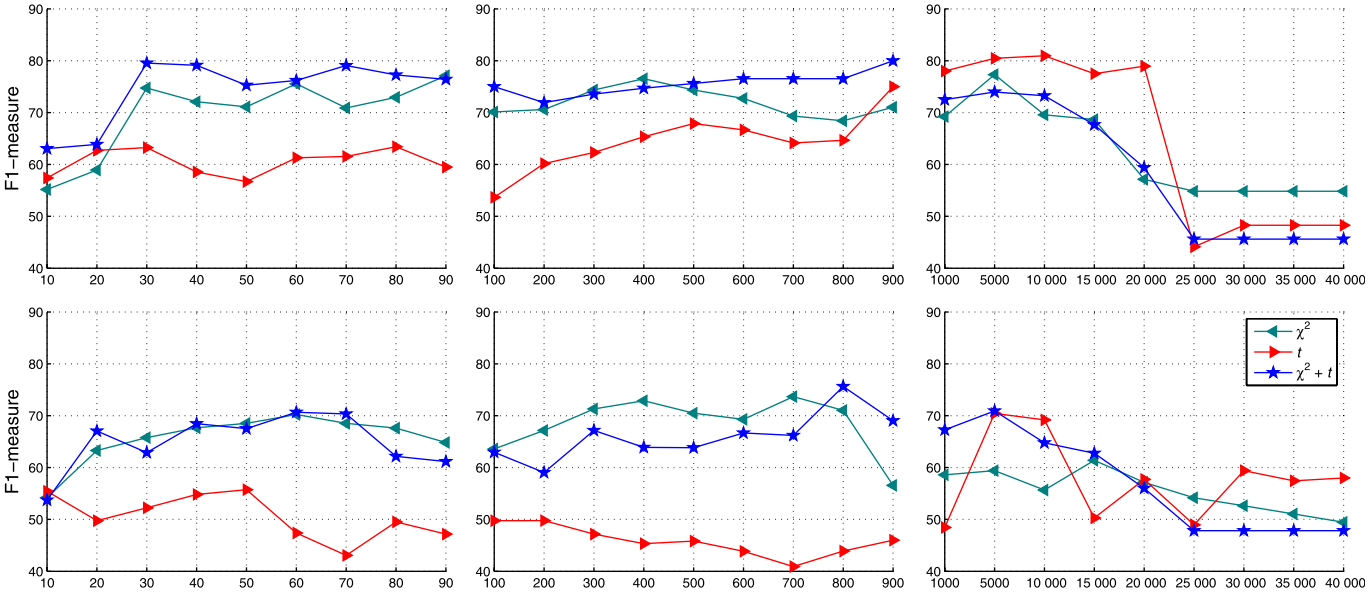

Figure 3 Performance results for various subsets of relevant features selected from the χ^2, t, and χ^2+t ranked lists of word n-grams. The size of each subset is indicated by the threshold values on the horizontal axis of each plot. On the vertical axes, the system performance is expressed in terms of F1-measure. The plots at the top correspond to results on the restricted dataset, while the plots at the bottom correspond to results on the unrestricted dataset.

Experimenting with all feature types

Figure 4 shows the results of a second set of experiments where we evaluated the impact of combining the word n-gram features with the assert feature and UMLS concepts. For clarity in the plots, we considered only the experiments involving the χ^2+t ranked lists. In this figure, we used a threshold value of 100 for the selection of the concept n-grams. As observed, the best results in these plots correspond to the experiment that combines all the feature types. To get a better insight into the results corresponding to various combinations of feature types, we computed the micro-averaged precision (microP), recall (microR), and F1-measure (microF$_1$) over all 27 threshold values associated with each experiment. For instance, the microF$_1$ values of the χ^2+t, χ^2+t+assert, χ^2+t+concepts, and χ^2+t +concepts+assert experiments on the restricted dataset are 70.76, 72.35, 71.65, and 75.97, respectively. Additional results and a more detailed ablation study (at the feature type level) are presented in the supplementary online appendix. The online appendix also lists, for each experiment, the best and worst F1 results (denoted as maxF$_1$ and minF$_1$, respectively) over the 27 threshold values for the word n-gram lists as well as experiments considering various threshold values for the concept n-gram lists.

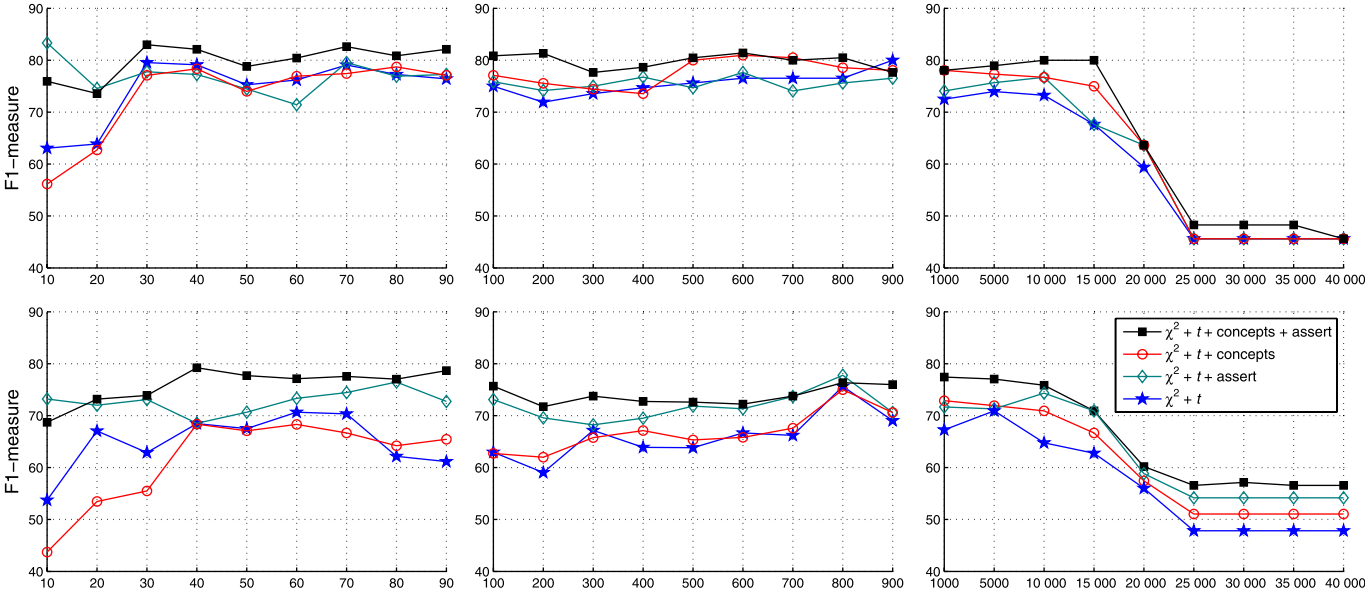

Figure 4 A study on the impact of the performance results when considering various combinations of word n-grams, Unified Medical Language System (UMLS) concepts, and the assert feature. For each experimental result, the number of relevant features selected from the χ^2+t lists of word n-grams is indicated by the threshold values on the horizontal axis of each plot. On the vertical axes, the system performance is expressed in terms of F1-measure. The threshold value for selecting the most relevant UMLS concepts is set to 100 in all the experiments from these plots. The plots at the top correspond to results on the restricted dataset, while the plots at the bottom correspond to results on the unrestricted dataset.

Table 3 The best performing results for each combination of feature types on the restricted dataset

Feature set	Test	w_{th}	uc_{th}	TP	FP	FN	TN	P	R	NPV	Spec	Acc	F_1
Baseline				17	6	27	186	73.90	38.60	87.32	96.90	86.00	50.70
concepts	χ^2	–	300	37	15	7	177	71.15	84.09	96.20	92.19	90.68	77.08*
concepts+assert	χ^2	–	300	37	12	7	180	75.51	84.09	96.26	93.75	91.95	79.57**
Words	*t*	10 000	–	34	6	10	186	85.00	77.27	94.90	96.88	93.22	80.95**
words+assert	χ^2+*t*	10	–	40	12	4	180	76.92	**90.91**	**97.83**	93.75	93.22	83.33**
words+concepts	*t*	10 000	50	37	8	7	184	82.22	84.09	96.34	95.83	93.64	83.15**
words+concepts+assert	*t*	10 000	5000	36	4	8	188	**90.00**	81.82	95.92	**97.92**	**94.92**	**85.71****

*p<0.01, **p<0.001; statistically significant differences in performance between the system configurations considered and the baseline.
Acc, accuracy; F_1, F1-measure; FN, false negatives; FP, false positives; NPV, negative predictive value; P, precision; R, recall; Spec, specificity; TN, true negatives; TP, true positives.

In the online appendix, the threshold value for the UMLS concepts is denoted as uc_{th}.

Finally, tables 3 and 4 list the best performing results for each combination of feature types for the restricted and unrestricted datasets, respectively. The results are reported in terms of true positives (TP), false positives (FP), false negatives (FN), true negatives (TN), precision (P), recall (R), negative predictive value (NPV), specificity (Spec), accuracy (Acc), and F1-measure (F_1). The best results associated with each performance metric from these two tables are emphasized in boldface. For each experiment that involves the word and UMLS concept features, the tables also list the threshold values corresponding to the best F1-measure (the primary measure considered for pneumonia identification). These two thresholds are denoted in the tables as w_{th} and uc_{th}. To determine whether the differences in performance between our proposed system configurations and the baseline are statistically significant, we employed a randomization test based on stratified shuffling.[28] As can be seen, the results produced by our system significantly outperform the baseline results. On both datasets, the best results were achieved when combining all three feature types. In particular, on the restricted dataset, the *t* experiment peaks at w_{th}=10 000 and uc_{th}=5000, whereas, on the unrestricted dataset, the χ^2 experiment has the best F1-measure when w_{th}=600 and uc_{th}=5000. In addition to the baseline considered, we also developed a rule-based system using the value assigned to each patient by the assert feature. Although this rule-based system achieved better results than the baseline (57.14 and 54.03 F1-measure on the restricted and unrestricted dataset, respectively), the differences in performance are not statistically significant. This is because both datasets have a relatively small number of instances. Of note as well, the rule-based system is significantly inferior to any of our systems that include machine learning.

Error analysis

There are several important limitations to our current dataset. First, it is not a complete set of reports for all patients (eg, none of the notes entered prior to the day of ICU admission were captured). Of the total of 426 patients, 35% did not have admit notes, 18% did not have discharge summaries, and 18% did not have any reports generated in the first 24 h of hospital stay. Second, the pneumonia annotation in this dataset was created for a different purpose, where the annotator (a medical expert) was asked to determine whether a patient had pneumonia within 48 h of admission to the ICU. As such, positive cases cover both community and hospital acquired pneumonia, which is confusing to the learner when it encounters a negative training example of hospital acquired pneumonia post-48 h of admission to the ICU. Third, the dataset is relatively small with a limited number of positive pneumonia cases.

Another source of errors is due to the limitations of our current system, which relies on features available from shallow processing of the text. The detection of a phenotype such as pneumonia often requires a deeper understanding of the reports that goes beyond assertion identification. For instance, for the reports that do not explicitly mention that the patient has pneumonia, one important decision factor is given by the pneumonia-related symptoms and lab results encoded in the ICU reports. However, lab results and other structured data that may be relevant for pneumonia identification are often not included in these reports. While the objective of the current study was to explore the text reports using NLP technologies, in our future work we plan to use a dataset that also includes informative structured data and radiology reports.

CONCLUSION

We presented a machine learning framework that is able to learn a model for pneumonia identification from narrative ICU reports. Statistical feature selection plays an important role in this framework, ranking features using statistical significance tests. As a result of this ranking, only the most relevant features for pneumonia identification will be selected by the feature extractor. We empirically proved that, by using this feature selection approach and considering only a small subset of informative features from the feature space, we can achieve significantly better results than a baseline which uses all the features from the same feature space. Consequently, this methodology significantly reduces the original feature space and

Table 4 The best performing results for each combination of feature types on the unrestricted dataset

Feature set	Test	w_{th}	uc_{th}	TP	FP	FN	TN	P	R	NPV	Spec	Acc	F_1
Baseline				28	20	38	340	58.30	42.40	89.95	94.40	86.40	49.10
concepts	χ^2	–	900	45	12	21	348	78.95	68.18	94.31	96.67	92.25	73.17**
concepts+assert	χ^2	–	1000	50	10	16	350	83.33	**75.76**	**95.63**	97.22	93.90	79.37**
words	χ^2+*t*	800	–	45	8	21	352	84.91	68.18	94.37	97.78	93.19	75.63**
words+assert	χ^2+*t*	800	–	49	11	17	349	81.67	74.24	95.36	96.94	93.43	77.78**
words+concepts	χ^2	700	50	49	7	17	353	87.50	74.24	95.41	98.06	94.37	80.33**
words+concepts+assert	χ^2	600	5000	49	5	17	355	**90.74**	74.24	95.43	**98.61**	**94.84**	**81.67****

**p<0.001; statistically significant differences in performance between the system configurations considered and the baseline.
Acc, accuracy; F_1, F1-measure; FN, false negatives; FP, false positives; NPV, negative predictive value; P, precision; R, recall; Spec, specificity; TN, true negatives; TP, true positives.

is able to discard most of the irrelevant and redundant features for the task of pneumonia identification. Furthermore, we showed that the addition of a feature that extracts the assertion value of all pneumonia expressions from our dataset improves the performance of our system for this task.

Contributors CAB designed and implemented the NLP system described in the paper, and extracted the experimental results achieved by this system. CAB also wrote 70% of the draft manuscript. FX contributed to the experimental setup of the feature selection algorithms, and consulted on the machine learning framework of the baseline system. LV contributed to the design of the assertion classifier, and provided assistance in using the SPLAT toolkit. MW had a significant contribution in the data annotation process, and validated the relevant features extracted by the feature selection algorithms. MYY had a major contribution in designing and implementing the baseline system, ran the baseline experiments, extracted the corpus statistics, and performed the error analysis. MYY also wrote the related work, dataset, and error analysis sections of the draft manuscript. All authors contributed to the final manuscript.

Funding This research was supported in part by P50 HL073996, RC2 HL101779, the Northwest Institute for Genetic Medicine, and Microsoft Research Connections.

Competing interests None.

Ethics approval Ethics approval was provided by the University of Washington Human Subjects Committee/Institutional Review Board.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **Yetisgen-Yildiz M,** Glavan BJ, Xia F, *et al*. Identifying patients with pneumonia from free-text intensive care unit reports. *Proceedings of Learning from Unstructured Clinical Text Workshop of ICML. Int Conf Mach Learn*. 2011.
2. **Hripcsak G,** Friedman C, Alderson PO, *et al*. Unlocking clinical data from narrative reports: a study of natural language processing. *Ann Intern Med* 1995;**122**:681—8.
3. **Demner-Fushman D,** Chapman WW, McDonald C. What can natural language processing do for clinical decision support? *J Biomed Inform* 2009;**42**:760—72.
4. **Fiszman M,** Chapman WW, Evans SR, *et al*. Automatic identification of pneumonia related concepts on chest x-ray reports. *Proc AMIA Symp* 1999:67—71.
5. **Chapman WW,** Haug PJ. Comparing expert systems for identifying chest x-ray reports that support pneumonia. *Proc AMIA Symp* 1999:216—20.
6. **Fiszman M,** Chapman WW, Aronsky D, *et al*. Automatic detection of acute bacterial pneumonia from chest x-ray reports. *J Am Med Inform Assoc* 2000;**7**:593—604.
7. **Chapman WW,** Fiszman M, Chapman BE, *et al*. A comparison of classification algorithms to automatically identify chest x-ray reports that support pneumonia. *J Biomed Inform* 2001;**34**:4—14.
8. **Aronsky D,** Fiszman M, Chapman WW, *et al*. Combining decision support methodologies to diagnose pneumonia. *Proc AMIA Symp* 2001:12—16.
9. **Mendonca EA,** Haas J, Shagina L, *et al*. Extracting information on pneumonia in infants using natural language processing of radiology reports. *J Biomed Inform* 2005;**38**:314—21.
10. **Haas JP,** Mendonca EA, Ross B, *et al*. Use of computerized surveillance to detect nosocomial pneumonia in neonatal intensive care unit patients. *Am J Infect Control* 2005;**33**:439—43.
11. **Elkin PL,** Froehling D, Wahner-Roedler D, *et al*. NLP-based identification of pneumonia cases from free-text radiological reports. *Proc AMIA Symp* 2008:172—6.
12. **Lutfiyya MN,** Henley E, Chang LF, *et al*. Diagnosis and treatment of community-acquired pneumonia. *Am Fam Physician* 2006;**73**:442—50.
13. **Mandell LA,** Wunderink RG, Anzueto A, *et al*. Infectious Diseases Society of America/American Thoracic Society consensus guidelines on the management of community-acquired pneumonia in adults. *Clin Infect Dis* 2007;**44**(Suppl 2):S27—72.
14. **Meyestre S,** Haug PJ. Natural language processing to extract medical problems from electronic clinical documents: performance evaluation. *J Biomed Inform* 2006;**39**:589—99.
15. **Aronson AR.** Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp* 2001:17—21.
16. **Chapman WW,** Bridewell W, Hanbury P, *et al*. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform* 2001;**34**:301—10.
17. **Hazlehurst B,** Frost HR, Sittig DF, *et al*. MediClass: a system for detecting and classifying Encounter-based clinical events in any electronic medical Record. *J Am Med Inform Assoc* 2005;**12**:517—29.
18. **Hazlehurst B,** Sittig DF, Stevens VJ, *et al*. Natural language processing in the electronic medical record: assessing clinician Adherence to tobacco treatment guidelines. *Am J Prev Med* 2005;**29**:434—9.
19. **Glavan BJ,** Holden TD, Goss CH, *et al*. Genetic variation in the FAS gene and associations with acute lung injury. *Am J Respir Crit Care Med* 2011;**183**:356—63.
20. **Fan RE,** Chang KW, Hsieh CJ, *et al*. LIBLINEAR: a library for large linear classification. *J Mach Learn Res* 2008;**9**:1871—4.
21. **Quirk C,** Choudhury P, Gao J, *et al*. MSR SPLAT, a language analysis toolkit. *In Proceedings of NAACL HLT 2012 Demonstration Session*. 2012.
22. **Yang Y,** Pedersen JO. A comparative study on feature selection in text categorization. *Proc Int Conf Mach Learn* 1997:412—20.
23. **Mladenić D,** Grobelnik M. Feature selection on hierarchy of web documents. *Decis Support Syst* 2003;**35**:45—87.
24. **Shang W,** Huang H, Zhu H, *et al*. A novel feature selection algorithm for text categorization. *Expert Syst Appl* 2007;**33**:1—5.
25. **Manning CD,** Schutze H. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press, 1999.
26. **Uzuner Ö,** South BR, Shen S, *et al*. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc* 2011;**18**:552—6.
27. **Chapman WW,** Chu D, Dowling JN. ConText: An algorithm for identifying contextual features from clinical text. *BioNLP 2007: Biological, Translational, and Clinical Language Processing*. 2007:81—8.
28. **Noreen E.** *Computer-intensive Methods for Testing Hypotheses*. New York: John Wiley & Sons, 1989.