Journal of Biomedical Informatics 52 (2014) 386–393

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Classification of CT pulmonary angiography reports by presence, chronicity, and location of pulmonary embolism with natural language processing

Sheng Yu [a,*], Kanako K. Kumamaru [b], Elizabeth George [b], Ruth M. Dunne [c], Arash Bedayat [b,d], Matey Neykov [e], Andetta R. Hunsaker [c], Karin E. Dill [f], Tianxi Cai [e], Frank J. Rybicki [b]

[a] Partners HealthCare Personalized Medicine, Brigham and Women's Hospital & Harvard Medical School, Boston, MA, United States
[b] Applied Imaging Science Laboratory, Department of Radiology, Brigham and Women's Hospital & Harvard Medical School, Boston, MA, United States
[c] Thoracic Imaging, Department of Radiology, Brigham and Women's Hospital & Harvard Medical School, Boston, MA, United States
[d] Department of Radiology, University of Massachusetts Medical School, Worcester, MA, United States
[e] Department of Biostatistics, Harvard School of Public Health, Boston, MA, United States
[f] Department of Radiology, University of Chicago, Chicago, IL, United States

ARTICLE INFO

Article history:
Received 1 April 2014
Accepted 1 August 2014
Available online 10 August 2014

Keywords:
Natural language processing
NILE
Nested modification structure
Pulmonary embolism
CT pulmonary angiography

ABSTRACT

In this paper we describe an efficient tool based on natural language processing for classifying the detail state of pulmonary embolism (PE) recorded in CT pulmonary angiography reports. The classification tasks include: PE present vs. absent, acute PE vs. others, central PE vs. others, and subsegmental PE vs. others. Statistical learning algorithms were trained with features extracted using the NLP tool and gold standard labels obtained via chart review from two radiologists. The areas under the receiver operating characteristic curves (AUC) for the four tasks were 0.998, 0.945, 0.987, and 0.986, respectively. We compared our classifiers with bag-of-words Naive Bayes classifiers, a standard text mining technology, which gave AUC 0.942, 0.765, 0.766, and 0.712, respectively.

1. Introduction

Patients with acute pulmonary embolism (PE) have a wide spectrum of clinical outcomes [1,2] with an overall mortality rate exceeding 10% [3]. CT pulmonary angiography (CTPA) directly visualizes emboli as filling defects and is the first-line imaging modality to evaluate patients with a clinical suspicion of acute PE [4]. As single center studies are rarely greater than 1000 patients [5–12] and do not account for variability between institutions, the establishment of large, multicenter, multinational databases will be advantageous to the research in PE. In order to facilitate a clinical study, positive PE cases need to be separated from the negative ones. Moreover, characteristics of PE, for example, the chronicity and the location of emboli, are also important information useful for clinical studies on subtypes of PE. Using diagnosis codes (e.g., ICD-9/ICD-10 codes) is one possible way to collect PE cases. However, the codes are not accurately reported to identify PE or to distinguish PE from deep vein thrombosis [13–16], and neither do they contain information on PE characteristics. To obtain detailed information on findings of PE, including its presence, as well as characteristics, we must rely on the description in the CTPA reports. However, extracting these information by manual chart review is too time-consuming and labor-intensive to be a viable approach to identify a large number of specific types of PE patients for research purposes. A reliable system that extracts information from CTPA reports automatically and accurately is expected to have significant impact on clinical research.

Natural language processing (NLP) is a promising technology for this information extraction task. The technology is not mature yet, as the prevalence of word sense ambiguity, shorthands (including symbols), and free mixing of language and semi-structured tables (plain text made to look tabular by using tabs, white spaces, or other symbols, whose semantics are fundamentally different from language) in general clinical notes still pose considerable difficulty to their interpretation [17–24]. However, diagnostic radiology reports are an ideal target for employing NLP, because they are composed almost purely of natural language, and they have a constrained vocabulary and a limited number of concepts for each imaging modality, so ambiguity is not very common.

A number of NLP software have been applied to radiology reports over the past thirty years. The Special Purpose Radiology Understanding System, focused on analyzing radiology reports were implemented at the LDS Hospital in Utah [25]. This was

* Corresponding author.
E-mail address: syu7@partners.org (S. Yu).

http://dx.doi.org/10.1016/j.jbi.2014.08.001
1532-0464/

followed by the Medical Language Extraction and Encoding System [26], focused initially on chest X-ray reports [27]. An information theory-based algorithm was validated on radiology reports [28]. A comparison of two specific algorithms [29] also confirmed that both machine learning and rule-based approaches, both popular in NLP, could perform well in analyzing radiology reports. Notably, their machine learning algorithm was trained on discharge summaries and showed excellent portability on radiology reports. Recent literature reports the application of NLP to identify the presence of PE from CTPA reports [30,31], the chronicity of PE, and the diagnostic quality of the exam [31].

It is hard not to notice that there has been no peer-reviewed report on the use of NLP to extract the most proximal location of the emboli in the pulmonary arterial tree. The location of PE is important. Massive central PE increases the risk of right ventricular overload and PE-related mortality [32,33], and the clinical significance and risk–benefit ratio of treatment of isolated subsegmental PE (ISSPE) has been questioned due to limited data on the natural history and outcomes of these relatively uncommon single peripheral clots [34,35]. One reason of this omission is that previous NLP technologies do not directly support such analysis. Several software, such as MedLEE [26] and cTAKES [36,37], can associate findings with locations if both are mentioned in a sentence. However, when radiologists describe, for example, a subsegmental embolus, arteries of segments and lobes are usually mentioned together, and previous technology do not allow us to distinguish whether the PE is subsegmental, segmental, or lobar. We built our program on top of the Narrative Information Linear Extraction (NILE) system [38]. NILE is an NLP library for semantic analysis of clinical narratives, and is developed upon the principles of linear interpretation using rules based on linguistic and clinical knowledge. One notable feature of NILE is that its location analyzer generates a nested modification structure that clarifies the relations among the multiple anatomical locations mentioned in a sentence. We utilized this feature as well as others to extract information of the presence, chronicity, and proximal extension of PE, and achieved satisfying accuracy.

The rest of the paper is organized as follows. Section 2 introduces the composition of our program for the extraction of information related to PE presence, chronicity and proximal extension, including the dictionary, variables, and algorithms that we used. Section 3 validates the classification output of our program against gold standard labels obtained from expert chart reviews, and compares its accuracy with the bag-of-words model — a standard and usually effective text mining approach. Section 4 discusses the advantages of the proposed tool for PE classification and the limitations of the current technology.

2. Method

2.1. Data

Our Institutional Review Board approved this HIPAA-compliant study; informed consent was waived. CTPA studies were identified using our hospital's electronic radiology information system, using the Current Procedural Terminology code for CTPA, and limiting the report type to "CT Pulmonary Embolism" and the report status to "Final" or "Revised". All consecutive 9413 retrieved CTPA examinations performed from 10/31/2006 to 3/31/2010, plus consecutive 917 CTPA examinations positive for PE (identified by the manual review of the CTPA reports) performed from 8/1/2003 to 10/30/2006 were included in the study. CTPA studies were performed by 16-, 64-, or 128-slice multi-detector CT scanners with a standard protocol using intravenous administration of 75–100 mL iodinated contrast media at 3–4 mL/s.

A CTPA was considered positive for PE when the official CT report confirmed or at least suggested the presence of PE. All reports that suggest the presence of PE but also mention limitations of images, e.g., poor opacification, motion artifact, increased noise, were considered positive for PE. The chronicity of PE (acute, subacute, chronic, acute on chronic, and unclassified) and the proximal extension of the embolus (central, lobar, segmental, or subsegmental pulmonary artery) were also determined based on the description in the Findings and Impressions sections of the CT reports. Briefly, the report was classified into acute, subacute, chronic, acute on chronic PE when these words or similar words appeared in the official report. When no word describing chronicity was available, the study was classified into "unclassified". All radiology reports were manually reviewed and classified by two radiologists independently. The Cohen's kappa coefficients for presence, chronicity, and proximal extension were 0.974, 0.868, and 0.912, respectively. Consensus was made where initial classifications disagreed.

Of the total sample of 10,330 CTPA reports, 2054 were positive for PE, and 8276 were negative. Among the positive samples, the fraction of manually validated acute, acute on chronic, subacute, chronic, and unclassified PE were 83.4%, 3.0%, 2.8%, 7.8%, and 2.9%, respectively. Proximal extension of the embolus was classified into central for 23.7%, lobar for 23.4%, segmental for 40.0%, and subsegmental for 12.9%, respectively.

2.2. NLP analysis and classification of CTPA reports

2.2.1. NLP library

Our report analysis program uses NILE for natural language processing. NILE is an NLP library developed for information extraction from clinical narratives. In named entity recognition (NER), NILE uses a prefix-tree matching algorithm that extracts the longest recognizable term from the left. For example, it would identify "heart failure" rather than "heart", because the former term is longer. It would then proceed from the word after the identified term, so "failure" would not be extracted, either. The data structure of NILE's dictionary is a prefix-tree with words as nodes. The processing time of NER is proportional to the document length, and is hardly affected by the dictionary size with the use of hash maps at each node. Compared to popular NER approaches that first identify phrases with shallow parsing then match them against a table-like dictionary, NILE's matching algorithm can be faster by orders of magnitude.

An entry of NILE's dictionary comprises a term, concept, and semantic role. The concept is treated as an identifier and is shared by synonyms. The semantic role indicates the function of the term in a sentence, and is used by the semantic analyzers later on. The roles are predefined by the program. There are three categories of semantic roles. Grammatical words are words that help structure the sentence, but with little meaning by themselves, such as conjunctions and prepositions. Meaning cues are words and phrases that express predefined meanings, e.g., "no" indicates negation and "maybe" indicates speculation. Finally, medical terms are concepts related to diagnosis or treatment, such as facts, modifiers, and anatomical locations, where facts can be disorders, findings, procedures, tests, substances (like drugs), etc.

After terms are identified, they are sent through a pipeline of semantic analyzers to determine the meanings associated with the mentions. The semantic analyzers are finite-state machines that analyze the sequence of observed terms, and apply rules that are based on grammatical and clinical knowledge. The essential semantic analyses that we needed from NILE were presence, general modification, and location modification. Table 1 illustrates NILE's semantic analysis capabilities with sentences from CTPA reports. Sentence 1 is an example of presence analysis. Presence analysis in NILE combines analyses of negation and speculation, and its value can be YES, NO, or MAYBE. Sentence 2 demonstrates basic location and modification analyses. The entities inside the parentheses modify the entity before them. Here, "bilateral" is a

general modifier, and the others are all location modifiers. It also demonstrates the handling of prefix- and suffix-sharing in identification of the lobe names. Sentence 3 shows advanced location analysis, where location modifications are nested. The nested parentheses read the same way as the simpler ones – ''right upper lobe'' modifies ''pulmonary artery'', and ''pulmonary artery'' and ''multiple'' modify ''filling defects''. This nested modification structure is important to the classification of proximal extension of the embolus, because if we did not know that the ''segmental branches'' was a direct modifier to the filling defects and the ''right upper lobe pulmonary artery'' was not, we would not be able to distinguish whether the location of the filling defects was in segmental artery or lobar artery. Sentence 4 also illustrates advanced location analysis where conjunction and nesting coexist, in the presence of imperfect grammar, typo, and term not in the dictionary. The presence of the second mention of filling defect is MAYBE, because the mention is modified by ''possible''. Sentences 5 and 6 show that NILE can distinguish whether a mention of PE is about its presence. ''Study'' is a cue word that suggest the mentions in the sentence are not about presence, but in Sentence 6, ''demonstrates'' terminates the scope of ''study''.

2.2.2. Dictionary

NILE comes with only a basic dictionary of grammatical words and meaning cues that it uses for semantic analysis. For our application, we populated NILE's dictionary with concepts and terms that may appear in CTPA reports. An entry of NILE's dictionary is a triple of term, concept code, and semantic role. Term is the natural language expression of the concept. A concept typically has multiple terms that are synonyms of each other or inflections of the base form. Terms of the same concepts share the same concept code, and are treated uniformly by the application. The semantic role of a term tells NILE's semantic analyzers how to understand the term in a sentence. The most basic semantic role in NILE is fact. A fact can be a disease, a symptom, a finding, a medicine, etc., and the terms are typically noun phrases. Other semantic roles that we used for the application included (regular) modifiers and anatomical locations. Anatomical locations are also modifiers, but in NILE they are handled by a dedicated location analyzer for nested modification analysis. After semantic analysis, modifiers and locations will be attached to the facts that they modify.

For the fact concepts, we manually located the Unified Medical Language System (UMLS) concepts that were important for detecting PE, including C0034065 Pulmonary Embolism, C1704212 Embolism, C0332555 Filling Defect, C0302148 Clot, and C0040053 Thrombosis, pulled their terms from the UMLS as entries of the dictionary, manually complemented the plural forms and removed the entries that were not natural language terms. Embolism and Pulmonary Embolism also included terms ''embolus'' and ''pulmonary embolus'', respectively, and Clot also included the term ''thrombus''. Physicians provided the terms of anatomical locations and other modifiers of PE. In our models, we used the following concepts:

- Facts: Pulmonary Embolism, Filling Defect, Clot, and Thrombosis. We merged the above UMLS concepts Pulmonary Embolism and Embolism. In general Embolism is a broader concept of Pulmonary Embolism; however, in the context of the selected CTPA reports, physicians usually write ''pulmonary embolism'' as ''embolism'' for short, and there is no ambiguity.
- Anatomical locations
  - Named arteries: Main Pulmonary Arteries, Lobar Pulmonary Arteries, Segmental Pulmonary Arteries, and Subsegmental Pulmonary Arteries. Arteries of the same level shared the same code. For instance, arteries of various segments (apical, anterior, posterior, etc.) were all coded as Segmental Pulmonary Artery.
  - Named sites: Segments and Lobes. Same as named arteries, sites of the same level shared the same code.
  - Other: Artery and Branch. These two concepts played special roles, which will be described in feature generation.
- General modifiers
  - Location related: Main, Segmental, and Subsegmental. We treated the word ''lobar'' as anatomical location. The word ''central'' is not a term of Main, because the meaning of ''central'' is ambiguous, and usually it means the location of PE within the lumen of the artery, as in ''central vs. eccentric'', rather than PE being in main pulmonary arteries.
  - Chronicity related: Acute, Non-acute, Acute on Chronic, Subacute, Chronic, and Previous.
  - Other words that may modify PE and filling defects, such as ''bilateral'', ''multiple'', and ''other''. We want to recognize these words to distinguish cases like ''no PE'' and ''no other PE''.

The above concepts contained 383 terms. In addition, we also loaded terms of other fact concepts that may appear in CTPA reports to the dictionary, but they were not examined by the program nor used in the classification models. The total number of custom terms in the end was 1123.

2.2.3. Report cleaning and sectioning

CTPA reports at our institution are stored in plain text, without tags to indicate their structure. For each report, we removed the artificial line breaks and restored the paragraphs with an ad hoc heuristic method. This step ensured that NILE processed on whole sentences to produce the correct semantic output. It was also a prerequisite to find the sections correctly.

The sections were identified by their headings, and headings again were identified by an ad hoc method (uppercase letters followed by a colon). The sections Technique, Exam, Procedure, Comparison, and those with similar names were ignored, as they do not contain findings. Sections regarding lower extremities, pelvis, thigh, and abdomen were also ignored, because imaging findings, e.g., thrombus, in these sections could be confounding information to the classification of PE. For the same reason, sentences regarding these locations in the Impressions section were ignored as well. Sections of History and Indication were treated separately from the other sections, as entity mentions in these two sections may have different meanings. For instance, ''concern for PE'' in these two sections does not imply anything about the presence of PE, but in other sections it confirms its presence. Also, all fact mentions in these two sections were treated as history, and we artificially added a modifier Previous to each fact mention.

2.2.4. Feature generation

After sectioning, the appropriate portions of the report were sent to NILE for semantic analysis. Here we describe how we converted the NLP output of a report to numeric features.

Fact features

The program counted mentions of the concepts Pulmonary Embolism, Filling defect, Clot, and Thrombosis. Each mention had a semantic attribute for presence, which could be YES, NO, or MAYBE. The program counted them respectively, thus each concept gave 3 variables. When counting the NO's, we only counted those with no modifiers other than ''acute''. For example, ''no other filling defects'' and ''no filling defects in the lower lobes'' did not count as NO because they do not imply ''no filling defects'' overall.

Chronicity features

For any occurrence of the above four fact concepts, if the presence attribute was not NO, the program also counted the following

modifiers: Acute, Non-acute, Acute on Chronic, Subacute, Chronic, and Previous. Recall that facts in the History and the Indication sections were treated as history and all bore a modifier of Previous. In addition, the program also counted mentions of "interval" and "previous/prior study", which are informative about chronicity. In total, the program generated 8 chronicity related variables.

Location features

Four location related features were generated: Central, Lobar, Segmental, and Subsegmental, i.e., the levels of proximal extension. For each mention of the above four fact concepts whose presence attribute was not NO, if locations or location modifiers were mentioned, then the program would use the location modification structures (examples see Table 1, Sentences 3 and 4) from NILE to analyze the most proximal location for each mention, and only the most proximal location of each mention would be counted. The program considers a location related regular modifier as a stronger description of the level than a location. For example, in "segmental emboli in lobar arteries", "segmental" binds more strongly with "emboli" than "lobar arteries" does. To find the most proximal location of a mention, the program checks and scores its direct modifiers, as described in Algorithm 1. In addition, it was common that arteries were expressed in ways that were not recorded in the dictionary. Algorithm 2 captures these arteries by using the nested location modification structure. For example, "filling defects in pulmonary arteries of the lateral basal segment" is interpreted as `filling defects:YES (pulmonary arteries (lateral basal segment))`, and the artery level can be recognized because it is modified by a segment. We applied this kind of composite concept recognition on Arteries and Branches. The level of an artery is determined by the site level. For a branch, if it is modified by a site, its level equals to the site level; if it is modified by an artery, its level is one level lower than the artery level (because it is a branch of that artery).

Algorithm 1. Finding the most proximal location

```
locList = new list
locList.add(0)
For modObj in modifiers {
    Switch(modObj) {
        case Subsegmental Pulmonary Artery:
  locList.add(1)
        case Segmental Pulmonary Artery:
  locList.add(2)
        case Lobar Pulmonary Artery: locList.add(3)
        case Subsegmental (regular modifier):
  locList.add(4)
        case Segmental (regular modifier):
  locList.add(5)
        case Main Pulmonary Artery: locList.add(6)
        case Main (regular modifier):
  locList.add(6)
        default: do nothing
        }
}
    mostProximalLoc = max(locList)
    Switch(mostProximalLoc){
    case 1,4: count as Subsegmental
    case 2,5: count as Segmental
    case 5: count as Lobar
    case 6: count as Central
    default: do nothing
}
```

Algorithm 2. Artery level recognition

```
function analyzeArteryLevel(artery) {
    modObj = artery.first_modifier
    Switch(modObj){
        case Segment: return Segmental Pulmonary
  Artery
        case Lobe: return Lobar Pulmonary Artery
        default: return null
    }
}
function analyzeBranchLevel(branch) {
    modObj = branch.first_modifier
    Switch(modObj){
        case Segment: return Segmental Pulmonary
  Artery
        case Lobe: return Lobar Pulmonary Artery
        case Segmental Pulmonary Artery: return
  Subsegmental Pulmonary Artery
        case Lobar Pulmonary Artery: return
  Segmental Pulmonary Artery
        case Main Pulmonary Artery: return Lobar
  Pulmonary Artery
        case Artery: return
  analyzeArteryLevel(modObj)-1
        default: return null
    }
}
```

The above 24 features are an numeric summary of a report. Each report was summarized as a vector in the above way and sent to statistical models for classification.

2.2.5. Classification

We trained a classifier for presence vs. absence of PE with all the 24 variables using all the 10330 samples. Using the 2054 positive samples, we trained a chronicity classifier (acute vs. others) with the 8 chronicity variables, and two location classifiers (central vs. others and subsegmental vs. others) with the 4 location variables. All classifiers were trained by fitting penalized logistic regression models with adaptive LASSO penalty [39] to alleviate over fitting. The tuning parameter for the penalized regression was selected based on the Bayesian Information Criterion [40].

2.3. Evaluation

To evaluate the performance of the prediction, true positive rate (TPR), false positive rate (FPR), positive predictive value (PPV), negative predictive value (NPV), as well as the receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC) were estimated. The PPV for PE present vs. absent was adjusted by accounting for a total of 4164 negative CTPA examinations performed from 8/1/2003 to 10/30/2006 that were not in the study population. The 0.632 bootstrap cross-validation [41,42] was used to correct for overfitting bias and the bootstrap [43] was used to estimate the standard errors for all statistics. For each algorithm, 1000 bootstrap replicates were used to obtain the standard errors and the confidence interval estimates.

We compare the performance of our NLP-based classifiers to that of bag-of-word multinomial Naive Bayes classifiers [44]. Bag-of-word classifiers are commonly used for text classification, and in many scenarios, despite their simplicity, they are very competent. The features were unigrams that were obtained by splitting

Table 1
NILE semantic analysis.

Sentence	Output
1. "No filling defects are seen to suggest pulmonary embolism"	Filling defects: NO Pulmonary embolism: NO
2. "Bilateral filling defects are seen within the distal main pulmonary arteries bilaterally, extending into the right upper, middle, lower, left upper, lingula, and left lower lobes"	Filling defects: YES (distal main pulmonary arteries; right upper lobe; right middle lobe; right lower lobe; left upper lobe; lingula; left lower lobe; bilateral)
3. "Multiple filling defects are seen in the segmental branches of the right upper lobe pulmonary artery"	Filling defects:YES (segmental branches (pulmonary artery (right upper lobe)); multiple)
4. "There are segmental and subsegmental filling defects in the right upper lobe, superior segment of the right lower lobe, and possible subsegmental filling defect in the anterolateral segment of the left lower lobe pulmonary arteries"	Filling defects: YES (right upper lobe; superior segment (right lower lobe); segmental; subsegmental) Filling defect: MAYBE (segment (pulmonary arteries (left lower lobe)); subsegmental)
5. "The study is of adequate technical quality for diagnosis of pulmonary embolism"	None
6. "A CT pulmonary angiogram is excellent quality study and demonstrates numerous pulmonary emboli seen centrally at the divisions of bilateral lobar pulmonary arteries and dissemination into multiple subsegmental branches in all lobes"	Pulmonary emboli: YES (lobar pulmonary arteries; subsegmental branches (all lobes); numerous; bilateral; multiple)

lines at white spaces, digits, and punctuations, and were counted through the whole reports, i.e., no report sectioning. Except lowercasing, we did not do any further refinement to the unigrams, such as stemming, lemmatization, or synonym grouping. Half of the reports were used for estimating the model parameters, and the other half were used for validation.

3. Results

The classifiers achieved high accuracy for all four tasks, with AUC being 0.998 ± 0.005, 0.945 ± 0.015, 0.987 ± 0.005, 0.986 ± 0.004, for PE present, acute PE, central PE, and subsegmental PE, respectively. Fig. 1 shows the ROC curves of the four classifiers (blue solid lines) with lower and upper 5% quantiles (blue dotted lines) estimated from bootstrapping. In comparison, the baseline bag-of-words models achieved AUC 0.942, 0.765, 0.766, and 0.712, respectively, and their ROC curves were plotted in red.

The logistic model assigns each report a score between 0 and 1 as predicted probability of being in a class, and the report is classified by whether the score is above a threshold. A higher threshold gives higher PPV but lower NPV. Our threshold values were chosen to balance the positive and negative predictive values according to radiologists' need. For PE present, we classified the report as PE positive if the predicted probability exceeded the threshold value of 0.5, yielding PPV of 0.95 and NPV of 0.99. Table 2 shows TPR, FPR, PPV, NPV, and F1 scores of all classifiers. Comparisons with baseline classifiers are also provided, where the threshold was chosen to be the one that best matches the FPR to the FPR of the corresponding NLP-based classifiers. In addition, the average F1 scores of the two radiologists are listed in column 'R'. Table 3 lists the beta coefficients of each model.

4. Discussion

This paper introduces and tests a new NLP application with excellent prediction capability to identify the presence, chronicity, and most proximal location of pulmonary embolus from CTPA reports. The application utilized the NILE NLP library, and achieved success with careful selection of dictionary entries, treatment of the reports, and interpretation of the NLP output. The performance of the classifiers had clear margins over the baseline bag-of-words classifiers, especially in the classification of subtypes, as shown by the ROC curves in Fig. 1. In the perspective of NLP methodology, the accuracy demonstrated in this application supports the linear semantic analysis approach advocated by NILE. It is worth pointing out that the semantic analyzers used in this project were written for general purposes, and we used them without customization (the only thing customized was the dictionary content). The fact that they worked well for radiology reports is good news to NILE and the approaches that it takes.

The best performance was from the presence vs. absence classification. The task was the easiest of the four, because even the bag-of-words classifier without NLP achieved AUC = 0.942. With NLP treatment, we were able to improve it to almost perfect, with AUC = 0.998. The performance of the chronicity classification was the lowest among the four, and got an AUC of 0.945, which is still good overall. The task was a hard one even for humans, as shown by the inter-rater agreement coefficients. A previous study [31] also confirmed its difficulty for NLP. We suspect that a cause for the low performance compared to the other three tasks was that chronicity information come from various sources that are spread (or hidden) across the documents, but our program only looked at chronicity modifiers of the four target concepts, with "interval" and "previous/prior study" as the only exceptions. In addition, the chronicity information are sometimes subtle hints and require reasoning to interpret. One example is that if the report says "as compared to [findings] on MM/DD", then a human reviewer could calculate the interval from the mentioned visit to the current one, and may conclude that the current visit is a follow-up. Other hints could be "partial resolution of [problem]" or "improved state of [problem]", which both contain temporal information. Unfortunately, NILE does not have a temporal analyzer at the moment, neither can it do reasoning like human.

The most innovative part of our program is the identification of the most proximal PE location, which is not seen in previous NLP applications. Our classifiers for central and subsegmental PE achieved AUC of 0.987 and 0.986, respectively, as compared to 0.766 and 0.712 by the bag-of-words classifiers. The success was largely due to the nested location modification analysis, a unique feature of NILE. The multilayer modification structure allowed us to recognize levels of arteries that were not recorded in the dictionary, and to distinguish direct and indirect location modifiers. Both of which were key to boost the recall and the precision. On the other hand, the location analyzer was not perfect. When we reviewed the reports whose proximal locations were incorrectly classified by the program, we found many of them were due to ambiguity, i.e., the same sentence structure could be interpreted in multiple ways. The most common ambiguity in nested location

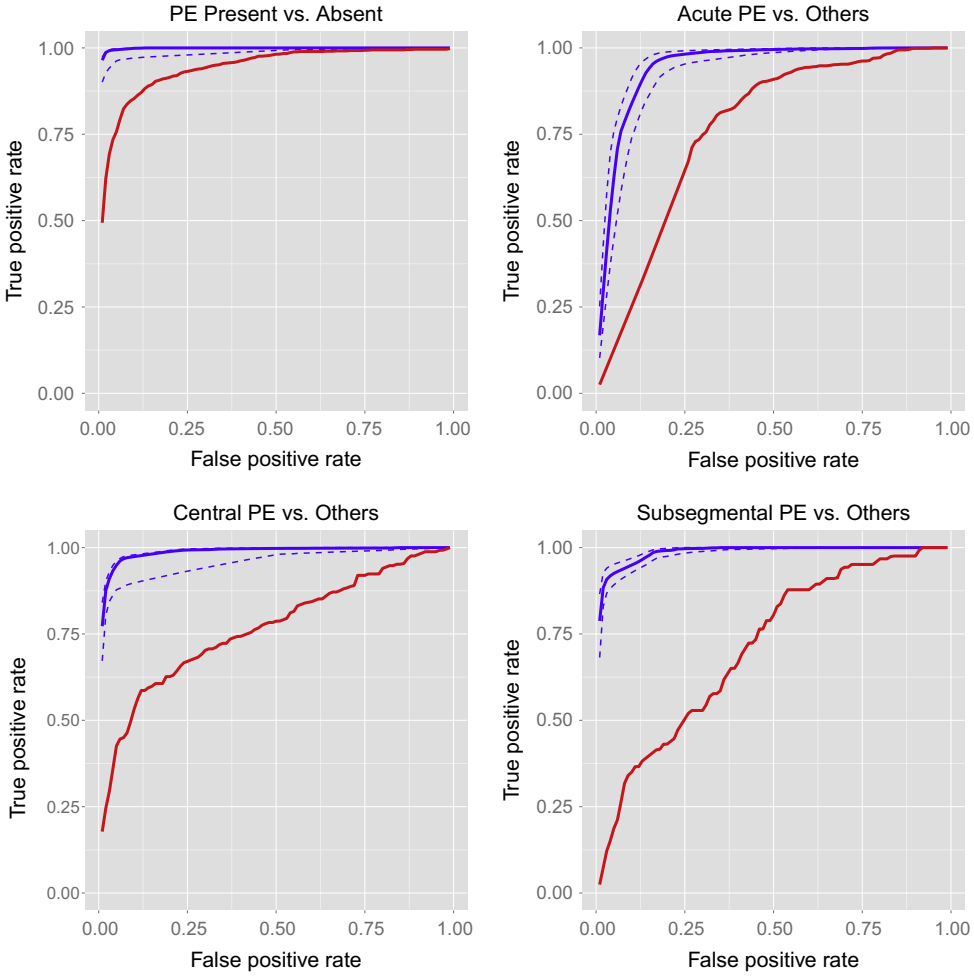

Fig. 1. ROC curves of PE classifications.

Table 2
Classifier performance (N – NLP, B – Bag-of-words).

Classification	TPR		FPR		PPV		NPV		F1 score		
	N	B	**N**	B	**N**	B	**N**	B	**N**	B	R
Present vs. absent	**.96**	.49	**.01**	.01	**.95**	.93	**.99**	.89	**.96**	.64	.98
Acute vs. others	**.98**	.67	**.25**	.27	**.96**	.93	**.90**	.29	**.97**	.78	.98
Central vs. others	**.90**	.29	**.03**	.03	**.91**	.76	**.97**	.81	**.91**	.42	.92
Subseg vs. others	**.79**	.02	**.01**	.01	**.91**	.25	**.97**	.88	**.84**	.04	.92

The bold font was used to highlight the proposed method of the paper, and the regular font was for the reference values.

modification is in conjunctions. For example, the Impression section of one CTPA report had the description "pulmonary embolus in the anterior segment of the left upper lobe pulmonary artery and in the right upper lobe pulmonary artery", which can be interpreted either as

```
pulmonary embolus:YES (anterior segment (pulmonary artery (left upper lobe)); pulmonary artery (right upper lobe)),
```

or as

```
pulmonary embolus:YES (anterior segment (pulmonary artery (left upper lobe); pulmonary artery (right upper lobe))).
```

The first interpretation considers "right upper lobe pulmonary artery" as a modifier to "pulmonary embolus", while the second considers it as a modifier to "anterior segment". The current version of NILE interprets the sentence as the first one, but based on descriptions in the Findings section of the report, the second interpretation was the intended one. This shows that linguistic knowledge alone is not enough for semantic analysis. A dedicated analyzer incorporated with clinical domain knowledge is hopeful to be more accurate.

The nested modification structure also reminds one of the parse tree from deep parsing. Our experience is that NILE is much more friendly and easier to use than deep parsers in medical NLP applications. As an illustration, we used the online version of the Stanford Parser [45] (http://nlp.stanford.edu:8080/parser/index.jsp) to parse the above sentence "Pulmonary embolus in the anterior segment of the left upper lobe pulmonary artery and in the right upper lobe pulmonary artery". The output is shown in Table 4. First of all, we observed that the output depended on whether there was a

Table 3
Beta coefficients of each model.

	Presence	Acute	Central	Subsegmental
Intercept	−3.053	3.627	−3.541	−2.098
Clot.YES	0.660	–	–	–
Clot.NO	0.011	–	–	–
Clot.MAYBE	−0.035	–	–	–
Thrombosis.YES	0.007	–	–	–
Thrombosis.NO	0.000	–	–	–
Thrombosis.MAYBE	−0.018	–	–	–
PE.YES	2.888	–	–	–
PE.NO	−2.421	–	–	–
PE.MAYBE	0.822	–	–	–
FillingDefect.YES	1.433	–	–	–
FillingDefect.NO	−0.929	–	–	–
FillingDefect.MAYBE	0.006	–	–	–
PreviousStudy	0.026	−0.104	–	–
Interval	−0.085	−0.318	–	–
Previous	−1.645	−1.410	–	–
Acute	0.854	0.017	–	–
Non-acute	0.646	−4.407	–	–
AcuteOnChronic	0.000	−0.850	–	–
Subacute	0.143	−4.019	–	–
Chronic	0.734	−4.175	–	–
Main	0.924	–	4.397	−3.812
Lobar	1.114	–	0.000	−2.161
Segmental	1.259	–	0.000	−3.018
Subsegmental	1.680	–	0.000	2.412

period at the end of the test sentence. Neither of the outputs was correct, and the output from NILE, although still wrong, was clearly superior to both. In addition to the tree structure, we also noticed that a good number of part-of-speech tags from both outputs were incorrect. Therefore it is very hard to do named entity recognition correctly using the result from the Stanford Parser. Finally, deep parsing is time consuming. The parsing time shown by the online version of the Stanford Parser was 798 ms for the sentence with the period and 785 ms without the period. In comparison, the total time inside NILE for processing the 10,330 reports was 2361 ms, i.e., 0.23 ms per report.

Since the entity level interpretation can get wrong, as demonstrated earlier, hard rules, such as `if PE.YES > 0 then classify as PE present`, are not the best classification criterion. Instead, statistical classification models can integrate information from all the features to find a more reliable decision rule. The logistic classifier that we chose is one of many models that can do this job adequately. It is a linear model whose result is easy to interpret compared to models like support vector machines and decision trees. The adaptive LASSO penalty can shrink beta coefficients of uninformative features to zero, automatically doing feature selection. From Table 3, we see that the signs and magnitudes of the coefficients are meaningful. In Present vs. Absent, the model used information not only from the presence information of the fact concepts, but also from their modifiers. This makes good sense, because if PE is not found, then the report is not likely to talk about its location or temporal attributes, or reversely, talking about attributes of PE is an evidence of its presence — except for Previous, which indicates that the mention is about history and got a large negative coefficient. A limitation of the logistic model (as with many other models) is that it depends on the frequencies of the mentions: if a report mentions subsegmental emboli in three sentences, then even if it also mentions a segmental embolus in another sentence, the model will still incorrectly classify the most proximal PE as subsegmental.

One limitation of the current study is that the database is from a single academic institution in the United States, and thus the accuracy of our NLP algorithm when applied to CTPA reports from other centers is uncertain. Although content in CTPA reports is relatively limited and in general has a constrained vocabulary, the degree of reporting variation is not clear and it is necessary to validate our NLP tool using data from other institutions. In Table 2, we only selected one specific threshold value to best balance the trade-off between PPV and NPV, but this should be changed according to the purpose of the NLP for each project. The discrepancy in the time periods for positive and negative CTPA data collection can be another limitation, while this does not significantly affect the

Table 4
Parse trees from the Stanford Parser.

with period

```
(ROOT
  (S
    (NP (NNP Pulmonary))
    (VP (VBZ embolus)
      (PP
        (PP (IN in)
          (NP
            (NP (DT the) (NN anterior) (NN segment))
            (PP (IN of)
            (NP (DT the) (JJ left) (JJ upper) (NN lobe) (JJ pulmonary) (NN artery)))))
        (CC and)
        (PP (IN in)
          (NP (DT the) (JJ right) (JJ upper) (NN lobe)))
        (NP (JJ pulmonary) (NN artery))))
  (..)))
```

without period

```
(ROOT
  (UCP
    (ADJP (RB Pulmonary) (JJ embolus)
      (PP (IN in)
        (NP
          (NP (DT the) (NN anterior) (NN segment))
          (PP (IN of)
          (NP (DT the) (JJ left) (JJ upper) (NN lobe) (JJ pulmonary) (NN artery))))))
    (CC and)
    (PP (IN in)
    (NP (DT the) (JJ right) (JJ upper) (JJ lobe) (JJ pulmonary) (NN artery)))))
```

NLP algorithm development and the value of PPV was adjusted accordingly.

Our approach can also be potentially ported to other applications. For general presence analysis, the key features for classification are the presence attribute of the target condition and critical evidences. For many subtype classifications, the modifiers are usually the key, but one may want to count them only when the presence attribute of the entity is not NO, as we did with PE. The nested location modification is useful in many cases, e.g., in determining symmetry of rheumatoid arthritis symptoms in physician notes (e.g., "swelling in both wrists"), and determining the anatomical location of brain aneurysm in radiology reports, which have similar patterns as PE CTPA reports when describing arteries. In order to use NILE for these tasks, one needs to prepare the concepts and terms for the dictionary, and write custom program to interpret NILE's semantic output for the target application.

In conclusion, we introduced an internally validated NLP application with excellent prediction capability to identify the presence, chronicity and most proximal location of PE from CTPA reports. This algorithm may potentially be used to create large multicenter databases for patients with PE.

5. Funding/support

The project was supported by R01-GM079330 from the National Institutes of Health (NIH), as well as by the Award No. U54-LM008748 from the National Library of Medicine (NLM). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NLM or NIH.

References

[1] Carson JL, Kelley MA, Duff A, Weg JG, Fulkerson WJ, Palevsky HI, et al. The clinical course of pulmonary embolism. New England J Med 1992;326(19):1240–5.
[2] Tapson VF. Acute pulmonary embolism. New England J Med 2008;358(10):1037–52.
[3] Goldhaber SZ, Visani L, De Rosa M. Acute pulmonary embolism: clinical outcomes in the international cooperative pulmonary embolism registry (ICOPER). Lancet 1999;353(9162):1386–9.
[4] Hunsaker AR, Lu MT, Goldhaber SZ, Rybicki FJ. Imaging in acute pulmonary embolism with special clinical scenarios. Circ: Cardiovasc Imag 2010;3(4):491–500.
[5] Cai B, Bedayat A, George E, Hunsaker AR, Dill KE, Rybicki FJ, et al. Malignancy and acute pulmonary embolism: risk stratification including the right to left ventricle diameter ratio in 1596 subjects. J Thorac Imag 2013;28(3):196–201.
[6] Kumamaru KK, Hunsaker AR, Bedayat A, Soga S, Signorelli J, Adams K, et al. Subjective assessment of right ventricle enlargement from computed tomography pulmonary angiography images. Int J Cardiovasc Imag 2012;28(4):965–73.
[7] Kumamaru KK, Hunsaker AR, Wake N, Lu MT, Signorelli J, Bedayat A, et al. The variability in prognostic values of right ventricular-to-left ventricular diameter ratios derived from different measurement methods on computed tomography pulmonary angiography: a patient outcome study. J Thorac Imag 2012;27(5):331–6.
[8] Kumamaru KK, Lu MT, Ghaderi Niri S, Hunsaker AR. Right ventricular enlargement in acute pulmonary embolism derived from ct pulmonary angiography. Int J Cardiovasc Imag (formerly Cardiac Imaging) 2013:1–4.
[9] Lu MT, Cai T, Ersoy H, Whitmore AG, Levit NA, Goldhaber SZ, et al. Comparison of ecg-gated versus non-gated ct ventricular measurements in thirty patients with acute pulmonary embolism. Int J Cardiovasc Imag 2009;25(1):101–7.
[10] Lu MT, Cai T, Ersoy H, Whitmore AG, Quiroz R, Goldhaber SZ, et al. Interval increase in right–left ventricular diameter ratios at CT as a predictor of 30-day mortality after acute pulmonary embolism: initial experience 1. Radiology 2008;246(1):281–7.
[11] Lu MT, Demehri S, Cai T, Parast L, Hunsaker AR, Goldhaber SZ, et al. Axial and reformatted four-chamber right ventricle–to–left ventricle diameter ratios on pulmonary ct angiography as predictors of death after acute pulmonary embolism. Am J Roentgenol 2012;198(6):1353–60.
[12] Parast L, Cai B, Bedayat A, Kumamaru KK, George E, Dill KE, et al. Statistical methods for predicting mortality in patients diagnosed with acute pulmonary embolism. Acad Radiol 2012;19(12):1465–73.
[13] Tamariz L, Harkins T, Nair V. A systematic review of validated methods for identifying venous thromboembolism using administrative and claims data. Pharmacoepidem Drug Safe 2012;21(S1):154–62.
[14] White RH, Garcia M, Sadeghi B, Tancredi DJ, Zrelak P, Cuny J, et al. Evaluation of the predictive value of ICD-9-CM coded administrative data for venous thromboembolism in the United States. Thromb Res 2010;126(1):61–7.
[15] White RH, Sadeghi B, Tancredi DJ, Zrelak P, Cuny J, Sama P, et al. How valid is the ICD-9-CM based AHRQ patient safety indicator for postoperative venous thromboembolism? Med Care 2009;47(12):1237–43.
[16] Zhan C, Battles J, Chiang Y-P, Hunt D. The validity of ICD-9-CM codes in identifying postoperative deep vein thrombosis and pulmonary embolism. Joint Comm J Qual Patient Safe 2007;33(6):326–31.
[17] Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF, et al. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform 2008;35:128–44.
[18] Xu H, Stetson PD, Friedman C. A study of abbreviations in clinical notes. AMIA annual symposium proceedings, vol. 2007. American Medical Informatics Association; 2007. p. 821.
[19] Wu Y, Denny JC, Rosenbloom ST, Miller RA, Giuse DA, Xu H. A comparative study of current clinical natural language processing systems on handling abbreviations in discharge summaries. AMIA annual symposium proceedings, vol. 2012. American Medical Informatics Association; 2012. p. 997.
[20] Kuhn IF. Abbreviations and acronyms in healthcare: when shorter is not sweeter. Pediatr Nurs 2007;33(5).
[21] Sheppard JE, Weidner LC, Zakai S, Fountain-Polley S, Williams J. Ambiguous abbreviations: an audit of abbreviations in paediatric note keeping. Arch Dis Child 2008;93(3):204–6.
[22] Pakhomov S, Pedersen T, Chute CG. Abbreviation and acronym disambiguation in clinical discourse. AMIA annual symposium proceedings, vol. American Medical Informatics Association; 2005. p. 589.
[23] Berman JJ. Pathology abbreviated: a long review of short terms. Arch Pathol Lab Med 2004;128(3):347–52.
[24] Liu H, Lussier YA, Friedman C. A study of abbreviations in the UMLS. In: Proceedings of the AMIA symposium, american medical informatics association; 2001. p. 393.
[25] Haug PJ, Ranum DL, Frederick PR. Computerized extraction of coded findings from free-text radiologic reports. work in progress. Radiology 1990;174(2):543–8.
[26] Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. J Am Med Inf Assoc 1994;1(2):161–74.
[27] Hripcsak G, Austin JH, Alderson PO, Friedman C. Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports 1. Radiology 2002;224(1):157–63.
[28] Dreyer KJ, Kalra MK, Maher MM, Hurier AM, Asfaw BA, Schultz T, et al. Application of recently developed computer algorithm for automatic classification of unstructured radiology reports: validation study 1. Radiology 2005;234(2):323–9.
[29] Uzuner Ö, Zhang X, Sibanda T. Machine learning and rule-based approaches to assertion classification. J Am Med Inf Assoc 2009;16(1):109–15.
[30] Raja AS, Ip IK, Prevedello LM, Sodickson AD, Farkas C, Zane RD, et al. Effect of computerized clinical decision support on the use and yield of ct pulmonary angiography in the emergency department. Radiology 2012;262(2):468.
[31] Chapman BE, Lee S, Kang HP, Chapman WW. Document-level classification of ct pulmonary angiography reports based on an extension of the context algorithm. J Biomed Inf 2011;44(5):728–37.
[32] Ribeiro A, Juhlin-Dannfelt A, Brodin L-Å, Holmgren A, Jorfeldt L. Pulmonary embolism: relation between the degree of right ventricle overload and the extent of perfusion defects. Am Heart J 1998;135(5):868–74.
[33] Wolfe MW, Lee RT, Feldstein ML, Parker JA, Come PC, Goldhaber SZ. Prognostic significance of right ventricular hypokinesis and perfusion lung scan defects in pulmonary embolism. Am Heart J 1994;127(5):1371–5.
[34] Goodman L. Small pulmonary emboli: what do we know? Radiology 2005;234(3):654.
[35] Le Gal G, Righini M, Parent F, Van Strijen M, Couturaud F. Diagnosis and management of subsegmental pulmonary embolism. J Thromb Haemostasis 2006;4(4):724–31.
[36] Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. component evaluation and applications. J Am Med Inf Assoc 2010;17(5):507–13.
[37] Dligach D, Bethard S, Becker L, Miller T, Savova GK. Discovering body site and severity modifiers in clinical texts. J Am Med Inf Assoc 2013 [amiajnl–2013].
[38] Yu S, Cai T. A short introduction to NILE, arXiv:1311.6063.
[39] Zou H. The adaptive lasso and its oracle properties. J Am Stat Assoc 2006;101(476):1418–29.
[40] Hastie T, Tibshirani R, Friedman J, Hastie T, Friedman J, Tibshirani R. The elements of statistical learning: data mining, inference and prediction. Springer; 2009.
[41] Efron B, Tibshirani R. Improvements on cross-validation: the 632+ bootstrap method. J Am Stat Assoc 1997;92(438):548–60.
[42] Jiang W, Simon R. A comparison of bootstrap methods and an adjusted bootstrap approach for estimating the prediction error in microarray classification. Stat Med 2007;26(29):5320–34.
[43] Efron B, Tibshirani RJ. An introduction to the bootstrap. CRC press; 1994.
[44] Kibriya AM, Frank E, Pfahringer B, Holmes G. Multinomial naive bayes for text categorization revisited. In: AI 2004: advances in artificial intelligence. Springer; 2005. p. 488–99.
[45] Klein D, Manning CD. Accurate unlexicalized parsing. Proceedings of the 41st annual meeting on association for computational linguistics, vol. 1. Association for Computational Linguistics; 2003. p. 423–30.