Natural Language Processing Technologies in Radiology Research and Clinical Applications[1]

Tianrun Cai, MD
Andreas A. Giannopoulos, MD
Sheng Yu, PhD[2]
Tatiana Kelil, MD
Beth Ripley, MD, PhD
Kanako K. Kumamaru, MD, PhD
Frank J. Rybicki, MD, PhD[2]
Dimitrios Mitsouras, PhD

Abbreviations: CDS = clinical decision support, cTAKES = Text Analysis and Knowledge Extraction, CUI = concept unique identifier, EMR = electronic medical record, iSCOUT = Searching Content with an Ontology-Utilizing Toolkit, LEXIMER = Lexicon Mediated Entropy Reduction, MedLEE = Medical Extraction and Encoding, NILE = Narrative Information Linear Extraction, NLP = natural language processing, SNOMED-CT = Systematized Nomenclature of Medicine–Clinical Terms, UMLS = Unified Medical Language System

RadioGraphics 2016; 36:176–191

Published online 10.1148/rg.2016150080

Content Codes: IN RS

[1]From the Applied Imaging Science Laboratory, Department of Radiology, Brigham and Women's Hospital, 75 Francis St, Boston, MA 02115 (T.C., A.A.G., K.K.K., F.J.R., D.M.); Harvard T.H. Chan School of Public Health, Boston, Mass (S.Y.); and Department of Radiology, Brigham and Women's Hospital, Boston, Mass (T.K., B.R.). Presented as an education exhibit at the 2014 RSNA Annual Meeting. Received March 24, 2015; revision requested August 20 and received September 18; accepted October 12. For this journal-based SA-CME activity, the authors, editor, and reviewers have disclosed no relevant relationships. **Address correspondence to** D.M. (e-mail: *dmitsouras@alum.mit.edu*).

[2]**Current address:** S.Y.: Tsinghua University Centre for Statistical Science, Beijing, China. F.J.R.: Department of Radiology, University of Ottawa, Faculty of Medicine, The Ottawa Hospital, Ottawa, Ontario, Canada.

Funding: The work was supported by the National Institutes of Health [grant number EB015868].

SA-CME LEARNING OBJECTIVES

After completing this journal-based SA-CME activity, participants will be able to:

■ Describe the set of technologies that compose present-day natural language processing in radiology.

■ List examples of how these technologies have been combined to achieve specific objectives in radiology research and, potentially, clinical practice.

■ Discuss current capabilities and possible future applications of use of natural language processing in radiology.

See www.rsna.org/education/search/RG.

The migration of imaging reports to electronic medical record systems holds great potential in terms of advancing radiology research and practice by leveraging the large volume of data continuously being updated, integrated, and shared. However, there are significant challenges as well, largely due to the heterogeneity of how these data are formatted. Indeed, although there is movement toward structured reporting in radiology (ie, hierarchically itemized reporting with use of standardized terminology), the majority of radiology reports remain unstructured and use free-form language. To effectively "mine" these large datasets for hypothesis testing, a robust strategy for extracting the necessary information is needed. Manual extraction of information is a time-consuming and often unmanageable task. "Intelligent" search engines that instead rely on natural language processing (NLP), a computer-based approach to analyzing free-form text or speech, can be used to automate this data mining task. The overall goal of NLP is to translate natural human language into a structured format (ie, a fixed collection of elements), each with a standardized set of choices for its value, that is easily manipulated by computer programs to (among other things) order into subcategories or query for the presence or absence of a finding. The authors review the fundamentals of NLP and describe various techniques that constitute NLP in radiology, along with some key applications.

Introduction

Guidelines for diagnostic imaging reports stress the importance of clear communication (1). To this end, electronic medical record (EMR) systems have been adopted to expedite communication and reduce risk for communication errors (2). Beyond improving the quality of patient care, EMR systems hold the promise of significantly advancing clinical research and practice by enabling analysis of the wealth of data contained in radiology reports (3)—for example, for clinical decision support (CDS), quality assurance and performance monitoring, hypothesis testing, and patient eligibility screening. Despite the movement toward structured radiology reporting (4,5), such as the Breast Imaging Reporting and Data System, the vast majority of reports at present use unstructured narratives separated into sections (eg, history and findings). Thus, although EMR systems offer electronic access to radiology reports, the concepts and events recorded within them are encumbered by the inherent ambiguity of human language, making searches difficult to automate.

TEACHING POINTS

- Natural language processing (NLP) is a computer-based approach that analyzes free-form text or speech by using a set of theories and technologies, including linguistics (ie, the scientific study of language form, meaning, and context) and statistical methods that infer rules and patterns from data, to convert the text into a structured format of hierarchically itemized elements with a fixed organization and standardized terminology for each element, such that the text is easily queried and manipulated.
- As a first step, NLP analyzes the text to identify individual concepts and their modification by other terms. When this process has been completed, each individual concept found in the text is ideally output as a separate item in a structured format that includes other important concepts that modify it (eg, anatomic location or chronicity). The primary NLP technologies used for this task are pattern matching and linguistic analyses.
- The second step is to determine whether the structured data extracted from a report contain one or more desired concepts and modifiers that indicate that the report possesses one or more specified characteristics with a given certainty (eg, positive for a specific disease). This step can be achieved by using a set of clinical rules developed by an expert with domain knowledge or, alternatively, by using statistical or machine learning approaches to automatically infer rules and patterns from a set of data.
- In general, the use of concepts identified by linguistic NLP as features in a machine learning–based classification algorithm can often yield better results compared with simple text features such as word *n*-grams, since a concept is likely to be more strongly associated with a desired classification compared with each individual synonymous term that can be used to describe it.
- Extraction of key information from free-text radiology reports with NLP has been used to enable large-scale testing of CDS, quality assurance and performance monitoring, and appropriate use of imaging, as well as to facilitate patient eligibility screening for clinical trials and hypothesis testing.

Natural language processing (NLP) is a computer-based approach that analyzes free-form text or speech by using a set of theories and technologies, including linguistics (ie, the scientific study of language form, meaning, and context) and statistical methods that infer rules and patterns from data, to convert the text into a structured format of hierarchically itemized elements with a fixed organization and standardized terminology for each element, such that the text is easily queried and manipulated (Fig 1) (6). Central to this process is the use of a standardized terminology for each concept that is of fundamental interest in a particular field. A concept is an intrinsically unique entity with an unambiguous meaning (eg, a specific disease such as myocardial infarction or a symptom such as chest pain) (Fig 2). Lexicons are collections of unique concepts accompanied by a preferred "term" (name) and a list of synonyms and derivational forms. One medical lexicon used in radiology is the UMLS Metathesaurus, which (for example) defines myocardial infarction by the unique alphanumeric code (or CUI) C0027051. It also includes synonymous terms for each concept (eg, "heart attack" and "cardiac infarction" for CUI C0027051) (7); specific semantic roles (types) for each concept (eg, disease, organ, or anatomic location); and relationships between concepts (eg, "is–a" relationships). Such a formal collection of concepts and their types and relationships is referred to as an ontology. Other useful lexicons and ontologies for NLP in radiology are SNOMED-CT, which is included in the UMLS Metathesaurus; and RadLex, which offers further radiology-specific terms, including devices and imaging techniques (8).

In this article, we review the fundamentals of NLP and describe various techniques that constitute NLP in radiology, along with some key applications.

Fundamentals of NLP

The overall goal of using NLP in radiology is to determine which concepts are mentioned in a clinical report and in what capacity. For example, a user may wish to identify reports with a particular imaging finding for outcome validation studies or patient eligibility screening. As a first step, NLP analyzes the text to identify individual concepts and their modification by other terms. When this process has been completed, each individual concept found in the text is ideally output as a separate item in a structured format (Fig 3) that includes other important concepts that modify it (eg, anatomic location or chronicity). The primary NLP technologies used for this task are pattern matching and linguistic analyses. The second step is to determine whether the structured data extracted from a report contain one or more desired concepts and modifiers that indicate that the report possesses one or more specified characteristics with a given certainty (eg, positive for a specific disease). This step can be achieved by using a set of clinical rules developed by an expert with domain knowledge or, alternatively, by using statistical or machine learning approaches to automatically infer rules and patterns from a set of data. Some of the technologies used for each of these steps are described in the following sections.

Pattern Matching

Pattern matching is the simplest, most fundamental technique for searching text and is an integral part of more complex NLP tasks. A pattern is a sequence of characters that can be matched, character for character, to a given text. For example, the pattern "he" can be matched twice in the sentence "*He* said *he*llo." To increase its generalizability, pattern matching makes use

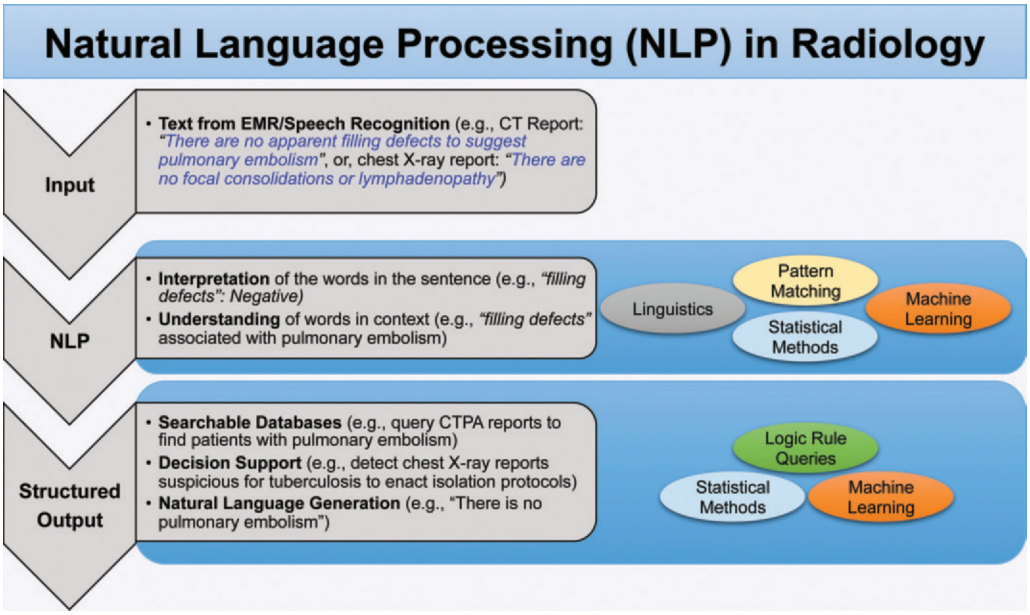
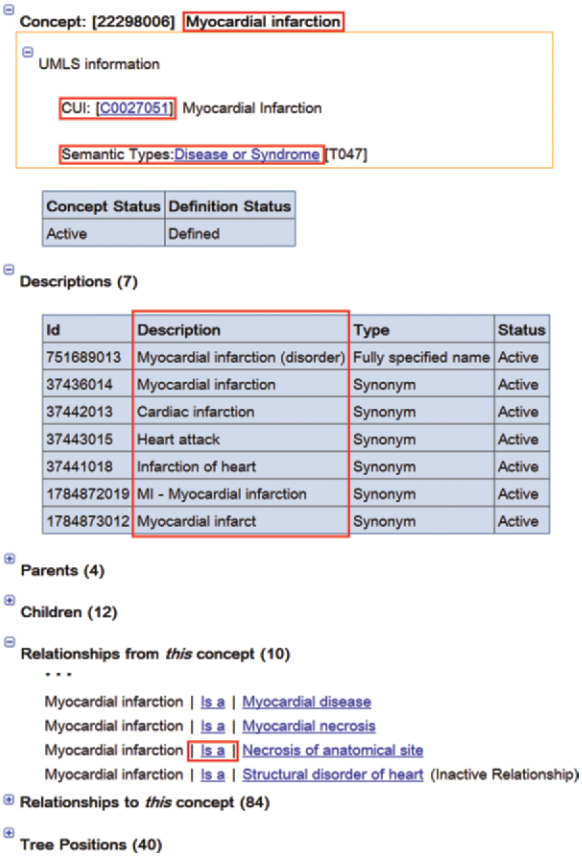

Natural Language Processing (NLP) in Radiology
Input
• Text from EMR/Speech Recognition (e.g., CT Report: "There are no apparent filling defects to suggest pulmonary embolism", or, chest X-ray report: "There are no focal consolidations or lymphadenopathy")
NLP
• Interpretation of the words in the sentence (e.g., "filling defects": Negative)
• Understanding of words in context (e.g., "filling defects" associated with pulmonary embolism)
Linguistics
Pattern Matching
Statistical Methods
Machine Learning
Structured Output
• Searchable Databases (e.g., query CTPA reports to find patients with pulmonary embolism)
• Decision Support (e.g., detect chest X-ray reports suspicious for tuberculosis to enact isolation protocols)
• Natural Language Generation (e.g., "There is no pulmonary embolism")
Logic Rule Queries
Statistical Methods
Machine Learning

Figure 1. Chart illustrates how NLP as understood in present-day radiology is a collection of various techniques that aim to extract information from natural language (eg, analyze a radiology report to extract concepts of interest and put them in a structured format) but that also use this output to (for example) index reports in a searchable database, provide patient- or report-level classification, or summarize findings in simpler natural language. CT = computed tomography, *CTPA* = CT pulmonary angiography.

Figure 2. Medical ontology (in this example, Systematized Nomenclature of Medicine–Clinical Terms [SNOMED-CT]) shows a unique concept and its description. SNOMED-CT provides a unique code for the concept (22298006) and its preferred name (myocardial infarction), the Unified Medical Language System (UMLS) concept unique identifier (CUI) and semantic type (disease or symptom), a list of synonyms (eg, cardiac infarction) for this concept, and relationships with other concepts.



⊟ Concept: [22298006] Myocardial infarction

⊟ UMLS information

CUI: [C0027051] Myocardial Infarction

Semantic Types:Disease or Syndrome [T047]

Concept Status	Definition Status
Active	Defined

⊟ Descriptions (7)

Id	Description	Type	Status
751689013	Myocardial infarction (disorder)	Fully specified name	Active
37436014	Myocardial infarction	Synonym	Active
37442013	Cardiac infarction	Synonym	Active
37443015	Heart attack	Synonym	Active
37441018	Infarction of heart	Synonym	Active
1784872019	MI - Myocardial infarction	Synonym	Active
1784873012	Myocardial infarct	Synonym	Active

⊞ Parents (4)

⊞ Children (12)

⊟ Relationships from *this* concept (10)

. . .

Myocardial infarction | Is a | Myocardial disease

Myocardial infarction | Is a | Myocardial necrosis

Myocardial infarction | Is a | Necrosis of anatomical site

Myocardial infarction | Is a | Structural disorder of heart (Inactive Relationship)

⊞ Relationships to *this* concept (84)

⊞ Tree Positions (40)

of so-called regular expressions, or sequences of characters and special symbols that explicitly define a character pattern to be searched for. The special symbols add multiple degrees of freedom to the search pattern specification (eg, by means of wildcards, character classes, quantifiers, and boundary matchers). In the programming language Java, for example, in the pattern “\bembol[a-z]*\b,” “\b” represents a word boundary, “[a-z]” refers to any lowercase letter, and “*” means “match [a-z] zero or more times.” Thus, the pattern will match any word that starts with “embol,” such as “embolus” and “emboli.” Another example would be “\bdila(?:ta)?tion\b,” which would match text containing either “dilation” or “dilatation,” as might be stated in connection with (for example) the aorta.

A strength of regular expression pattern matching in medical NLP lies in the fact that terms related to a given concept often share a common root, or stem. For example, “pneumonia” and “pneumonitis” share the common stem “pneumon.” This stem can be matched to a multitude of terms that are most likely related to concepts involving the lungs. Although knowledge of the

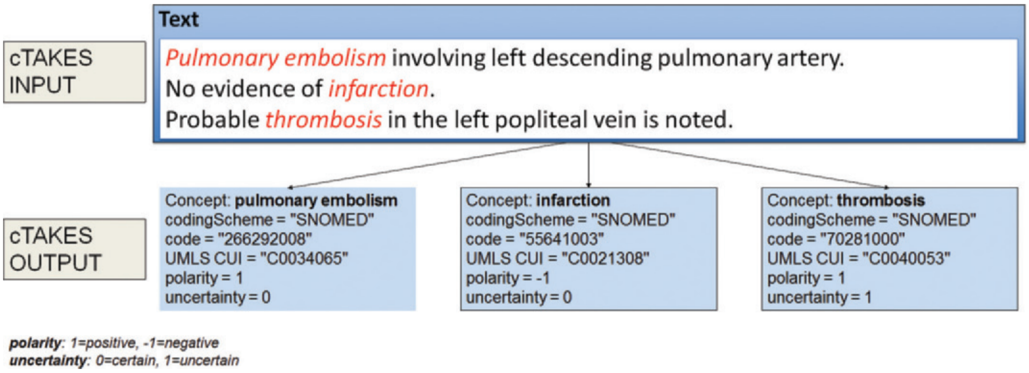


Figure 3. Diagram illustrates Text Analysis and Knowledge Extraction (cTAKES), an NLP system designed specifically for extracting information from clinical text. Text from a radiology report when input into cTAKES is analyzed to produce a list of individual concepts identified from a terminology of medical terms (in this example, both SNOMED-CT code and UMLS Metathesaurus CUI). Each concept is also assigned a "polarity" based on whether cTAKES recognizes the finding mentioned as present or absent (eg, no evidence of infarction is assigned a polarity of –1). A degree of certainty is also assigned. In this example, because of the word "probable," the corresponding concept is coded as uncertain.

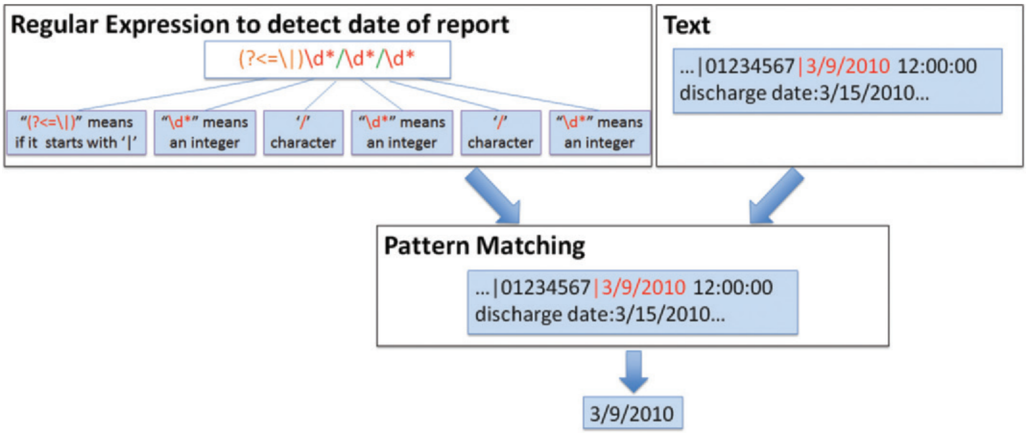

Regular Expression to detect date of report
(?<=\|)\d*/\d*/\d*
"(?<=\|)" means if it starts with '|'
"\d*" means an integer
'/' character
"\d*" means an integer
'/' character
"\d*" means an integer
Text
...|01234567|3/9/2010 12:00:00 discharge date:3/15/2010...
Pattern Matching
...|01234567|3/9/2010 12:00:00 discharge date:3/15/2010...
3/9/2010

Figure 4. Diagram illustrates a pattern matching process designed to extract report dates. A regular expression (upper left) is designed to detect the date in the header of each report stored in our EMR system. Reports have a header that consists of a numeric string (the EMR number) enclosed by the character "|" and followed by a date (upper right). When the pattern matching process encounters a character sequence that matches this pattern, the date is displayed (bottom).

stem of interest allows one to perform more general searches, the inverse process—word stemming—is also useful in NLP. Word stemming uses the knowledge of language morphology to reduce a given word to its root. In this manner, reports can be standardized by (for example) replacing each word with its stem. The aggregate of stems in a report can then be more readily searched for the stems related to the concept of interest (9–11).

Another common use of pattern matching is to break text into "chunks" and "tokens," or conceptually meaningful subparts. Subparts can be report sections, individual sentences, or words. For example, word segmentation (breaking text into individual words) can be performed on the basis of pattern matching spaces and punctuation, or a single text file containing concatenated reports from a set of patients can be broken into individual reports with a regular expression designed to find the individual report "header." Figure 4 shows the regular expression designed to find the EMR number and extract the date for the output format of the EMR system at our institution.

Regular expression pattern matching is often used to accomplish limited linguistic analyses (described in the following section). One example of a linguistic task would be to determine whether a concept contained in a sentence is described as being present, absent, uncertain, or an alter-association (ie, pertaining to a different subject, such as the history of a family member) (12). A more limited task would be to detect whether a concept falls within the scope of a negation phrase. For example, the NegEx algorithm uses pattern matching to search for negation lexical cues within a small number of words before and after the mention of a UMLS sign, symptom, or

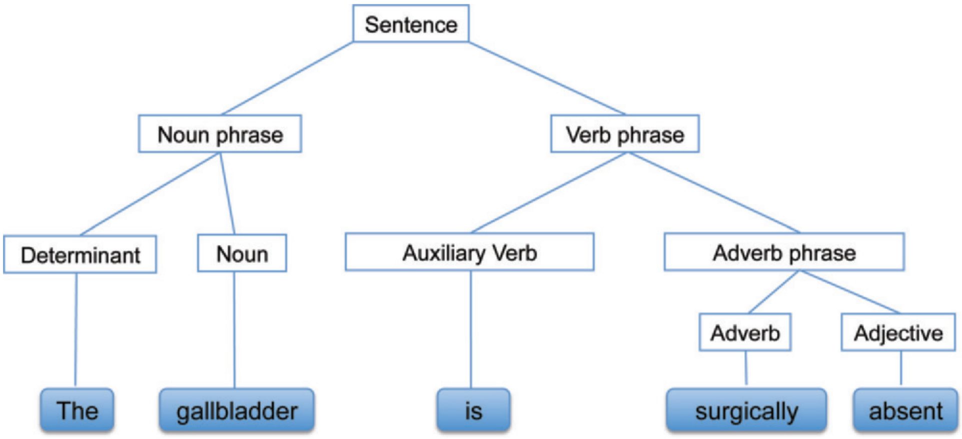
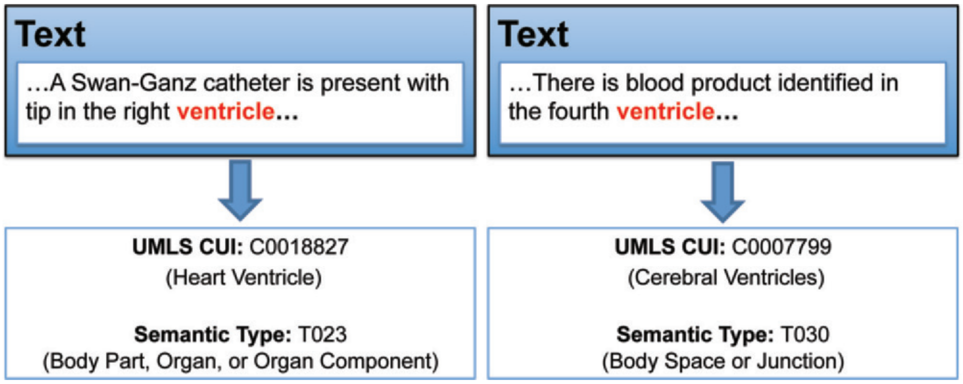


Figure 5. Diagram illustrates the syntactic analysis of the sentence "The gallbladder is surgically absent." Each word (except "The") is assigned a part-of-speech designation using grammatical rules. Linguistic NLP systems often perform such analyses to identify sentence subparts that might correspond to specific medical concepts.



Text

...A Swan-Ganz catheter is present with tip in the right **ventricle...**

UMLS CUI: C0018827
(Heart Ventricle)

Semantic Type: T023
(Body Part, Organ, or Organ Component)

Text

...There is blood product identified in the fourth **ventricle...**

UMLS CUI: C0007799
(Cerebral Ventricles)

Semantic Type: T030
(Body Space or Junction)

Figure 6. A challenge in NLP is that ambiguous terms can be interpreted in more than one way depending on the context in which they are used. For example, this diagram shows how the word "ventricle" can refer to two distinct concepts in the UMLS Metathesaurus terminology. Beyond distinct UMLS CUIs, these particular concepts also have distinct semantic types, broad categories of concepts that are described in the UMLS Semantic Network. Each concept may be assigned to one or more semantic types.

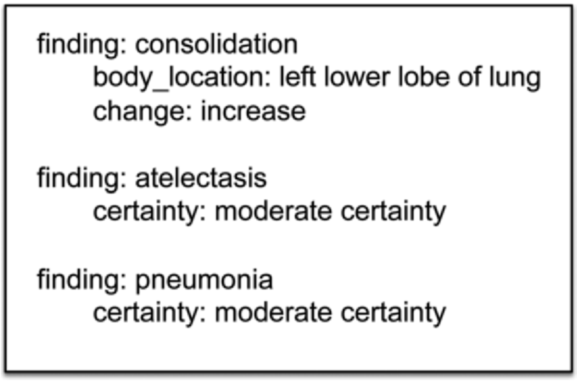
disease concept, such as the word "no" preceding the mention of "myocardial infarction" or the word "negative" succeeding it. This simple method has a specificity of 94.5% and a sensitivity of 77.8% for detecting negations (13).

Linguistic Approach

Linguistic NLP systems treat words as symbols that have been put together based on grammatical rules that define what associations between symbols are meaningful—for example, using their part-of-speech designation (eg, verb, noun, adjective) (Fig 5). A computer algorithm uses this knowledge, both syntactic (ie, the rules that control the arrangement of words in a sentence) and semantic (ie, knowledge regarding the different meanings of words in the context of a sentence), to infer what concepts are mentioned and how each concept modifies other concepts.

In addition to syntax and semantics, natural language is based on other components such as phonetics, morphology (the formation and internal structure of words), and pragmatics (pairing words or sentences with concepts for which they would be appropriate) (14). However, because the use of language in clinical reports is more limited than that in general text (13), most NLP systems in radiology achieve sufficient accuracy with the syntactic and semantic approaches only. Instead, the major shortcomings are ambiguity, wherein an expression can be interpreted in two or more distinct ways depending on context (Fig 6); incorrect grammar usage in a fast-paced environment; and misspellings. Despite these challenges, linguistic NLP systems theoretically offer more complete information than pattern matching–based systems and have thus been preferred whenever complex (eg, temporal, anatomic) relationships are of interest (15).

One example of a linguistic NLP package used in radiology is Medical Extraction and Encoding



```
finding: consolidation
      body_location: left lower lobe of lung
      change: increase

finding: atelectasis
      certainty: moderate certainty

finding: pneumonia
      certainty: moderate certainty
```

Figure 7. Simplified example of the structured format generated by an NLP system (MedLEE) as a result of processing the text "increased consolidation of the left lower lobe compatible with atelectasis or pneumonia." MedLEE has been used to extract information from radiology reports for a variety of research and CDS purposes. (Reprinted, with permission, from reference 18.)

(MedLEE) (16), originally developed at New York Presbyterian Hospital to process chest radiography reports using semantic grammar (17,18). Its output is a list of findings, along with any modifiers for those findings, returned in a structured format. Fields include temporal, anatomic, and certainty modifiers for each finding (Fig 7), whose values can be obtained from existing terminologies (eg, the UMLS Metathesaurus) and custom-built dictionaries developed for specific tasks. The open-source NLP package cTAKES similarly relies on significant linguistic components (19,20) in conjunction with statistical and machine learning approaches (discussed in the following section).

Linguistic analysis does not recognize medical concepts in a terminology such as the UMLS Metathesaurus in and of itself. For example, "myocardial infarction" and "heart attack" refer to the same disease concept but could be considered distinct findings; one is an "infarction" in the "myocardium" anatomic location, the other an "attack" in the "heart" anatomic location (17). Matching these findings to a standardized terminology, wherein both would result in identification of the same disease concept of myocardial infarction, requires an additional, nontrivial step. Friedman et al (17) summarized techniques for achieving this task and developed the approach used in MedLEE, wherein a standardized terminology is first analyzed to generate a structured format for each variation of each concept in the terminology, including synonyms and multiword variations. Each of these structures is stored and directly compared with those extracted by MedLEE from a given text. The National Library of Medicine's MetaMap is a similar, freely available tool that also relies on NLP to identify UMLS Metathesaurus concepts mentioned in text (21). MetaMap works by generating all possible variants of a finding encountered in the text (eg, replacing an occurrence of "eye" with "ocular," "oculus," and "optic") and then scoring each variant against all concepts contained in the terminology, with the highest-scoring concept being identified as the match.

Statistical and Machine Learning Approaches

Statistical and machine learning methods infer rules and patterns directly from data. They are deeply interwoven into all aspects of NLP. For example, statistical methods are used in a cTAKES component to predict whether a concept such as a disorder is described as present, absent, possible, or part of past medical or family history. Another example is the Statistical Assertion Classifier (StAC) algorithm, which achieves similar results as NegEx but with no linguistic knowledge, instead using a machine learning algorithm to "learn" what negations are by examining reports for which a human has previously determined whether a negation is present (22). Uses of machine learning to achieve linguistic tasks are typically hidden from the user. Instead, the reader will most often directly encounter a discussion of machine learning methods whenever NLP output is used to predict a document- or patient-level classification.

Many methods can be used to accomplish the classification task. The simplest strategy is to use a clinical "logic rule" that is "true" when a report contains a combination of findings. For example, if pneumonia, pneumonitis, infiltrates, or consolidations and other opacities are identified in a chest radiography report by an NLP system, the report is likely positive for pneumonia (23). Such clinical rules are developed on the basis of expert knowledge and can be readily understood and extended by others; however, they are often cumbersome to generate, since they must synthesize a multitude of concepts and all of their sensible combinations. Statistical and machine learning methods perform classification tasks by analyzing data to automatically determine what features are associated with (for example) a positive versus a negative result (Fig 8).

There are three elements in statistical and machine learning methods: features, training data, and models. Features are any characterizations of the subjects of analysis. For example, if one wishes to infer a rule for obesity, body weight and height are relevant features. One of the simplest and most useful features in NLP is the *n*-gram (ie, *n* consecutive words in a text). For instance, unigrams (n = 1) are the individual words in a text, and bigrams (n = 2) are every pair of consecutive words. By examining the *n*-grams in a text, it is possible to guess the topic and, thus, classify it. For example,

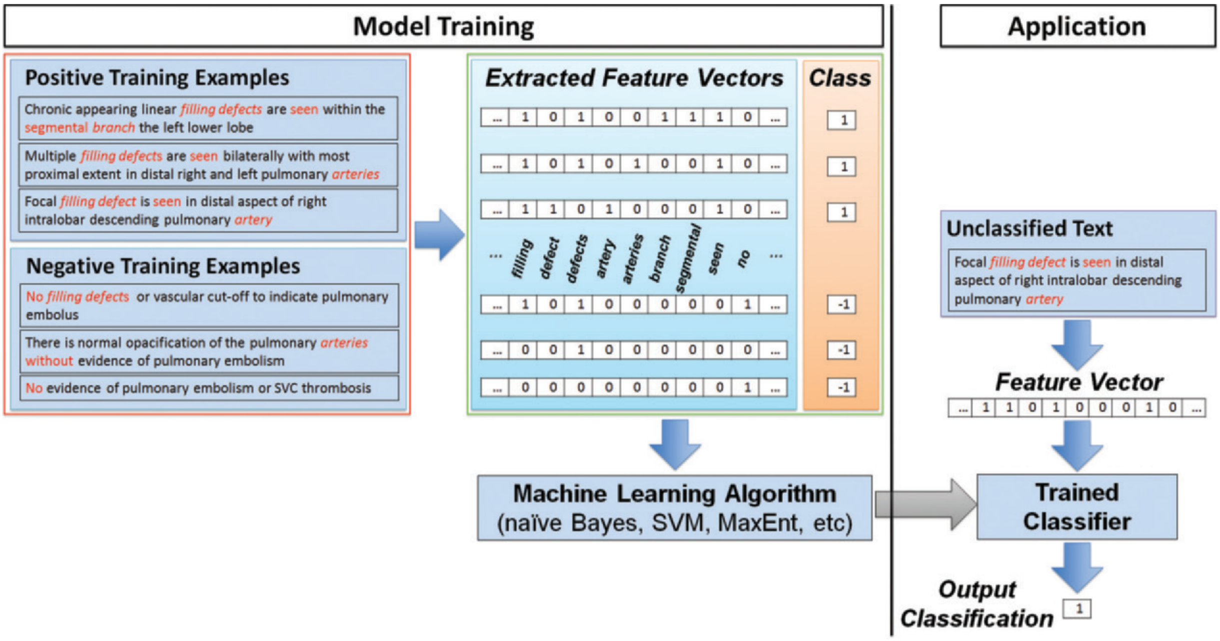

Model Training
Application
Positive Training Examples
Chronic appearing linear *filling defects* are seen within the segmental *branch* the left lower lobe
Multiple *filling defects* are seen bilaterally with most proximal extent in distal right and left pulmonary *arteries*
Focal *filling defect* is seen in distal aspect of right intralobar descending pulmonary *artery*
Negative Training Examples
No *filling defects* or vascular cut-off to indicate pulmonary embolus
There is normal opacification of the pulmonary *arteries* without evidence of pulmonary embolism
No evidence of pulmonary embolism or SVC thrombosis
Extracted Feature Vectors
Class
... 1 0 1 0 0 1 1 1 0 ... 1
... 1 0 1 0 1 0 0 1 0 ... 1
... 1 1 0 1 0 0 0 1 0 ... 1
... filling defect defects artery arteries branch segmental seen no ...
... 1 0 1 0 0 0 0 0 1 ... -1
... 0 0 1 0 0 0 0 0 0 ... -1
... 0 0 0 0 0 0 0 0 1 ... -1
Machine Learning Algorithm (naïve Bayes, SVM, MaxEnt, etc)
Unclassified Text
Focal *filling defect* is seen in distal aspect of right intralobar descending pulmonary *artery*
Feature Vector
... 1 1 0 1 0 0 0 1 0 ...
Trained Classifier
Output Classification
1

Figure 8. Diagram illustrates how machine learning algorithms are an integral part of linguistic NLP systems. Most important, these algorithms, such as support vector machine *(SVM)* or maximum entropy *(MaxEnt)* models, are used for patient- or report-level classification. They rely on analyzing a set of features used to describe each training example to determine a model that best separates positive (class = 1) from negative (class = –1) examples. Features are typically thought of as vectors whose entries can be as simple as the frequency with which individual words appear in each example, but they can also be based on the structured information extracted from each example using linguistic NLP systems. Following this model training, the trained classifier is applied to a new text of unknown classification by extracting the same features used to train it. *SVC* = superior vena cava.

"consolidations" is likely to appear in a chest radiography report that is positive for pneumonia, but the likelihood that the report identifies pneumonia is reduced if the consolidations are described as "chronic" and, similarly, is increased if they are concerning for "infection." Although *n*-grams are powerful features used in many practical systems such as speech recognition, concepts identified by NLP can be more predictive as features, since NLP reduces synonymous findings to standardized names (15). However, this is not always the case because words that are relevant for some characterizations may not have concepts, such as the words "if" and "further," which have been found to be highly predictive for identifying reports with follow-up recommendations (24), thus missing important features.

Statistical and machine learning methods require a second element: training data that include a criterion standard classification (the "correct answer"). These data are then used to establish a link between the features and the class. A large number of criterion standard labels, although desirable for ensuring stability of the fitted model, may not be realistic due to the cost associated with manual chart review. In practice, the number of training data for different learning tasks may vary; typically, however, a few hundred data are sufficient for most tasks (25), provided that the number of candidate features is not exceedingly large (eg, <100). Care must also be taken in the choice of data because the performance of the resulting classifier may depend on the training set (eg, favoring the majority class of training examples) (9)—for instance, yielding a classifier that will more readily produce a false-negative result than a false-positive result, if the training examples are mostly negative ones.

The third element of statistical and machine learning is the model used to relate the class of interest to the features. One category of models is based on probability theory. A simple model that is often used in document classification is the naïve Bayes model, which assumes that the values of the features are independent from each other once the class is fixed (conditional independence). Under this assumption, the joint probability of the observed features can easily be calculated, and the class that maximizes this joint probability is chosen as the prediction. Despite its simplicity, use of the naïve Bayes model with *n*-grams can be difficult to outperform in many scenarios, even with more advanced techniques. Another category of models seeks to establish a more direct link between features and class in the form of a mathematic function, wherein the function parameters are estimated from the training data in a procedure called "model fitting." One such model is known as logistic regression, whose result is a number between 0 and 1 that can be interpreted as the probability that the document belongs to a given class. The classification is not limited to a binary format, and the multiclass logistic regression is also known as the maximum entropy classifier. Another powerful classification model that has become very popular in recent years is the support vector machine, which implicitly maps the features to a much higher dimensional space so as to derive many complex features automatically from the existing one, giving the model much better adaptivity. Both the maximum entropy and support vector machine models are often encountered in radiology NLP applications (22,24,26–29).

Although machine learning–based techniques offer the promise of fully automated organization and extraction of information from radiology reports, there are pitfalls to avoid when using these models. For instance, it is generally undesirable to use all possible features, such as all words or *n*-grams in the reports being analyzed, because using too many uninformative features causes the model to overly adapt to the training data, a phenomenon known as overfitting. Another issue is that fitted models may not be portable; if, for example, the language pattern of reports at a medical institution is different from that of the training data (eg, the reports use different terms or are simply longer), the distribution of the features may be different, and the method may need to be retrained before application at the new institution. Finally, a potential disadvantage is that the model cannot always be as easily evaluated by a human as can (for example) a set of clinical logic rules.

Applications of NLP in Radiology

Most applications of NLP in radiology use a combination of the aforementioned technologies, typically in the form of a cascade (pipeline). For example, simple pattern matching might first be used to separate report sections and sentences and potentially retain only those that are relevant to the task. Subsequently, linguistic NLP can be used to (for example) detect concepts and modifiers such as negations. Finally, a carefully crafted clinical logic rule or a statistical or machine learning algorithm is typically applied to group reports into desired classes. Although the technologies are very different, they can (as mentioned earlier) be used to achieve the same tasks. For example, pattern matching or machine learning can be used for linguistic tasks, whereas linguistic NLP tools most often rely on some machine learning and certainly pattern matching components. Similarly, classification can be as simple as pattern matching for a single keyword or concept. In the following sections, we review results that have been achieved in radiology with NLP and demonstrate how the different technologies were used to achieve them.

Chest Radiography and CT

One of the first explorations of NLP in radiology was by Knirsch et al (30), who compared MedLEE with expert review in identifying chest radiography reports that were suspicious for tuberculosis in patients with a subsequent positive culture. An agreement of 89%–92% was achieved by focusing on whether six selected keywords (eg, “infiltrate”) appeared in the report (30). The system was incorporated into a CDS infrastructure for implementing respiratory isolation protocols, yielding improved isolation rates compared with the clinical protocol. MedLEE also enabled one of the earliest large-scale applications of NLP in radiology for the testing of four hypotheses, such as the observations that *(a)* lung cancer occurs more commonly in the right lung, and *(b)* the frequency of bullet and stab wounds decreased along with the reduction in U.S. crime rates in 889,921 chest radiography reports obtained over a 10-year period (31). The system had a sensitivity of 81% and a specificity of 99% for identifying 24 abnormal findings in a subset of 150 reports (31), results that were considered similar to those achieved with human coders. This study also found that NLP was more accurate than financial discharge ICD-9-CM (International Classification of Diseases, Ninth Revision, Clinical Modification) coding for pneumothorax: Over a period of 1 month, financial discharge coding had a sensitivity of 17% versus 100% for NLP, whereas both had nearly 100% specificity compared with expert assignment (31). Despite its limitations, ICD-9-CM is often used to identify patient cohorts for epidemiologic, cost, and outcome analyses. NLP can deliver significant advantages in this arena (32,33). Dublin et al (34) leveraged a linguistic NLP system named ONYX to identify chest radiography reports that could not be immediately classified as consistent or inconsistent with pneumonia but that instead required manual review because of complex statements (eg, those indicating change in status, such as an improving or resolved infiltrate). The system classified 88% of 5000 reports as consistent or inconsistent with pneumonia with a sensitivity of 75% and a specificity of 95%, classifying the remaining reports as requiring manual review (34).

As mentioned earlier, machine learning approaches can often achieve results similar to those of expert knowledge. This was shown in one study in the setting of classifying chest radiography reports for the presence of acute lung injury, a diagnosis that is rarely reported explicitly but rather is identified on the basis of multiple clinical features (29). In one arm of the study, reports were scored on the basis of a weighted sum of keywords provided by experts. Keywords were detected using pattern matching in conjunction with NegEx to exclude negations, since those would have otherwise resulted in negative findings such as "no pulmonary edema" being matched with the expert-provided keyword "pulmonary edema." In the other arm, a machine learning algorithm was trained using word unigrams and character *n*-grams ranging from one to 14 characters. In a study by Solti et al (35), machine learning with six-character *n*-grams performed as well as or better than the expert-provided keywords in 857 reports. An example of a six-character sequence that machine learning ranked high for positive reports was "y opac," likely analogous to the term "patchy opacities" that experts identified as an important keyword (35).

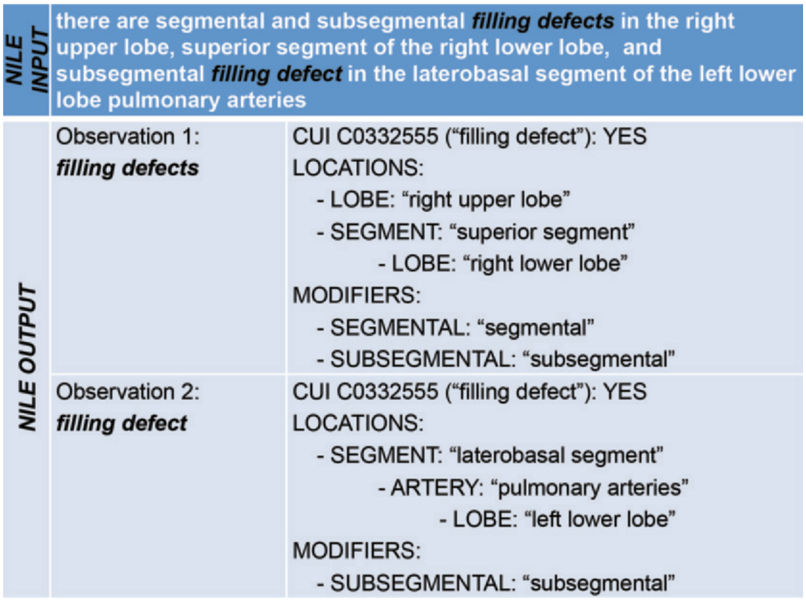
One advantage of using linguistic NLP to detect concepts defined in standardized medical terminologies as features for the subsequent classification task is that concepts can be quickly applied and reused by others. Elkin et al (23) used the Multi-threaded Clinical Vocabulary Server with SNOMED-CT coding to identify pneumonia in chest radiography and CT reports by using a simple logic rule query that consolidated concepts relevant to pneumonia, including infiltrates, consolidations, and other pulmonary densities. Compared with manual review, this approach had a high sensitivity (100%) and specificity (98%) (23). Mendonça et al (18) used MedLEE to identify cases of health care–associated pneumonia by extending a logic rule query previously developed to detect the mention of acquired pneumonia in adult chest radiography reports. The final query, consolidating 38 findings and modifier-finding combinations, differentiated health care–associated pneumonia with high specificity (99%) but only moderate sensitivity (71%) compared with chart review in 1277 neonates admitted to an intensive care unit over a 2-year period (18). The majority of false-positive results were due to negative microbiology cultures despite correct NLP identification of radiographic findings corresponding to pneumonia. In contrast, misclassifications traced to the NLP system were primarily due to grammatical errors, misspellings, and abbreviations, which remain a difficult problem for NLP.

Another advantage of linguistic NLP is that it allows the combining of findings from diverse EMR data. Friedman et al (36) categorized patients with community-acquired pneumonia into one of five risk classes by independently analyzing both discharge summaries and chest radiography reports using MedLEE. The overall system had 80% accuracy for categorizing patients into the correct risk class and 100% accuracy for categorizing them into a risk class no more than one risk class above or below the correct one (36). NLP accuracy was higher for chest radiography reports (96%) than for discharge summaries (93%), likely due to the smaller variability of potential findings in radiography reports, as well as the difficulty of extracting numeric data such as vital signs from discharge summaries, a task that had an accuracy of only 85% (36). Correctly identifying numeric findings also remains a difficult problem for NLP.

Despite the advantages of linguistic NLP, limited linguistic analyses using pattern matching and statistical and machine learning techniques can offer similar accuracies for well-defined tasks. Fiszman et al (37) used a system called SymText to detect acute bacterial infection in chest radiography reports to replace a keyword search ("pnemoni*," "aspirati*," "infiltr*") previously used in a CDS system that aids in selection of appropriate antibiotics in conjunction with laboratory and microbiologic data. SymText was developed to process chest radiography reports for extraction of 76 different radiographic findings and 89 different disease concepts using a word unigram–based statistical model (Bayesian network) designed with syntactic, semantic, and clinical knowledge. A logic rule of findings extracted by SymText was used to determine the presence or absence of pneumonia in 292 reports. Compared with classification by a majority vote of three physicians, this system had 95% sensitivity and 85% specificity, compared with 94% and 91%, respectively, for consensus interpretation by four physicians and 83% and 74%, respectively, for simple keyword searches performed with pattern matching (37).

Pulmonary Embolism

A recent application of NLP has been to classify CT pulmonary angiography report findings with respect to pulmonary embolism disease outcome, such as severity based on thrombus location (eg, central versus subsegmental). Chapman et al (38) developed an application called peFinder for classification of disease presence, chronicity, and certainty, as well as examination technical quality. The peFinder application is based on an extension of NegEx to detect lexical cues other than negations and define how each cue modifies a preceding or succeeding concept. This simple system had high sensitivity (86%–98%) and specificity (89%–93%) for each task except chronicity (60% and 99%, respectively) (38). Yu et al (15) used an NLP system called Narrative Information Linear Extraction (NILE) that combines linguistic and machine learning approaches to improve identification of pulmonary embolism location. Although many full-linguistic NLP packages such as MedLEE and cTAKES can identify the anatomy mentioned with a finding, pulmonary embolism presents a particular challenge because of the relevance of the inherently nested anatomic structure of the vasculature (Fig 9). Using NILE's output as features in



NILE INPUT	there are segmental and subsegmental ***filling defects*** in the right upper lobe, superior segment of the right lower lobe, and subsegmental ***filling defect*** in the laterobasal segment of the left lower lobe pulmonary arteries	
NILE OUTPUT	Observation 1: ***filling defects***	CUI C0332555 ("filling defect"): YES LOCATIONS: - LOBE: "right upper lobe" - SEGMENT: "superior segment" - LOBE: "right lower lobe" MODIFIERS: - SEGMENTAL: "segmental" - SUBSEGMENTAL: "subsegmental"
	Observation 2: ***filling defect***	CUI C0332555 ("filling defect"): YES LOCATIONS: - SEGMENT: "laterobasal segment" - ARTERY: "pulmonary arteries" - LOBE: "left lower lobe" MODIFIERS: - SUBSEGMENTAL: "subsegmental"

Figure 9. Chart illustrates a simplified example of the structured format generated by the NILE NLP system, which combines linguistic and clinical knowledge. NILE can identify concepts and recognize the anatomic relationships between location modifiers and these concepts. This information can then be used to classify pulmonary embolism into (for example) central, segmental, or subsegmental categories.

a machine learning classifier, the system was able to achieve a receiver operating characteristic area under the curve of 0.998 to detect the presence of pulmonary embolism, 0.945 to detect acute pulmonary embolism, and 0.945 and 0.987 to detect central and subsegmental pulmonary embolisms, respectively (15).

Both of these studies compared the use of linguistic NLP-extracted features versus word *n*-grams for machine learning–based classification. Although the use of word *n*-grams was also effective in detecting the presence of pulmonary embolism, linguistic NLP was superior for chronicity and location, tasks that require understanding of temporal and anatomic relationships. In general, use of concepts identified by linguistic NLP as features in a machine learning–based classification algorithm can often yield better results compared with simple text features such as word *n*-grams because a concept is likely to be more strongly associated with a desired classification compared with each individual synonymous term that can be used to describe it. However, for certain tasks, the benefit may be less pronounced. For example, machine learning–based classification of acute orbital fractures in emergency department CT reports obtained in 3710 consecutive patients who presented with blunt orbital trauma was only slightly improved with use of linguistic NLP to extract features (sensitivity of 93.3% versus 92.5% and specificity of 96.9% versus 93.3%, respectively) (26).

Conversely, whenever NLP can provide a benefit for machine learning–based classification, even limited linguistic analysis can suffice (eg, using pattern matching to determine "high-value" features). In one study, accuracy for each of three classification tasks in thromboembolic diagnoses (presence, CT technique, and clinically relevant incidental findings) was uniformly increased regardless of the machine learning algorithm used (naïve Bayes model, support vector machine, or maximum entropy) when pattern matching was used to identify relevant concepts and their relationships (eg, "nodule" as a condition and "lingula" as an anatomic structure) (28). The authors of that study also used an NLP-based automated anonymization tool named MEDINA (MEDical Information Anonymization) to identify and replace patient and physician information and to shift dates by a uniform random number (28). Such use of NLP is of particular relevance to research because one could potentially develop a technique to preserve temporal information in the collective EMR data of each individual patient being de-identified. Other investigators have relied on manual or purpose-written software to achieve patient-linked de-identification (26).

A direct application of NLP for detection of pulmonary embolism–positive reports is the validation of new clinical algorithms. One study confirmed the recommendations of a prospective European trial to improve the low specificity of ᴅ-dimer for determining appropriate use of CT pulmonary angiography for ruling out acute pulmonary embolism in aging populations (39). Using an NLP framework named GATE (General

Architecture for Text Engineering) with an accuracy of 98% in detecting pulmonary embolism compared with manual review, the authors were able to quickly validate two age-adjusted D-dimer cutoffs in a U.S. population (39). Dunne et al (40) used the same NLP system to assess the effect on the use and yield of CT pulmonary angiography in inpatients with suspected pulmonary embolism before and after implementation of a CDS system that uses evidence-based guidelines to assist in making the decision to order a study. Ideally, effective CDS systems will not only decrease the number of studies performed but also increase the diagnostic yield by eliminating unnecessary examinations. Application of NLP stands to quickly confirm these benefits. The authors tested this hypothesis in the 31-month period following CDS implementation, reporting a 12.3% decrease in monthly orders for CT pulmonary angiography, along with a nonsignificant increase in monthly yield of 16.3% (40). The nonsignificance of increase in yield may have been due to a concurrent campaign to promote venous thromboembolism prophylaxis for hospitalized patients at the authors' institution, which may have decreased disease prevalence.

Cancer

NLP was used early on in oncology for detection of findings suspicious for breast cancer in mammography reports (41) and cancer-related findings in chest radiography reports (42). The latter application is another example of the extensibility of classification systems based on expert-created logic rules using linguistic NLP. The authors extended a logic rule–based commercial NLP system designed to extract billing codes from reports (LifeCode; A-Life Medical, San Diego, Calif) with rules for findings and modifiers relevant to cancer. The system correctly identified 4347 of 5139 findings in 500 reports in 6 minutes, compared with 20 hours for manual coding by a board-certified internist (42).

NLP can also be used for more granular tasks than extraction of diagnoses. Two important examples that have been explored are cancer progression and recurrence. Cheng et al (9) used NLP to classify brain tumor progression in 778 consecutive follow-up magnetic resonance (MR) imaging reports (238 patients) that referenced a prior CT or MR imaging study (Fig 10). Classifications of status (progressed, stable, or regressed), magnitude of change (mild, moderate, or marked), and certainty of change (uncertain, possible, or probable) were performed by separate NLP systems; status classification was performed with a powerful machine learning algorithm (support vector machine) by using word stems plus negations detected by NegEx as features to detect the individual tumors being described in each report. This allowed the remaining two tasks to be constrained to within one sentence of each tumor status finding and accomplished with a simple pattern matching approach. This cascaded approach yielded a sensitivity and specificity of 80.6% and 91.6%, respectively, for overall status classification; 79.3% and 89.4%, respectively, for magnitude classification; and 68.6% and 85.9%, respectively, for certainty classification. Similarly, Carrell et al (43) used cTAKES to consolidate pathology and radiology reports plus clinical notes to detect cancer recurrence in women with early-stage invasive breast cancer. A custom-built dictionary with 1360 entries was created for pathologic findings. The dictionary for radiology reports and clinical notes included 4891 findings and more complex logic query rules necessary to integrate indirect evidence, such as a change in imaging findings over time. The system was able to reduce the number of patient charts that had to be manually reviewed to identify confirmed cases of breast cancer recurrence by 90%, while missing 8% of recurrent cases, similar to manual review (43).

Recommendation Practices and Communication of Critical Results

Some of the largest-scale applications of NLP in radiology to date have been the assessment of recommendations for additional imaging. Dreyer et al (25) developed an NLP system named Lexicon Mediated Entropy Reduction (LEXIMER) to classify reports based on whether they contain clinically important findings and recommendations. LEXIMER uses linguistic stemming to reduce each sentence to its root meaning and empiric principles to assign an "importance" weight to each resulting phrase (eg, later phrases in the impression section of a report have a higher likelihood of summarizing results in prior phrases and thus may have higher value). Trained using a machine learning approach on 200 CT and MR imaging reports, the system achieved a sensitivity and specificity of 98.9% and 94.9%, respectively, for detection of clinically important findings, and of 98.2% and 99.9%, respectively, for detection of recommendations for additional imaging in 1059 consecutive radiology reports across all major imaging modalities and subspecialties in a single hospital radiology department (25).

This system was used for the largest NLP-based analysis of rates of recommendations for additional imaging with respect to 11 factors, including radiologist experience, imaging modality, and body part examined, in 5.9 million radiology reports spanning multiple imaging modalities over 13 years in a large urban academic radiology de-

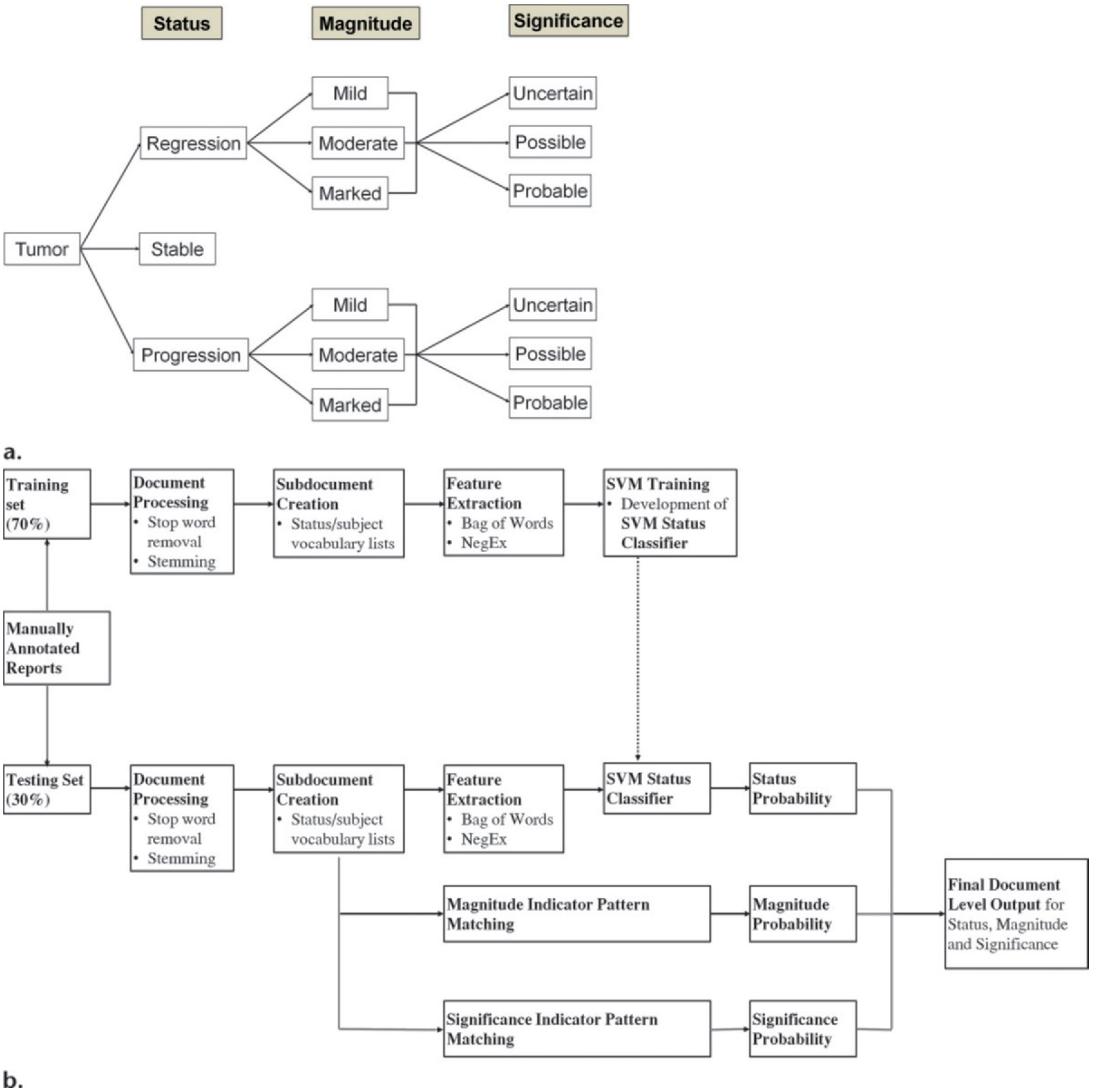


Figure 10. Retrieval of information regarding tumor progression from unstructured brain MR imaging reports. **(a)** Diagram illustrates the desired classification scheme for extracting structured information regarding disease status, magnitude of change, and significance of change. **(b)** Diagram illustrates how an NLP system is developed for a classification task using machine learning– and/or rule-based methods. *SVM* = support vector machine. (Fig 10 reprinted, with permission, from reference 9.)

partment (44). In addition to observing a doubling of the proportion of examinations that contained at least one recommendation for additional imaging during the study period (from 6% to 12%), the authors assessed correlations with imaging modality, body area examined, ordering service, and radiologist specialty (44). Such information can be used for quality improvement, performance quantification, and (potentially) feedback for individual radiologists on their recommendation rates. An extension of LEXIMER to extract the recommended time frame and suggested imaging technique for follow-up (with an estimated accuracy of 94.3% and 93.2%, respectively) found that 12.5% of 4.2 million reports generated between 1995 and 2004 at one institution contained a recommendation for subsequent action, 71.4% of which were for further imaging (45). The increase in recommendations for high-cost CT, MR imaging, and sonography (21%) outpaced the increase in volume of these examinations (14.4%) (45).

NLP-based identification of recommendations in radiology reports is also driven by quality assurance efforts to ensure expedient communication. Yetisgen-Yildiz et al (24) noted that miscommunication is the second most common cause of radiologist malpractice suits and developed a cascade of linguistic and machine learning methods to determine the likelihood that a sentence contained a recommendation. They compared unigrams with and without their part-of-speech assignment and the report sections in which they appeared as features. With use of all features, the system had a sensitivity of 64.6% and an accuracy of 99.7% in 800 reports representing various imaging modalities at one medical center, and the system was only marginally better than use of unigrams alone (62.8% sensitivity, 99.7% accuracy) (24). Importantly, this work assessed how class prevalence in the machine learning training set affected performance. Specifically, a training dataset containing roughly equal numbers of positive and negative examples yielded high sensitivity and low positive predictive value (97.3% and 25.1%, respectively), whereas with a training dataset with the expected prevalence of positive examples (1:165 ratio of sentences positive for recommendations in that study), sensitivity was negatively impacted but positive predictive value rose significantly (64.6% and 82.0%, respectively) (24). Reliance on the training set characteristics is one caveat associated with machine learning approaches and one reason why logic query rules based on expert knowledge are sometimes favored over machine learning.

Dutta et al (46) explored the related challenge of ensuring communication of incidental findings in radiology studies ordered in the emergency department. An initial attempt using keyword pattern matching to detect discharge imaging recommendations unrelated to the chief complaint achieved high sensitivity (98.6%) but low specificity (74%); when extended with logic rules to ensure that at least one keyword from two or more designated categories (eg, both the word "followup" and the word "CT") appeared in the same sentence, sensitivity was reduced to 88.9% and specificity increased to 98.2% (46). When the list of keywords was extended to reduce false-negative results and negation exclusion was added to reduce false-positive results, sensitivity and specificity were more balanced at 97.2% and 95.2%, respectively (46). Using their final system, the authors discovered that only 49% of discharge-relevant imaging recommendations made over a 32-day period in a tertiary care center with 24-hour emergency department radiology coverage were documented in the discharge paperwork (46).

A related quality assurance application of NLP concerns communication of critical findings to health care providers. Lakhani et al (47) developed NLP algorithms to detect nine such findings (acute pulmonary embolism, cholecystitis, appendicitis, ectopic pregnancy, testicular torsion, new or tension pneumothorax, unexplained intraperitoneal free air, increasing or new intracranial hemorrhage, and malpositioned enteric and endotracheal tubes) in radiology reports. An extensive algorithm for each finding, using regular-expression pattern matching for relevant word stems and combinations of words in proximity (eg, the word "pulmonary" near the word "embolism"), had an average accuracy of 93.3% across the findings (range, 81%–100%) (47). The technique was combined with an NLP algorithm to detect documentation of communication and was used to analyze 9.3 million radiology reports in a single institution, revealing that documentation of communication rose from 19% to more than 72% from 1990 to 2011, starting in 1997. Further analysis revealed that one of nine radiologists had a significantly lower rate of documenting communication for three selected findings (48).

The simpler task of automatic report indexing using NLP has significant applications in education (described in the following section) but can also be powerful for limited tasks. Many indexing tools use simple pattern matching to identify reports containing expanded versions of the query terms provided by a user, in conjunction with use of NegEx to exclude negative reports. Query expansion is readily performed using synonyms from lexicons such as RadLex to (for example) automatically search for "renal cyst" in addition to "kidney cyst" when the latter query is entered. One such tool, known as Information from Searching Content with an Ontology-Utilizing Toolkit (iSCOUT), was used to identify abdominal CT reports with a finding of a renal mass over a 1-year period at a teaching hospital to assess radiologists' adherence to management guidelines and institutional communication policies for this finding (49). The estimated positive predictive value from a subset of the reports identified by iSCOUT was 93.6% (49). Analysis of 97 reports (all 57 reports containing a critical finding and 40 reports randomly selected from those with a noncritical finding) revealed lower adherence to recommendation guidelines (73%) and communication policies (84.2%) for critical than for noncritical results (100% and 100%, respectively) (49).

Education

Enhancing access to training and imaging finding repositories is an ideal application of indexing radiology reports using NLP. Early on, Hersh et al (50) described the modification of a pattern matching NLP system to index radiology images using UMLS Metathesaurus concepts. As is often the case with pattern matching systems, excluding matches to uncommon semantic types for radiology (eg, chemical substances) and adding negation detection increased the positive predictive value significantly, but only from 14% to 30%, highlighting the difficulty of this important task (50). Do et al (11) described a similarly simple NLP application using word stemming–based pattern matching and negation detection to search reports for RadLex terms and to subsequently retrieve the images corresponding to a report from a picture archiving and communication system on demand. Rather than attempting to achieve highly accurate matches, for educational purposes, reports matching a user query were ranked in order of relevance to the search term, much like the ranking given to Web pages by Google. More recently, Dang et al (51) developed a similar system called Render, which integrates the more powerful LEXIMER NLP engine to achieve more relevant results for a given search.

Conclusion

Automated extraction of key information from free-text radiology reports with NLP has been used to enable large-scale testing of CDS, quality assurance and performance monitoring, and appropriate use of imaging, as well as to facilitate patient eligibility screening for clinical trials and hypothesis testing. Trained personnel can perform the requisite information extraction task at the cost of significant amounts of time and resources; thus, use of NLP in conjunction with statistical and machine learning classification algorithms is an attractive alternative.

Certainly, radiology reports introduce unique challenges for NLP. For example, if ambiguous terms such as “suggestive of” are mentioned and are accepted as favoring the diagnosis of a finding, an automated system’s balance of sensitivity and specificity may be altered with a bias, whereas an expert may be able to consistently infer disposition from context (43). Ambiguity of abbreviations is another example. The drive toward structured reporting in radiology is poised to enhance NLP accuracy and thus presents exciting opportunities in terms of what can potentially be achieved. In breast imaging, for example, extraction of the Breast Imaging Reporting and Data System final assessment category for positive versus negative classification by NLP using a simple pattern matching technique achieved a sensitivity of 100% and a positive predictive value of 96.6% at one center (52), arguably better results than those achieved with many of the applications described earlier.

Nonetheless, numerous NLP systems have already been used for radiology report classification, with an accuracy often similar to that of humans. NLP stands to simplify large-scale hypothesis testing of millions of existing imaging reports in a matter of minutes, an achievement that would otherwise not be possible. However, the field is still nascent, so that the choice of technologies and how they are selected and cascaded depends largely on the tools available at each institution and the expertise and knowledge of the developer of a system, rather than on the underlying requirements of the application itself. It remains important to explore the effects of these choices because in some cases similar tasks may be achieved with similar accuracies but with use of entirely different techniques. For example, simple pattern matching had an accuracy of 97% in detecting recommendations for additional imaging for incidental findings in 1635 emergency department radiology reports (46), whereas an arguably complex system achieved an accuracy of 99.6% in detecting recommendations for subsequent action in 1059 consecutive reports across all imaging modalities in an academic hospital radiology department (25). The Table lists and describes a number of tools that have been discussed in this article. Increased awareness of the applications of NLP in radiology may help drive future research to establish optimal approaches for specific applications and to achieve better accuracy for increasingly granular tasks, such as determining anatomic relationships and temporal changes.

References

1. National Guideline Clearinghouse. ACR practice guideline for communication of diagnostic imaging findings. Rockville, Md: Agency for Healthcare Research and Quality (AHRQ). http://www.guideline.gov/content.aspx?id=32541. Published 2014. Accessed March 16, 2015.
2. Jha AK, DesRoches CM, Campbell EG, et al. Use of electronic health records in U.S. hospitals. N Engl J Med 2009;360(16):1628–1638.
3. Tang PC. Key capabilities of an electronic health record system. Washington, DC: Committee on Data Standards for Patient Safety, Board on Health Care Services, Institute of Medicine, 2003.
4. Travis AR, Sevenster M, Ganesh R, Peters JF, Chang PJ. Preferences for structured reporting of measurement data: an institutional survey of medical oncologists, oncology registrars, and radiologists. Acad Radiol 2014;21(6):785–796.
5. Larson DB, Towbin AJ, Pryor RM, Donnelly LF. Improving consistency in radiology reporting through the use of department-wide standardized structured reporting. Radiology 2013;267(1):240–250.
6. Liddy E. Natural language processing. In: Encyclopedia of library and information science. 2nd ed. New York, NY: Decker, 2001.
7. National Library of Medicine (U.S.). UMLS reference manual [Internet], Chapter 1, Introduction to the UMLS. http://www.ncbi.nlm.nih.gov/books/NBK9675/. Published September 2009. Accessed March 20, 2015.
8. Langlotz CP. RadLex: a new method for indexing online educational materials. RadioGraphics 2006;26(6):1595–1597.
9. Cheng LT, Zheng J, Savova GK, Erickson BJ. Discerning tumor status from unstructured MRI reports: completeness of information in existing reports and utility of automated natural language processing. J Digit Imaging 2010;23(2):119–132.
10. Dang PA, Kalra MK, Blake MA, et al. Natural language processing using online analytic processing for assessing recommendations in radiology reports. J Am Coll Radiol 2008;5(3):197–204.
11. Do BH, Wu A, Biswal S, Kamaya A, Rubin DL. RADTF: a semantic search–enabled, natural language processor–generated radiology teaching file. RadioGraphics 2010;30(7):2039–2048.
12. Friedman C, Liu H, Shagina L, Johnson S, Hripcsak G. Evaluating the UMLS as a source of lexical knowledge for medical language processing. Proc AMIA Symp 2001:189–193.

Resources for NLP in Radiology

Resource	Description
RadLex	Comprehensive lexicon for indexing and retrieval of radiology information resources that replaces the ACR Index for Radiological Diagnoses (available at *http://www.radlex.org*)
SNOMED-CT	Comprehensive multilingual clinical terminology of clinical findings, symptoms, diagnoses, procedures, body structures, organisms and other causes, substances, pharmaceuticals, devices, and specimens; concepts are organized in a hierarchy, and relationships that link concepts are also included (available at *http://www.ihtsdo.org/snomed-ct*)
UMLS	Compilation of tools and resources including the Metathesaurus, a large comprehensive ontology of biomedical and health care–related concepts and relationships combining various sources such as SNOMED-CT; developed by the National Library of Medicine; facilitates implementation of NLP in biomedical data analysis and informatics research (available at *https://www.nlm.nih.gov/research/umls/*)
MedLEE	NLP system developed at New York Presbyterian Hospital for extracting and encoding clinical information in medical narratives including radiology reports, discharge summaries, and pathology reports; a commercial system based on this technology is available from Health Fidelity, Palo Alto, Calif (*http://healthfidelity.com*)
cTAKES	Open-source NLP system that processes clinical notes to identify types of named entities (drugs, diseases and disorders, signs and symptoms, anatomic sites, and procedures); source code can be modified to develop customized tools (available at *http://ctakes.apache.org*)
Apache OpenNLP	Machine learning–based tool kit that supports common NLP tasks (available at *http://opennlp.apache.org/*)
MetaMap	Program for mapping biomedical text to UMLS Metathesaurus concepts; also used for semi- and fully automatic indexing of biomedical literature at the National Library of Medicine (available at *http://metamap.nlm.nih.gov*)
LEXIMER	Machine learning NLP system developed at Massachusetts General Hospital; trained to identify and classify information in radiology reports, including findings and recommendations; has been licensed to Nuance (Burlington, Mass) and included in the RadCube for Radiology tool (available at *http://www.nuance.com/for-healthcare/index.htm*)
NegEx	Pattern matching–based NLP tool used to detect the negation status of an indexed phrase in a sentence (available at *https://code.google.com/p/negex/*)
iSCOUT	NLP toolkit that can search for specific terms in a file containing concatenated radiology reports, using a lexicon such as RadLex to expand the user search term to synonyms and other related terms (available at *http://sourceforge.net/projects/iscout/*)
WEKA	Open-source environment for data mining using machine learning algorithms; can be used as a stand-alone program or called from Java language programs (available at *http://www.cs.waikato.ac.nz/ml/index.html*)
MALLET	Package for statistical NLP, document classification, clustering, topic modeling, information extraction, and other machine learning applications for general texts (available at *http://mallet.cs.umass.edu/*)
eHOST	Open-source tool for manual annotation of clinical texts; supports encoding standard clinical vocabularies such as SNOMED-CT (available at *http://code.google.com/p/ehost*)
Stanford University NLP	Comprehensive list of tools and resources for statistical NLP; includes the Stanford Log-linear Part-of-Speech Tagger (available at *http://nlp.stanford.edu/links/statnlp.html*)
SMILE Text Analyzer/Stemmer	Simple online part-of-speech tagger and word stemming–based algorithms (available at *https://smile-pos.appspot.com/* and *http://smile-stemmer.appspot.com*)
Porter stemming algorithm	Widely used algorithm for English word stemming (available at *http://tartarus.org/~martin/PorterStemmer/index.html*)
Protégé	Freely distributed open-source vocabulary management tool with tools for viewing and editing annotations (available at *http://protege.stanford.edu/*)
BRAT	Freely available open-source tool for collaborative structured annotation of text (available at *http://brat.nlplab.org/*)
MEDINA	Application for de-identifying French EMRs (available at *https://medina.limsi.fr/index-en.html*)

Note.—ACR = American College of Radiology, eHOST = Extensible Human Oracle Suite of Tools, MALLET = MAchine Learning for LanguagE Toolkit, MEDINA = Medical Information Anonymization, WEKA = Waikato Environment for Knowledge Analysis.

13. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform 2001;34(5):301–310.
14. Bender EM. Linguistic fundamentals for natural language processing: 100 essentials from morphology and syntax. San Rafael, Calif: Morgan & Claypool, 2013.
15. Yu S, Kumamaru KK, George E, et al. Classification of CT pulmonary angiography reports by presence, chronicity, and location of pulmonary embolism with natural language processing. J Biomed Inform 2014;52:386–393.
16. Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. J Am Med Inform Assoc 1994;1(2):161–174.
17. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc 2004;11(5):392–402.
18. Mendonça EA, Haas J, Shagina L, Larson E, Friedman C. Extracting information on pneumonia in infants using natural language processing of radiology reports. J Biomed Inform 2005;38(4):314–321.
19. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010;17(5):507–513.
20. Garla V, Lo Re V 3rd, Dorey-Stein Z, et al. The Yale cTAKES extensions for document classification: architecture and application. J Am Med Inform Assoc 2011;18(5):614–620.
21. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp 2001:17–21.
22. Uzuner O, Zhang X, Sibanda T. Machine learning and rule-based approaches to assertion classification. J Am Med Inform Assoc 2009;16(1):109–115.
23. Elkin PL, Froehling D, Wahner-Roedler D, et al. NLP-based identification of pneumonia cases from free-text radiological reports. AMIA Annu Symp Proc 2008:172–176.
24. Yetisgen-Yildiz M, Gunn ML, Xia F, Payne TH. A text processing pipeline to extract recommendations from radiology reports. J Biomed Inform 2013;46(2):354–362.
25. Dreyer KJ, Kalra MK, Maher MM, et al. Application of recently developed computer algorithm for automatic classification of unstructured radiology reports: validation study. Radiology 2005;234(2):323–329.
26. Yadav K, Sarioglu E, Smith M, Choi HA. Automated outcome classification of emergency department computed tomography imaging reports. Acad Emerg Med 2013;20(8):848–854.
27. Martinez D, Ananda-Rajah MR, Suominen H, Slavin MA, Thursky KA, Cavedon L. Automatic detection of patients with invasive fungal disease from free-text computed tomography (CT) scans. J Biomed Inform 2015;53:251–260.
28. Pham AD, Névéol A, Lavergne T, et al. Natural language processing of radiology reports for the detection of thromboembolic diseases and clinically relevant incidental findings. BMC Bioinformatics 2014;15(1):266.
29. Solti I, Cooke CR, Xia F, Wurfel MM. Automated classification of radiology reports for acute lung injury: comparison of keyword and machine learning based natural language processing approaches. Proceedings (IEEE Int Conf Bioinformatics Biomed) 2009;2009:314–319.
30. Knirsch CA, Jain NL, Pablos-Mendez A, Friedman C, Hripcsak G. Respiratory isolation of tuberculosis patients using clinical guidelines and an automated clinical decision support system. Infect Control Hosp Epidemiol 1998;19(2):94–100.
31. 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. Radiology 2002;224(1):157–163.
32. Min JK, Kang N, Shaw LJ, et al. Costs and clinical outcomes after coronary multidetector CT angiography in patients without known coronary artery disease: comparison to myocardial perfusion SPECT. Radiology 2008;249(1):62–70.
33. Shreibati JB, Baker LC, Hlatky MA. Association of coronary CT angiography or stress testing with subsequent utilization and spending among Medicare beneficiaries. JAMA 2011;306(19):2128–2136.
34. Dublin S, Baldwin E, Walker RL, et al. Natural language processing to identify pneumonia from radiology reports. Pharmacoepidemiol Drug Saf 2013;22(8):834–841.
35. Solti I, Cooke C, Xia F, Wurfel M. Peeling away the black box label: clinical validation of a MaxEnt machine learning character *n*-gram feature set for acute lung injury. Presented at the AMIA Summit on Translational Bioinformatics, San Francisco, Calif, March 11, 2010.
36. Friedman C, Knirsch C, Shagina L, Hripcsak G. Automating a severity score guideline for community-acquired pneumonia employing medical language processing of discharge summaries. Proc AMIA Symp 1999:256–260.
37. Fiszman M, Chapman WW, Aronsky D, Evans RS, Haug PJ. Automatic detection of acute bacterial pneumonia from chest X-ray reports. J Am Med Inform Assoc 2000;7(6):593–604.
38. 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 Inform 2011;44(5):728–737.
39. Gupta A, Raja AS, Ip IK, Khorasani R. Assessing 2 D-dimer age-adjustment strategies to optimize computed tomographic use in ED evaluation of pulmonary embolism. Am J Emerg Med 2014;32(12):1499–1502.
40. Dunne RM, Ip IK, Abbett S, et al. Effect of evidence-based clinical decision support on the use and yield of CT pulmonary angiographic imaging in hospitalized patients. Radiology 2015;276(1):167–174.
41. Jain NL, Friedman C. Identification of findings suspicious for breast cancer based on natural language processing of mammogram reports. Proc AMIA Annu Fall Symp 1997:829–833.
42. Mamlin BW, Heinze DT, McDonald CJ. Automated extraction and normalization of findings from cancer-related free-text radiology reports. AMIA Annu Symp Proc 2003:420–424.
43. Carrell DS, Halgrim S, Tran DT, et al. Using natural language processing to improve efficiency of manual chart abstraction in research: the case of breast cancer recurrence. Am J Epidemiol 2014;179(6):749–758.
44. Sistrom CL, Dreyer KJ, Dang PP, et al. Recommendations for additional imaging in radiology reports: multifactorial analysis of 5.9 million examinations. Radiology 2009;253(2):453–461.
45. Dang PA, Kalra MK, Blake MA, Schultz TJ, Halpern EF, Dreyer KJ. Extraction of recommendation features in radiology with natural language processing: exploratory study. AJR Am J Roentgenol 2008;191(2):313–320.
46. Dutta S, Long WJ, Brown DF, Reisner AT. Automated detection using natural language processing of radiologists' recommendations for additional imaging of incidental findings. Ann Emerg Med 2013;62(2):162–169.
47. Lakhani P, Kim W, Langlotz CP. Automated detection of critical results in radiology reports. J Digit Imaging 2012;25(1):30–36.
48. Lakhani P, Kim W, Langlotz CP. Automated extraction of critical test values and communications from unstructured radiology reports: an analysis of 9.3 million reports from 1990 to 2011. Radiology 2012;265(3):809–818.
49. Maehara CK, Silverman SG, Lacson R, Khorasani R. Renal masses detected at abdominal CT: radiologists' adherence to guidelines regarding management recommendations and communication of critical results. AJR Am J Roentgenol 2014;203(4):828–834.
50. Hersh W, Mailhot M, Arnott-Smith C, Lowe H. Selective automated indexing of findings and diagnoses in radiology reports. J Biomed Inform 2001;34(4):262–273.
51. Dang PA, Kalra MK, Schultz TJ, Graham SA, Dreyer KJ. Render: an online searchable radiology study repository. RadioGraphics 2009;29(5):1233–1246.
52. Sippo DA, Warden GI, Andriole KP, et al. Automated extraction of BI-RADS final assessment categories from radiology reports with natural language processing. J Digit Imaging 2013;26(5):989–994.