17
CHAPTER

Natural Language Processing for Biosurveillance

Wendy W. Chapman
Center for Biomedical Informatics
University of Pittsburgh, Pittsburg, Pennsylvania

1. INTRODUCTION

Data useful for biosurveillance are often only available in a free-text format that can be easily read and understood by a human but not by a computer. Natural language processing (NLP) refers to automated methods for converting free-text data into computer-understandable format (Allen, 1995). This conversion is necessary so that information stored in free-text format can contribute to detection and characterization of outbreaks.

2. THE ROLE OF NLP IN BIOSURVEILLANCE

Detection algorithms count the number of occurrences of a variable in a given spatial location over a given time period to look for anomalous patterns. Detection algorithms require structured data, that is, data in a format that can be interpreted by a computer. By far the most common structured data formats are relational database tables. An example of a structured data element is the number of units of cold and cough medicine sold over the last 24 hours in a particular county. Many other examples of structured data elements were described in earlier chapters and in Part IV.

Much data that could potentially be useful in biosurveillance are unstructured. These include symptoms reported by a patient when presenting at an emergency facility, physical and radiological findings recorded by a physician, and queries to healthcare related web sites. In order to use these data for biosurveillance, the information must be converted. We focus our discussion on the use of NLP to encode information from textual patient records for input to outbreak detection algorithms.

3. EXAMPLE USE OF NLP

As an example, assume we want to develop an expert system that generates the probability a patient presenting to an emergency department (ED) has severe acute respiratory syndrome (SARS) given their free-text medical records. According to the World Health Organization and Centers for Disease Control and Prevention case definitions of SARS, the required input variables for diagnosing SARS are whether the patient (1) has an acute respiratory finding (Respiratory Fx), (2) is febrile (Fever), (3) has an abnormal chest radiograph consistent with consolidation or pneumonia (CXR), and (4) has recently traveled to a country currently affected by SARS (Travel). Values for the first three variables are often described in electronic textual patient records. Figure 17.1 shows excerpts from a patient's medical record generated during an ED visit, including the triage chief complaint, history and physical exam, and chest radiograph report. A human physician reading these textual records could easily determine the correct values for the first three variables. Information from the fourth variable (Travel) may not be accessible anywhere in the patient's medical record unless the dictating physician had been concerned about SARS and had dictated the travel history. Retrieving the value for the Travel variable may require a semi-automatic technique in which patients with a high probability of SARS based on the three clinical values could be interviewed regarding their travel history.

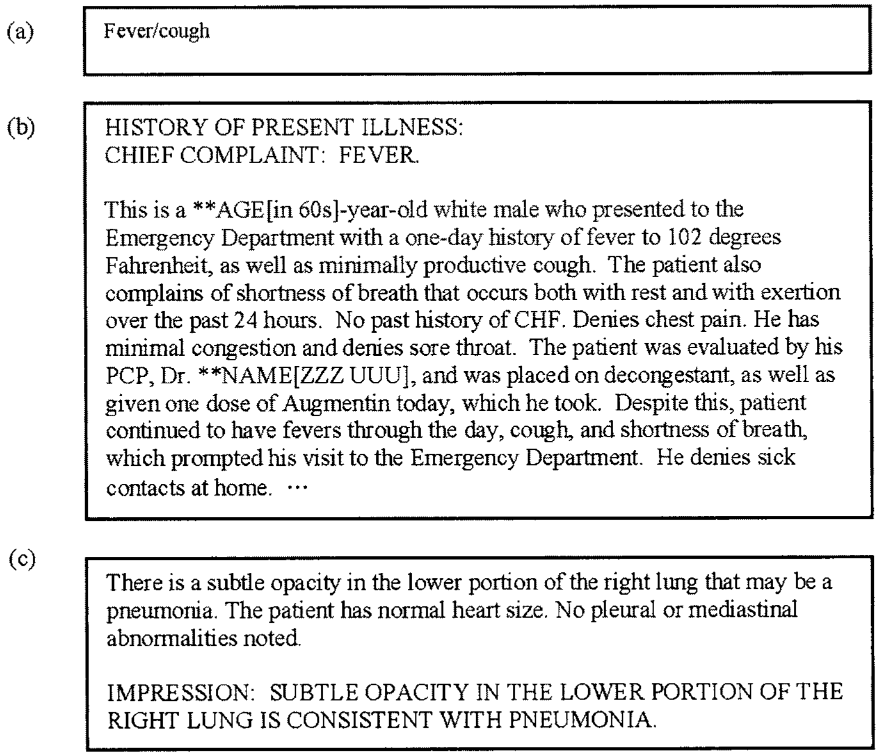
A simple expert system may only use the variable Respiratory Fx, monitoring the number of patients presenting to the ED with respiratory complaints. A physician may be able to determine the true value of the Respiratory Fx variable from information in the triage chief complaint. A more complex expert system may use all three variables. According to the medical record in Figure 17.1, a physician would assign values to the variables in a more complex expert system as follows. Respiratory Fx: yes, because the patient had a chief complaint of cough, he complained to the ED physician of productive cough and shortness of breath, and the shortness of breath was probably not cardiac in nature, given that the patient did not have a history of CHF and denied chest pain; Fever: yes, because the chief complaint and the ED report said the patient was febrile; and CXR: yes, because the radiograph report described an opacity consistent with pneumonia.

It would be impractical to hire physicians to read medical records and extract values for the variables required by our expert system. Therefore, if we want to know the variables' values, we must determine them automatically using natural language processing.

4. HOW HARD IS NLP?

NLP can be relatively easy or difficult depending on how complex the text is and on what variables you want to extract. For example, it is relatively easy to extract symptoms from free-text chief complaints using simple methods, because chief complaints are short phrases describing why the patient came

ISBN 0-12-369378-0



(a)

Fever/cough

(b)

HISTORY OF PRESENT ILLNESS:
CHIEF COMPLAINT: FEVER.

This is a **AGE[in 60s]-year-old white male who presented to the Emergency Department with a one-day history of fever to 102 degrees Fahrenheit, as well as minimally productive cough. The patient also complains of shortness of breath that occurs both with rest and with exertion over the past 24 hours. No past history of CHF. Denies chest pain. He has minimal congestion and denies sore throat. The patient was evaluated by his PCP, Dr. **NAME[ZZZ UUU], and was placed on decongestant, as well as given one dose of Augmentin today, which he took. Despite this, patient continued to have fevers through the day, cough, and shortness of breath, which prompted his visit to the Emergency Department. He denies sick contacts at home. ···

(c)

There is a subtle opacity in the lower portion of the right lung that may be a pneumonia. The patient has normal heart size. No pleural or mediastinal abnormalities noted.

IMPRESSION: SUBTLE OPACITY IN THE LOWER PORTION OF THE RIGHT LUNG IS CONSISTENT WITH PNEUMONIA.

FIGURE 17.1 De-identified excerpts from a patient's electronic medical record. (a) Chief complaint recorded by a triage nurse. (b) First paragraph of a history and physical exam dictated by the ED physician. (c) Impression section of a transcribed chest radiograph report.

to the ED. It is not possible to extract diagnoses from chief complaints, because information in a chief complaint is recorded before the patient even sees a physician. Once a patient is examined by a physician, the patient's diagnosis may be recorded in a dictated report. Extracting information from dictated reports is much more difficult, because a report tells a complex story about the patient involving references to time and negation of symptoms that are not present in chief complaints.

There are many types of technologies used in NLP. In general, the selection of technology depends on the linguistic characteristics of the text. There are some linguistic characteristics that are so difficult to process that effective NLP methods do not exist for them. For example, few NLP systems can accurately extract information that is being conveyed by use of a metaphor. Fortunately, metaphor is not a frequent characteristic in the data sources of potential value in biosurveillance.

In the remainder of this chapter we will discuss (1) the linguistic characteristics of clinical texts that should be considered when implementing NLP for biosurveillance, (2) the types of NLP technologies researchers are using to successfully model information in text, (3) evaluation methods for determining how successful an NLP application is in the domain of outbreak and disease surveillance, and (4) the feasibility of using NLP to encode information for biosurveillance expert systems.

5. LINGUISTIC CHARACTERISTICS OF CLINICAL TEXT—WHAT MAKES NLP HARD?

According to Zelig Harris (Friedman et al., 2002), the informational content and structure of a domain form a specialized language called a sublanguage. The sublanguage of patient medical records exhibits linguistic characteristics that influence an NLP system's ability to extract information from the text. When a physician reads a patient's medical reports, she understands the linguistic characteristics of the text and can make reasonable inferences from the record. For instance, a physician will not assign the Respiratory Fx a value of yes if the respiratory finding described in the report is described as occurring in the patient's past history. For an NLP application to determine the values of clinical variables from patient records the same way a physician would, the application must account for or model the linguistic characteristics of the clinical text. Some important linguistic characteristics of the sublanguage of patient reports are (1) linguistic variation, (2) polysemy, (3) negation, (4) contextual information, (5) finding validation, (6) implication, and (7) co-reference.

5.1. Linguistic Variation

Natural language provides us with freedom to express the same ideas in different ways. Humans are generally capable of understanding the meaning of a natural language expression in spite of such variation; however, the freedom that accompanies natural language makes computerized understanding of the language difficult. In patient reports, a patient's clinical state can be expressed differently due to the linguistic characteristics of derivation, inflection, and synonymy.

Derivation and inflection change the form of a word (the word's morphology) while retaining the underlying meaning of the word. The adjective "mediastinal" can be derived from the noun "mediastinum" by exchanging the suffix -um for the suffix -al. Similar rules can be used to derive the adjective "laryngeal" from the noun "larynx" or to derive the noun "transportation" from the verb "transport."

There are other forms of linguistic variation to contend with. The two most important are inflectional rules (which change a word's form, such as by pluralization of a noun or tense change of a verb) and synonymy (in which different words or phrases mean the same thing).

Physicians reading reports are seldom confused by derivation, inflection, or synonymous expressions. An NLP application attempting to determine whether a patient has shortness of breath, for example, must account for linguistic variation in order to identify "dyspnea," "short of breath," or "dyspneic" as evidence of shortness of breath.

5.2. Biomedical Polysemy

Terms that have the identical linguistic form but different meanings are polysemous. Biomedical polysemy manifests itself in different ways (Roth and Hole, 2000, Liu et al., 2001). Some words in clinical texts have different biomedical meanings or word senses. For instance, the word "discharge" has two word senses—one word sense meaning a procedure for being released from the hospital, as in "prior to discharge," and one word sense meaning a substance that is emitted from the body, as in "purulent discharge."

Acronyms and abbreviations with more than one meaning may be the most frequently occurring type of biomedical polysemy. A striking example of this is the acronym "APC," which has more than thirty unique biomedical definitions, including activated protein c, adenomatosis polyposis coli, antigen-presenting cell, aerobic plate count, advanced pancreatic cancer, age period cohort, and alfalfa protein concentrated. According to one study (Wren and Garner, 2002), 36% of the acronyms in MEDLINE are associated with more than one definition. The number of unique acronyms in MEDLINE is increasing at the rate of 11,000 per year, and the number of definitions associated with unique acronyms is increasing at 44,000 per year. In the sublanguage of patient reports, the type of report is helpful in disambiguating the correct meaning of an acronym or abbreviation, because the report type indicates the type of medical specialty. In this way, "APC" in a microbiology lab report is more likely to mean aerobic plate count, whereas "APC" in a discharge summary may be referring to advanced pancreatic cancer.

Triage chief complaints are full of abbreviations created by clerks and triage nurses to keep the complaint short (Travers and Haas, 2003). Some of the abbreviations are standard and are easily understood by physicians, such as "rt" for "right" and "h/a" for headache. But many abbreviations in chief complaints are unique to the sublanguage of chief complaints or perhaps even to a single hospital or registration clerk. For example, "appy" is commonly used to describe an "appendectomy," and in one hospital "gx" indicates the patient came to the ED by ground transportation.

Depending on the particular clinical variables that we want to extract or encode from text, understanding the meaning or word sense of polysemous words in the patient reports can be critical to success.

5.3. Negation

One of the primary goals in differential diagnosis is to definitively rule out as many hypotheses as possible in order to concentrate on the most probable set of diagnoses. One study (Chapman et al., 2001a) estimated that between 40% and 80% of all findings were explicitly negated in ten different report types, with surgical pathology and operative notes demonstrating the least amount of negation and mammograms and chest radiograph reports demonstrating the most. Explicit negations are indicated by negation terms such as "no," "without," and "denies." Findings can also be implicitly negated. For example, "The lungs are clear upon auscultation" indicates that rales/crackles, rhonchi, and wheezing are all absent. We focus on explicit negation, which is the most common type of negation in patient reports.

In most cases, a physician can easily determine from a report whether a finding is negated in the text. In the sentence, "The patient denies chest pain but has experienced shortness of breath," a physician would assign the clinical variable chest pain the value of no and the variable shortness of breath the value of yes. The types of information a human uses to identify explicitly negated findings include (1) negation terms, (2) scope of the negation term, and (3) expressions of uncertainty.

5.3.1. Negation Terms

Explicit negations are triggered by negation terms that may precede the finding being negated, as in "The chest x-ray revealed no abnormalities," or may follow the observation, as in "The patient is tumor free." Consistent with Zipf's law (Manning and Schutze, 1999), which states that there exist a few very common words, a middling number of medium-frequency words, and many low-frequency words, very few negation phrases account for the majority of negation in patient reports. Two studies on automated negation (Mutalik et al., 2001, Chapman et al., 2001a) found that a few negation phrases accounted for approximately 90% of negation in different report types: "no," "denies/denied," "without," and "not."

The other 10% of negated observations are triggered by a potentially huge number of low-frequency negation phrases.

Once a human identifies a negation term, he must decide whether a relevant finding in the sentence is being negated by that term, that is, whether the finding is within the scope of the negation term. For example, in sentences (1) and (2) the words "source" and "change" are being negated by "not" instead of the findings "infection" and "pain."

(1) This is not the source of the *infection*.
(2) There has not been much change in her *pain*.

5.3.2. *Expressions of Uncertainty*

Unfortunately, differential diagnosis is not a clear-cut science in which physicians are completely confident in what findings or diseases a patient has, and the language used in patient reports expresses the dictating physician's uncertainty on a continuum ranging from certain absence to certain presence. Consider the implications of sentences (5) to (12). The first sentence expresses certainty that pneumonia is absent, whereas the last sentence expresses certainty that pneumonia is present. The intervening sentences express different amounts of uncertainty about a diagnosis of pneumonia. A sophisticated expert system may try to incorporate uncertainty of the variables into its decision making. A simpler expert system may only allow variables to be present or absent. In that case, determining whether pneumonia is negated in the sentences below depends on the goal of the expert system. An expert system designed to be especially sensitive may accept a finding with uncertainty to be present and may set the value of pneumonia to yes for all but the first two sentences. An expert system designed to be specific may consider uncertainty about the variable an indication of negation and may only set the value of pneumonia to yes for the last two.

(5) The chest x-ray ruled out *pneumonia*.
(6) We performed a chest x-ray to rule out *pneumonia*.
(7) Cannot rule out *pneumonia*.
(8) It is not clear whether the opacity is atelectasis or *pneumonia*.
(9) Radiographic findings may be consistent with *pneumonia*.
(10) Discharge diagnosis: possible *pneumonia*.
(11) The patient has *pneumonia*.
(12) He did have sputum that grew out klebsiella *pneumonia* during his admission.

5.4. Contextual Information

Information contained in a single word or phrase is not always sufficient for understanding the value of a clinical variable; the context around the phrase is often essential in understanding the patient's clinical state. Among other things, contextual information is important for determining when the finding occurred and what anatomic location was involved.

Any expert system attempting to increase timeliness in outbreak detection must distinguish between findings that occurred in past history and current problems. For example, one of the variables in our SARS detector is whether the patient has an acute respiratory finding. The definition of acute is not straightforward. However, at the least, an NLP application attempting to determine the value of this variable should be able to accurately assign the value yes to pleuritic chest pain in sentence 13 and no to pneumonia in sentence 14.

(13) The patient presents today with *pleuritic chest pain*.
(14) She has a past history significant for *pneumonia*.

A physician reading a report uses contextual clues like the structure of a report to discriminate between acute or current findings and those in the past history. For example, a finding described in an ED report within a section that is titled "Past Medical History" is probably a historical finding. A human may also use linguistic cues within sentences to determine whether a finding is current. For instance, in sentence 15, a physician would know that myocardial infarction occurred in the past history but that chest pain is a current finding.

(15) He has a past history significant for *myocardial infarction*, and presents to the ED today with *chest pain*.

Determining what findings are described in a patient report also entails discriminating current findings from future or hypothetical findings. In sentence 16, the instance of fever is described as a hypothetical finding, but shortness of breath is described as a finding that probably occurred at the current hospital visit.

(16) She should return for *fever* or exacerbation of her *shortness of breath*.

Some findings can occur with multiple anatomic locations. For detection of SARS, our expert system needs to know whether the edema described in sentence 17 was found in the lung or in the skin.

(17) Chest is *edematous*.

Sometimes the anatomic location is explicitly stated, as in sentence 18. Other times, the anatomic location is not explicitly stated (e.g., sentence 19). The context around the finding is important for disambiguating the anatomic location—even when a location is not explicitly stated.

(18) The *lump* on her back has not changed.
(19) Chest x-ray showed no *mass*.

5.5. Finding Validation

Not all terms representing findings or diseases in a patient report are actual findings in the patient; some findings must have a particular value in order to be considered positive. The variable of oxygen desaturation may be useful in our SARS detector, but a physician may not describe oxygen desaturation with those words. Instead he may say "the patient's O_2 saturation is low" or "the patient is satting at 85% on room air." The qualitative value of "low" in the first example and the quantitative value of "85%" are what let the reader know the patient has oxygen desaturation. Similarly, the presence of the word "temperature" does not inform the reader of whether the patient has a fever–the variable together with its value provide the requisite information to the reader.

5.6. Implication

The main audience of patient reports consists of other physicians. For this reason, understanding what is said in a dictated medical report is difficult for a human reader without domain knowledge. Researchers compared laypeople against physicians at reading chest radiograph reports and judging whether the report described radiological evidence of acute bacterial pneumonia (Fiszman et al., 1999). Not surprisingly, laypeople performed much worse than physicians. As long as the report stated explicitly that the findings were consistent with pneumonia, the laypeople agreed with the physicians in their judgment, but pneumonia was mentioned in only one-third of the positive reports. In the remaining two-thirds of the reports, the evidence for pneumonia was inferred by the physicians and missed by the laypeople.

Implication in medical reports can occur at the sentence level and at the report level. A simple example is the sentence, "The patient had her influenza vaccine." If our SARS expert system had a variable for influenza, even a layperson reading the previous sentence could determine that the value for the variable would probably be no, because the patient was vaccinated. This inference requires domain knowledge that a vaccine generally prevents the target disease. In the radiology study reported above, evidence for pneumonia in positive reports was not always explicitly stated by the radiologist. Instead, the radiologist described "hazy opacities" or "ill-defined densities" in the lobes of the lung, which can be inferred to mean localized infiltrates. Once the inference at the sentence level has been correctly made, a physician reading the radiology report can integrate the findings described throughout the entire report and can infer that because the chest x-ray shows localized infiltrates not explained by other causes, there is evidence for acute bacterial pneumonia. Domain knowledge about words, combinations of words, and combinations of findings make it possible for a physician to make inferences from reports that a lay person—or an expert system—may not be able to make without training in knowledge of the domain.

5.7. Coreference

As described above, sometimes information across sentences must be combined to truly understand the patient's clinical state. A single entity (which could be a finding, a person, or some other object mentioned in a report) may be referred to in more than one sentence. True to the human inclination towards conciseness, once an entity has been evoked, we can refer to the entity with shortened phrases, including pronouns (e.g., "it," "he," or "she") or definite noun phrases (e.g., "the finding," or "her mother"). When two expressions refer to the same entity, they corefer. Determining which referring expressions refer to which referent is important in understanding a clinical report.

5.8. Summary of Linguistic Issues

We have described some of the linguistic characteristics of the sublanguage of patient medical records, including linguistic variation, polysemy, negation, contextual information, finding validation, implication, and coreference. If we want to automatically determine an individual patient's values for the variables used in our expert system, we must address these linguistic characteristics, using the types of information a physician uses to understand the meaning of the words and sentences in the reports. Below we describe some of the techniques current natural language processing research employs for extracting information from clinical texts.

6. TECHNOLOGIES FOR NATURAL LANGUAGE PROCESSING

NLP techniques fall into two broad classes: statistical and symbolic. Statistical techniques use information from the frequency distribution of words within a text to classify or extract information. Symbolic techniques use information from the structure of the language (syntax) and the domain of interest (semantics) to interpret the text to the extent necessary for encoding the text into targeted categories. Although some NLP applications exclusively use one or the other technique, many applications use both statistical and symbolic techniques. In this section, we give a brief background of NLP research in the medical domain and describe some statistical and symbolic NLP techniques used for classifying, extracting, and encoding information from biomedical texts, focusing on techniques useful for addressing the linguistic characteristics of patient medical reports described in the previous section.

6.1. Brief Background of NLP in Medicine

Over the last few decades researchers have actively applied NLP techniques to the medical domain (Friedman and Hripcsak, 1999, Spyns, 1996). NLP techniques have been used for a variety of applications, including quality assessment in radiology (Fiszman et al., 1998, Chapman et al., 2001b); identification of structures in radiology images (Sinha et al., 2001a, Sinha et al., 2001b); facilitation of structured reporting (Morioka et al., 2002, Sinha et al., 2000) and order entry (Wilcox et al., 2002, Lovis et al., 2001); encoding variables required by automated decision-support systems such as guidelines (Fiszman and Haug, 2000), diagnostic systems (Aronsky et al., 2001), and antibiotic therapy alarms (Fiszman et al., 2000); detecting patients with suspected tuberculosis (Jain et al., 1996, Knirsch et al., 1998, Hripcsak et al., 1999); identifying findings suspicious for breast cancer (Jain and Friedman, 1997), stroke (Elkins et al., 2000), and community acquired pneumonia (Friedman et al., 1999b); and deriving comorbidities from text (Chuang et al., 2002).

Probably the most widely used and evaluated NLP system in the medical domain is MedLEE, which was created at Columbia Presbyterian Medical Center (Friedman, 2000, Friedman et al., 1994, 1998, 1999a). MedLEE extracts clinical information from several types of radiology reports, discharge summaries, visit notes, electrocardiography, echocardiography, and pathology notes. MedLEE has been shown to be as accurate as physicians at extracting clinical concepts from chest radiograph reports (Hripcsak et al., 1995, 2002).

NLP has only recently been applied to the domain of outbreak and disease surveillance, and most of the research has focused on processing free-text chief complaints recorded in the ED (Olszewski, 2003, Ivanov et al., 2002, Ivanov et al., 2003, Travers et al., 2003, Travers and Haas, 2003, Chapman et al., 2005a).

Below we describe some of the statistical and symbolic NLP techniques implemented in the medical domain.

6.2. Statistical NLP Techniques

Statistical text classification techniques use the frequency distribution of words to automatically classify a set of documents or text fragments into one of a discrete set of predefined categories (Mitchell, 1997). For example, a text classification application may classify MEDLINE abstracts into one of many possible MeSH categories or may classify websites by topic. Various statistical models have been applied to the problem of text classification, including regression models, Bayesian belief networks, nearest neighbor algorithms, neural networks, decision trees, and support vector machines. The basic element in all text classification algorithms is the frequency distribution of the words in the text. Applications of text classification of free-text patient medical records include retrieving records of interest to a specific research query (Aronis et al., 1999, Cooper et al., 1998), assigning ICD-9 admission diagnoses to chief complaints (Gundersen et al., 1996), and retrieving medical images with specific abnormalities (Hersh et al., 2001). In the domain of biosurveillance, text classification techniques have been applied to triage chief complaints and chest radiograph reports. CoCo (Olszewski, 2003) is a naive Bayesian text classification application that classifies free-text triage chief complaints into syndromic categories, such as respiratory, gastrointestinal, or neurological, based on the frequency distribution of the words in the chief complaints. For example, the chief complaint "cough" would be assigned a higher probability of being respiratory than of being gastrointestinal or neurological, because chief complaints in the training corpus that contained the word "cough" were classified most frequently as respiratory. The IPS system (Aronis et al., 1999, Cooper et al., 1998) was used to create a query for retrieving chest radiograph reports describing mediastinal findings consistent with inhalational anthrax (Chapman et al., 2003). The IPS system uses likelihood ratios to identify words that discriminate between relevant and not relevant documents.

Statistical NLP techniques have been applied to the problem of biomedical polysemy. Given a word or phrase with multiple meanings, the statistical distribution of the neighboring words in the document could be helpful in disambiguating the correct meaning or sense of the word. As an example, consider the word "discharge," which has two word senses: a procedure for being released from the hospital (Disch1) and a substance emitted from the body (Disch2). If we applied a statistical learning technique to text containing the word "discharge," we may learn that Disch1 occurs significantly more often with the neighboring words "prescription," "upon," "home," "today," and "instructions," and that Disch2 occurs more often with the words "purulent," "rashes," "swelling," and "wound."

Beyond text classification, statistical techniques can be used for complex NLP tasks. For instance, Taira and Soderland (1999) have developed an NLP system for radiology reports that uses mainly statistical techniques to encode detailed information about radiology findings and diseases, including the finding, whether it was present or absent, and its anatomic location.

Because of the complexity of patient medical reports, purely statistical techniques that only rely on words and their frequencies are less common than hybrid or purely symbolic techniques that leverage knowledge about the structure or meaning of the words in the text in order to classify, extract, or encode information in clinical documents.

6.3. Symbolic NLP Techniques

Linguistics is the study of the nature and structure of language, including the pronunciation of words (phonetics), the way words are built up from smaller units (morphology), the way words are arranged together (syntax), the meaning of linguistic utterances (semantics), and the relation between language and context of use (pragmatics) including the relationships of groups of sentences (discourse). As humans, we combine all of this linguistic knowledge with knowledge of the world to understand natural language. Symbolic NLP techniques also utilize this linguistic information in attempting to interpret free-text. Below we describe NLP techniques that take advantage of syntactic, semantic, and discourse knowledge in order to address the linguistic characteristics of clinical texts.

6.3.1. Syntax: The Way Words Are Arranged Together

Every word in a language has at least one part of speech. The most common parts of speech in English are noun (e.g., "tuberculosis," "heart"), verb (e.g., "see," "prescribe"), adjective (e.g., "severe," "red"), adverb (e.g., "quickly," "carefully"), determiner (e.g., "the," "some"), preposition (e.g., "of," "in"), participle (e.g., "up," "out"), and conjunction (e.g., "and," "but"). The difficulty in automatically assigning a part of speech to words in a sentence is that some words can have more than one part of speech. For example, the word "discharge" can be a verb or a noun. Automated part-of-speech taggers use either rules or probability distributions learned from hand-tagged training sets to assign parts of speech and perform with an accuracy of 96–97% on general English texts, such as newspaper articles, scientific journals, and books. Part-of-speech distribution in patient reports is different than that of nonclinical texts. For example, discharge summaries contain more nouns and past tense verbs and fewer proper nouns (e.g., people and company names) and present tense verbs (Campbell and Johnson, 2001). Not surprisingly, training a part-of-speech tagger on medical texts improves its accuracy when assigning parts of speech to patient reports (Campbell and Johnson, 2001, Coden et al., 2005). Publicly available part-of-speech taggers trained on medical documents are just beginning to become available (Smith et al., 2004).

A word's part of speech can sometimes be helpful in understanding which word sense is being used in a sentence. Returning to the example of the word "discharge," a statistical analysis of the distribution of "discharge" in patient reports may show that if "discharge" is being used as a verb, the word sense is more likely Disch1 (release from hospital).

Syntactic rules use the part of speech to combine words into phrases and phrases into sentences. For instance, an adjective followed by a noun is a noun phrase, an auxiliary verb followed by a verb is a verb phrase, and a preposition followed by a noun phrase is a prepositional phrase. Phrases can be combined so that a noun phrase followed by a prepositional phrase creates another noun phrase and a noun phrase followed by a verb phrase creates a sentence. This process of breaking down a sentence into its constituent parts is called parsing. Automated parsers employ a grammar consisting of rules or probability distributions for generating combinations of words and a lexicon listing the possible parts of speech for the words. Automated parsers may attempt to produce a deep parse that connects all the words and phrases together into a sentence (Figure 17.2[a]) or a partial parse (also called a shallow parse), which combines words into noun phrases, verb phrases, and prepositional phrases but does not attempt to link the phrases together (Figure 17.2[b]). A deep parse gives you more information about the relationships among the phrases in the sentence but is more prone to error. A partial parse is easier to compute without errors and may be sufficient for some tasks.

As with part-of-speech tagging, the syntactic characteristics of patient reports differ from those of nonclinical texts (Campbell and Johnson, 2001). A publicly available parser trained on medical texts does not yet exist. Szolovitz (2003) showed that the Link Grammar Parser (available at *www.link.cs.cmu.edu/link/*) only recognized 38% of the words in a large sample of ED reports. For this reason, he adapted the SPECIALIST Lexicon distributed by the National Library of Medicine to the format required for the Link Grammar Parser

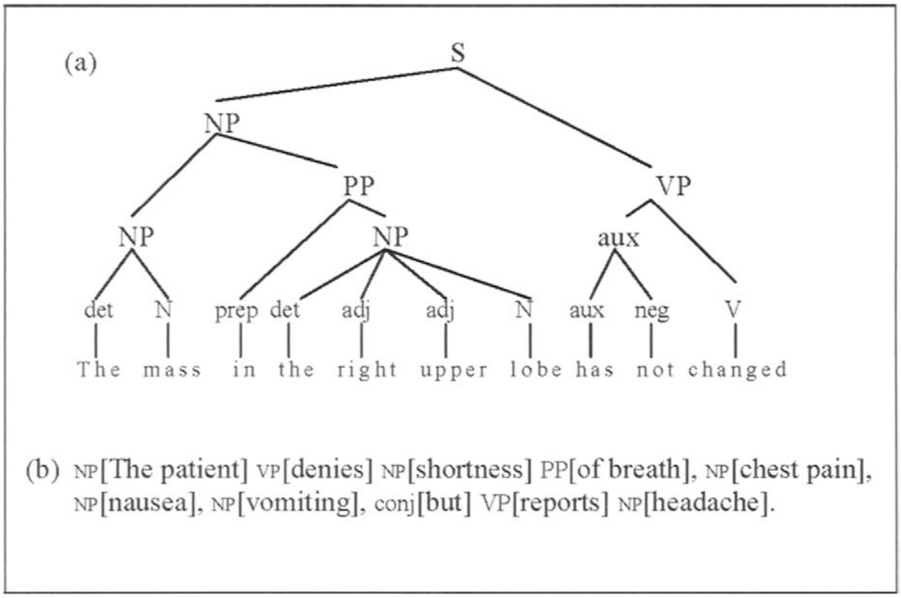

(a)
S
NP
PP
VP
NP
NP
aux
det
N
prep
det
adj
adj
N
aux
neg
V
The
mass
in
the
right
upper
lobe
has
not
changed
(b) NP[The patient] VP[denies] NP[shortness] PP[of breath], NP[chest pain], NP[nausea], NP[vomiting], conj[but] VP[reports] NP[headache].

FIGURE 17.2 (a) The tree structure of a deep parse in which words are combined into phrases and phrases are combined into a sentence. det, determiner; N, noun; prep, preposition; adj, adjective; aux, auxiliary verb; v, verb. (b) A partial parse that only labels simple phrases and conjunctions (conj) without linking the phrases together.

and provided over 200,000 new entries for the Link Grammar Lexicon, quintupling the size of the original Lexicon (*available at www.medg.lcs.mit.edu/projects/text/*).

The syntactic structure of a sentence can provide information about the semantic relationships among the words. For example, in Figure 17.2(a) a relationship between the mass and the right upper lobe is indicated by the fact that the prepositional phrase "in the right upper lobe" is attached to the noun phrase "the mass." Statistical methods that rely on whether or not a word or phrase occurs in the sentence without requiring a syntactic relation between the constituents may mistakenly infer a location relation between a noun phrase and prepositional phrase. For instance, in sentence 22, the noun "mass" and the prepositional phrase "in the right upper lobe" both occur in the sentence, but without syntactic knowledge there is no way to know the phrases are actually unrelated.

(22) There is no change in the mass, but the infiltrate in the right upper lobe has increased..

6.3.2. Semantics: The Meaning of Linguistic Utterances

Understanding the syntactic relationships among words in a sentence does not assure understanding the meaning of the sentence. Sentence (23) shows a famous example by Noam Chomsky, the father of modern linguistics, of a perfectly grammatical sentence that has no meaning.

(23) Colorless green ideas sleep furiously.

Understanding a patient report requires not only knowledge of the syntactic relation of the words but also knowledge of the meaning of the words in the report, knowledge of semantic relations between the words, and knowledge of the relationships between the words and the ideas they represent.

Lexical Semantics Refers to Meaning of Words. Understanding the meaning of the words used in a patient medical report is the first step to understanding what is wrong with the patient. The National Library of Medicine's (NLM) Unified Medical Language System® (UMLS®) has created several resources to "facilitate the development of computer systems that behave as if they 'understand' the meaning of the language of biomedicine and health" (*www.nlm.nih.gov/research/umls/about_umls.html*). The NLM freely distributes three UMLS knowledge sources: the Metathesaurus®, the Semantic Network, and the SPECIALIST Lexicon. The three knowledge sources can assist NLP applications in understanding the meaning of the words in clinical reports.

The Metathesaurus is a vocabulary database of biomedical and health related concepts containing over 900,000 concepts compiled from more than 60 different source vocabularies. The Metathesaurus integrates existing vocabularies (such as SnoMed and ICD-9), which provide terms and sometimes hierarchies relating the terms. The Metathesaurus organizes the terms into concepts, organizes the concepts into hierarchies, and relates concepts to each other. If a concept from a new source vocabulary already exists in the Metathesaurus, the concept is added as a synonym. The Metathesaurus is the most complete collection of biomedical concepts and their synonyms.

The Semantic Network provides a consistent categorization of all concepts represented in the Metathesaurus, which are the nodes in the network, and provides a useful set of relations among these concepts, which are the arcs in the network. Every concept in the Metathesaurus is assigned at least one of 135 different semantic types (e.g., finding, anatomical structure, pathologic function, etc.). The Semantic Network contains 54 relationships among the semantic types, such as "part of," "is-a," and "caused by."

An NLP application for our SARS detector may find the phrase "shortness of breath" in a patient report, which is a synonym for the Metathesaurus concept Dyspnea. Other synonyms for the concept Dyspnea are "difficulty breathing," "SOB," and "breathlessness." The concept Dyspnea has the semantic type of Sign or Symptom and has children like "hypoventilation," "paroxysmal dyspnea," and "respiratory insufficiency." A knowledge base with synonyms and semantic information can be helpful in identifying variables and their values from text.

The SPECIALIST Lexicon is a general English lexicon that includes many of the biomedical terms in the Metathesaurus together with the most commonly occurring English words. As of 2003, the SPECIALIST contained almost 300,000 entries. A lexical entry for each word or term records information about spelling variants, derivation, inflection, and syntax. Using the SPECIALIST, we could know, for example, that the term "mediastinal" is the adjectival form of the noun "mediastinum" and that the following phrases are equivalent: "mediastinal widening," "widened mediastinum," and "wide mediastinum."

The Semantic Relationships Among Words Are Also Important. An NLP technique with a syntactic model of a sentence and a semantic model of the words in a sentence has a better chance of understanding relationships among the words in the sentence. For example, a noun phrase comprising an adjective followed by a noun signifies a relationship between the noun, which is the head of the phrase, and the adjective, which is the modifier. Precisely what that relationship is depends on the meaning of the words in the phrases. Consider the phrase "atrial fibrillation." The UMLS semantic type for "atrial" is Body Part, Organ, or Organ Component, and the semantic types of "fibrillation" are Disease or Syndrome and Sign or Symptom. An NLP application that modeled both the syntactic and the semantic information in this phrase could have a rule that stated: *If a syntactic modifier has the semantic type Body Part, Organ, or Organ Component, and the head has the semantic type Disease or Syndrome or Sign or Symptom, the semantic relationship is* Head-*has-location*-Modifier.

An application that validated whether a term mentioned in a report actually is a finding could benefit from modeling semantic and syntactic relationships. For instance, the NLM system FindX (Sneiderman et al., 1996) contains rules based on the semantic type of the words in a modifier-head relation to validate the finding. For example, one rule states: *An abnormality or anatomical site modified by a SNOMED adjective is a finding*, validating "*chest* clear to auscultation" as a finding. Another rule says: *A diagnostic or laboratory procedure modified by a SNOMED adjective or a numeric value is a finding.* This rule correctly validates arterial blood gas as a finding in (24) and invalidates it in (25).

(24) *Arterial blood gas* 7.41/42/43/27
(25) We suggest *arterial blood gas* preoperatively.

Semantic modeling of syntactically related words can also be useful in understanding implicit information in a report. MPLUS (Medical Probabilistic Language Understanding System) is an NLP system that uses Bayesian networks to model the relationship between the words in a report and the ideas or concepts the words represent (Christensen et al., 2002). Figure 17.3 shows a simplified network for radiological findings. The syntactic parse helps determine which words in the sentence should be slotted together into the Bayesian network (i.e., which words are syntactically related). When a new phrase or sentence is slotted into the network, MPLUS can make inferences about the meaning of the words in a sentence in spite of the different combinations of words that can be used to describe the same concept. For example, the phrases "hazy opacity in the left lower lobe" and "ill-defined densities in the lower lobes" both indicate localized infiltrates—even though the word "infiltrate" was not used by the radiologist.

6.4. Discourse: Relationships Among Sentences

Sentences in a patient report are not meant to stand alone–they often convey a story about the differential diagnosis and treatment process for a patient. Some of the variables our example SARS expert system would need cannot be obtained without integrating and disambiguating information from the entire report. Once the individual variables have been located in a report, some type of discourse processing must integrate values for the variables to answer questions such as: (1) Were the relevant findings reported for the patient or for someone else (e.g., a family member, as in "patient's mother died at the age of 48 with an MI")? (2) Did the relevant findings occur at the current hospital visit (versus past history or hypothetical findings)? (3) Is it likely the patient has a respiratory disease or disorder? Three discourse techniques that may help answer these questions are section identification, co-reference resolution, and diagnostic modeling.

Patient reports are semistructured, depending on the type of report and the institution from which the report is generated. For instance, ED reports may contain sections for chief complaint, past history, history of present illness, physical exam, radiologic or lab findings, hospital course, discharge diagnosis, and plan. The section in which a finding is described can provide information important to understanding the meaning of the report. For example, our SARS detector may have a variable for pneumonia history, a variable for radiological

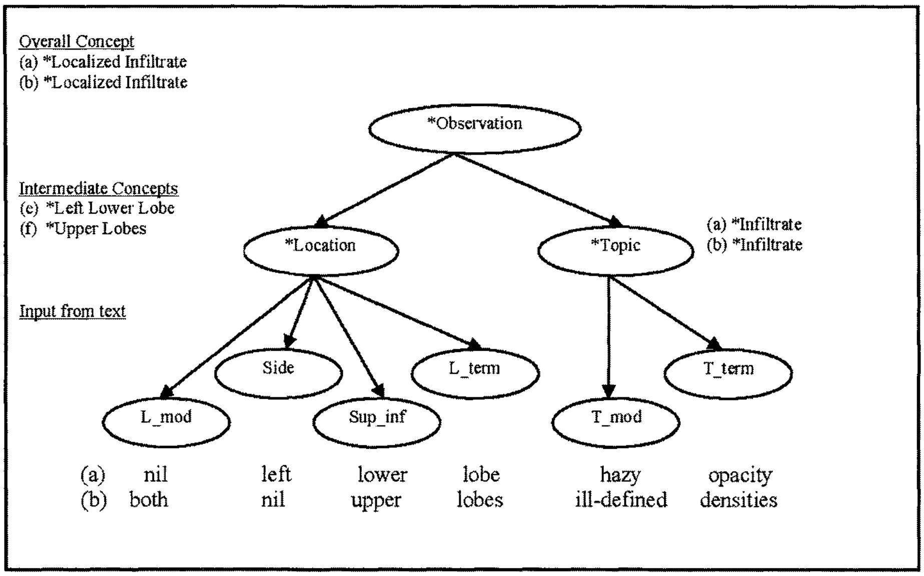


FIGURE 17.3 Partial Bayesian network for radiological findings. Words in leaf nodes come directly from text, concepts in other nodes (shown with asterisks[*]) are inferred based on training examples. Two sentences are slotted in the network: (a) There is a hazy opacity in the left lower lobe. (b) Both upper lobes show ill-defined densities.

evidence of pneumonia, and a variable for a pneumonia diagnosis. An instance of pneumonia described in the radiological findings section of a report is likely to provide radiological evidence of pneumonia, whereas an instance of pneumonia in the discharge diagnosis section probably indicates a diagnosis. Report section identification can also assist in understanding whether the finding occurred in the past history, the current visit, or as a hypothetical finding (e.g., a finding described in the plan section is more likely to be a hypothetical finding), can identify findings for family members (e.g., a finding in the social history section may not be the patient's finding), and can provide insight regarding the anatomic location of an ambiguous finding (e.g., a mass described in the radiology finding section is probably a pulmonary mass).

Patient reports tell a story involving various findings, physicians, patients, family members, medications, and treatments that are often referred to more than once in the text. Identifying which expressions are really referring to the same entity is important in integrating information about that entity. Useful discourse clues for identifying coreferring expressions include how close the expressions are within the text (e.g., a referring expression is more likely to refer to a referent in the previous sentence than to a referent five sentences back), overlapping words (e.g. "the pain" is more likely to refer to "chest pain" than to "atelectasis"), and the semantic type of the entities (e.g., "she" can only refer to a human entity, not to a finding or disease).

Integrating the clinical information within a report to determine the clinical state of the patient (e.g., the likelihood the patient has SARS) requires a diagnostic model relating the individual variables or findings to the diagnosis. Many diagnostic models have been used in medicine, including rule sets, decision trees, neural networks, and Bayesian networks. Diagnostic models are also helpful for determining the values of individual variables. For example, a Bayesian network can model which radiological findings occur with which diseases. With this type of semantic model, even if a report did not mention pneumonia, for example, the model could infer that acute bacterial pneumonia is probable given the radiologic finding of a localized infiltrate (Chapman et al., 2001c).

None of the NLP techniques we have described perform perfectly, but some of the techniques described in this section are easier to address than others. For instance, automatic part-of-speech taggers perform similarly to human taggers. The ability to perform inference on information in a report as a physician does is more complex, entailing both semantic and discourse modeling.

Although the task is difficult, developing NLP techniques for classifying, extracting, and encoding individual variables from patient medical reports is feasible and has been accomplished to different extents by many groups. Successful extraction of variables in spite of imperfect syntactic and semantic techniques can occur for many reasons, including access to the UMLS databases and tools, structure and repetition within reports, and modeling a limited domain. NLP research over the years has revealed that NLP techniques perform better in narrower domains. For instance, modeling the lexical semantics of the biomedical domain is easier than modeling the lexical semantics of all scientific domains, and modeling the lexical semantics of patient reports related to SARS would be easier than modeling all clinical findings in patient reports.

Most of the studies in NLP have focused on the ability of the technology to extract and encode individual variables from the reports. Fewer studies have integrated NLP variables from an entire report to diagnose patients or have evaluated whether an NLP-based expert system can improve patient care. Below we discuss different levels of evaluation of NLP technology related to biosurveillance.

7. EVALUATION METHODS FOR NLP IN BIOSURVEILLANCE

The first step in evaluating an NLP application is to validate its ability to classify, extract, or encode features from text (feature detection). Most evaluations of NLP technology in the biomedical domain have focused on this phase of evaluation. Once we validate feature detection performance, we can evaluate the ability of the encoded features to diagnose individual cases of interest (case detection). Finally, we can perform summative evaluations addressing the ability to detect epidemics (epidemic detection). Figure 17.4 shows how the three levels of evaluation relate to one another, using the diagnostic system for SARS as an example.

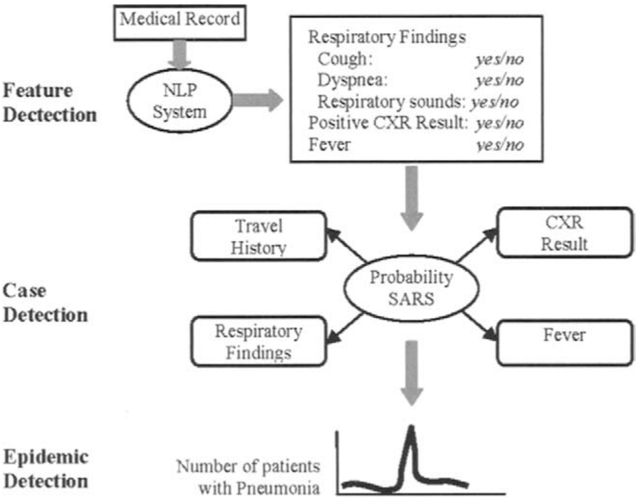


FIGURE 17.4 Relationship between the three levels of evaluation for biosurveillance. Evaluations of feature detection quantify how well variables and their values are automatically encoded from text. Evaluations of case detection quantify the ability to accurately diagnose a single patient from the variables encoded from text, which may or may not be combined with other variables. Evaluations in epidemic detection quantify whether the variable being monitored by detection algorithms can detect outbreaks.

7.1. Feature Detection

The first type of NLP evaluation should measure the application's ability to detect features from text. The question being addressed when quantifying the performance of feature detection for the domain of biosurveillance is: *How well does the NLP application determine the values to the variables of interest from text?* For our SARS detector, examples of feature detection evaluations include how well the NLP application can determine whether a patient has a respiratory-related chief complaint, whether an ED report describes fever in a patient, or whether a patient has radiological evidence of pneumonia in a radiograph report.

Figure 17.5 illustrates the evaluation process for feature detection. Studies of feature detection do not evaluate the truth of the feature in relation to the patient–that is, whether the patient actually had the finding of interest–but only evaluate how well the technique interpreted the text in relation to the feature. Therefore the reference standard for an evaluation of feature detection is generated by experts who read the same text processed by the NLP application and assign values to the same variables. If the reference standard and the NLP application both believe the chest radiograph report describes the possibility of pneumonia, the NLP system is considered correct–even if the patient turned out to not have pneumonia.

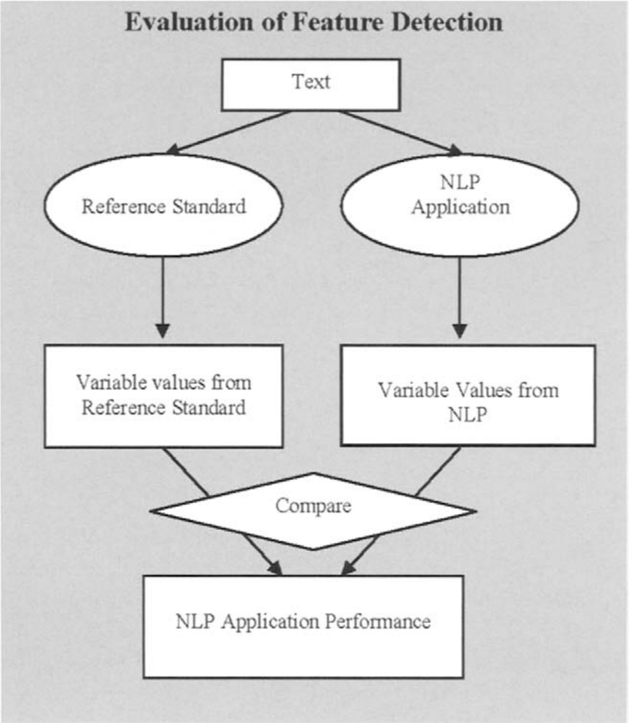

Evaluation of Feature Detection
Text
Reference Standard
NLP Application
Variable values from Reference Standard
Variable Values from NLP
Compare
NLP Application Performance

FIGURE 17.5 In an evaluation of feature detection, the NLP application and the reference standard independently extract the relevant variable values from the same text. Performance metrics are calculated by comparing the NLP output against that of the reference standard.

Several studies have evaluated how well NLP applications can encode findings and diseases, such as atelectasis, pleural effusions, CHF, stroke, and pneumonia from radiograph reports (Hripcsak et al., 1995, Friedman et al., 2004, Fiszman et al., 2000, Elkins et al., 2000). The reference standard for these studies was physician encodings of the variables, and the studies showed that the NLP applications performed similarly to physicians. One study (Chapman et al., 2004) evaluated how well the variable fever could be automatically identified in chief complaints and ED reports compared to a reference standard of physician judgment from the ED report. The application identified fever from chief complaints with 100% sensitivity and 100% specificity, and from ED reports with 98% sensitivity and 89% specificity.

Other studies have evaluated how well NLP technology can classify chief complaints into syndromic categories (e.g., respiratory, gastrointestinal, neurological, rash, etc.). Olszewski (2003) evaluated CoCo, a naive Bayesian classifier (Mitchell, 1997) that classifies chief complaints into one of eight syndromic categories. Chapman et al. (2005a) evaluated a chief complaint classifier (MPLUS [Christensen et al., 2002]) that used syntactic and semantic information to classify the chief complaints into syndromic categories. The reference standard for both studies was a physician reading the chief complaints and classifying them into the same syndromic categories. Performance of the NLP applications was measured with the area under the receiver operating characteristic (ROC) curve (areas under curve [AUC]), with AUCs ranging from 0.80 to 0.97 for CoCo and 0.95 to 1.0 for MPLUS, suggesting that NLP technology is quite good at classifying chief complaints into syndromes.

Studies of feature detection do not make claims about whether the NLP technology can accurately diagnose patients with the target findings, syndromes, or diseases. The conclusions only relate to the application's ability to determine the correct values for the variables given the relevant input text. Once feature detection has been validated, the next step is to apply the technology to the problem of diagnosing the patients and evaluate the technology's accuracy at case detection.

7.2. Case Detection

The question being addressed when measuring the case detection ability of an NLP application for the domain of biosurveillance is: *How well does the NLP application identify relevant patients from textual data?* For our SARS detector, examples of case detection evaluations include how well the NLP application can determine whether a patient has a respiratory syndrome, whether a patient has a fever, whether a patient has radiological evidence of pneumonia, or whether a patient has SARS.

Figure 17.6 illustrates the evaluation process for a study on case detection. The reference standard is generated by expert diagnosis of the patients. The source of the expert diagnosis

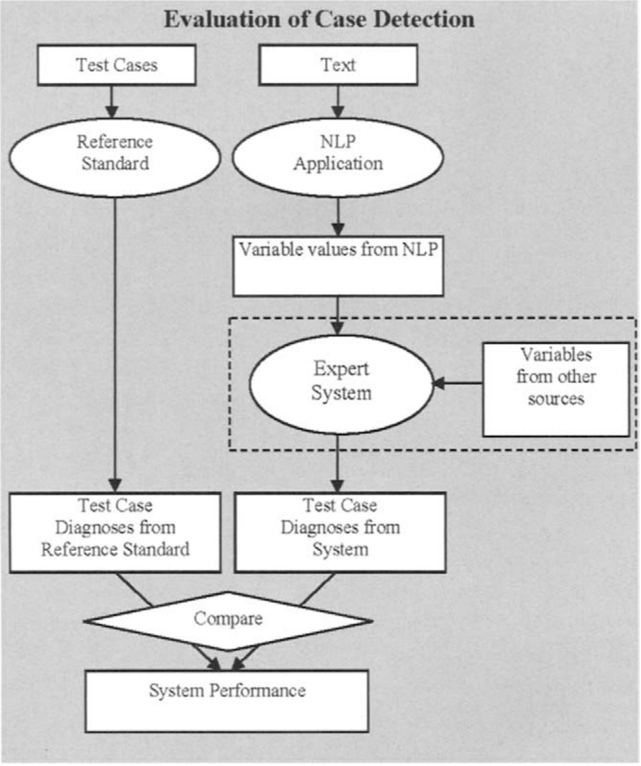

Evaluation of Case Detection
Test Cases
Text
Reference Standard
NLP Application
Variable values from NLP
Expert System
Variables from other sources
Test Case Diagnoses from Reference Standard
Test Case Diagnoses from System
Compare
System Performance

FIGURE 17.6 In an evaluation of case detection, the NLP application extracts relevant variable values from text, which may or may not be combined with variables from other sources (dashed box) to diagnose patients. The reference standard reviews test cases independently and generates a reference diagnosis. Performance metrics are calculated by comparing the diagnoses generated in part or in whole by the NLP application against that of the reference standard.

depends on the finding, syndrome, or disease being diagnosed, and may comprise review of textual patient reports or complete medical records, results of laboratory tests, autopsy results, and so on.

One of the first case detection studies involving an NLP-based system evaluated the ability of a computerized protocol to detect patients suspicious for Tuberculosis with data stored in electronic medical records (Hripcsak et al., 1997, Knirsch et al., 1998). In a prospective study, the system correctly identified 30 of 43 patients with TB. The computerized system also identified four positive patients not identified by clinicians. Aronsky et al. (2001) showed that a Bayesian network for diagnosing patients with pneumonia performed significantly better with information from the chest radiograph encoded with an NLP system than it did without that information (AUC 88% without NLP vs. 92% with NLP).

Several studies have evaluated how well automatically classified chief complaints can classify patients into syndromic categories (Espino and Wagner, 2001, Ivanov et al., 2002, Beitel et al., 2004, Chapman et al., 2005b, Gesteland et al., 2004). The studies used either ICD-9 discharge diagnoses or physician judgment from medical record review as the reference standard for the syndromic categories. The majority of the studies have focused on more prevalent syndromes—respiratory and gastrointestinal—but a few studies have evaluated classification into more rarely occurring syndromes, such as hemorrhagic and botulinic. Results suggest that syndromic surveillance from free-text chief complaints can diagnose patients into most syndromic categories with sensitivities between 40% and 77%, in spite of the limited nature of chief complaints.

In the section on feature detection, we described a study that evaluated the ability of an NLP application to determine whether chief complaints and ED reports described fever (Chapman et al., 2004). The fever study also measured the case detection accuracy of fever diagnosis from chief complaints and ED reports. The NLP application for identifying fever in chief complaints performed with perfect sensitivity and specificity in the feature detection evaluation. However, when quantifying how well the automatically extracted variable of fever from chief complaints identified patients who had a true fever based on reference standard judgment from the ED report, the chief complaint fever detector only performed with a sensitivity of 61%. The specificity remained at 100%. On the one hand, whenever a chief complaint mentioned fever, the patient actually had a fever, so there were no false-positive diagnoses from chief complaints. On the other hand, despite the fact that the NLP technology did not make any mistakes in determining if fever was described in a chief complaint, the chief complaints themselves did not always mention fever when the patient was febrile in the ED. As demonstrated by this study, coupling evaluations on feature detection with evaluations on case detection can inform us about the source of diagnostic errors, which could be the NLP technology, the input data itself, or a combination of the two.

7.3. Epidemic Detection

The question being addressed when measuring the epidemic detection performance of an NLP application in the domain of biosurveillance is *How well does the NLP application contribute to detection of an outbreak?* Evaluating epidemic detection is difficult. The first requirement for an epidemic detection study is reference standard identification of an outbreak. Outbreaks of respiratory and GI illnesses, such as influenza, pneumonia, and gastroenteritis, occur yearly throughout the country. Outbreaks of other infectious or otherwise concerning diseases, such as anthrax, West Nile virus, hemorrhagic fever, or SARS, rarely occur in the United States. Once an outbreak is identified, the next requirement for an epidemic detection evaluation is having access to textual data for an adequate sample of patients living in the geographical area of the outbreak.

One example of an evaluation of epidemic detection involving NLP technology was performed by Ivanov et al. (Ivanov et al., 2003). The evaluation used ICD-9 discharge diagnoses to define retrospective outbreaks of pediatric respiratory and gastrointestinal syndromes over a five year period (1998–2001) in four contiguous counties in Utah. Outcome measures were reported for correlation between chief complaint classifications and ICD-9 classifications and for timeliness of detection. Figure 17.7 from the Ivanov publication shows the time series plot of respiratory illness admissions (reference standard) and chief complaints. It is evident from the plot that chief complaints generated the same type of signal that the reference standard generated. Chief complaint classification detected three respiratory outbreaks with 100% sensitivity and specificity, and time series of chief complaints correlated with hospital admissions and preceded them by an average of 10.3 days.

A study by Irvin (Irvin et al., 2003) showed that numeric chief complaints could correctly detect an influenza outbreak between 1999 and 2000 with one false positive alarm. Although the chief complaints were numeric instead of textual, the same study design could be applied to free-text chief complaint classification for known outbreaks.

Evaluating feature detection is an important first step in evaluation of NLP techniques to ensure that the technology is working as expected. However, to truly understand the impact of NLP in outbreak and disease surveillance, evaluations of case detection and epidemic detection must also be performed.

8. SUMMARY

Natural language processing techniques are far from perfect. However, the question is not whether the techniques perform perfectly but whether the performance is good enough to contribute to disease and outbreak detection. For instance, a few errors in part-of-speech tagging or negation identification may not substantially decrease the ability of an NLP application to determine whether a patient has a fever. Evaluation studies of NLP in biosurveillance are still young, but we have learned a few things about how variables extracted from free-text medical records with NLP can contribute to outbreak detection.

First, we have learned that automated classification of free-text chief complaints, while not perfect, is sufficient to detect one-third to two-thirds of positive syndromic cases. Moreover, chief complaints for pediatric patients are accurate and timely at detecting respiratory and gastrointestinal outbreaks. Second, we have learned that ED reports can provide more detailed information about the state of a patient than chief complaints. For example, we can detect 40% more patients with fever from ED reports than from chief complaints (Chapman et al., 2004). Third, several researchers have shown that identification of radiological variables required for detection of many public health threats (including SARS and inhalational anthrax) from chest radiograph reports is feasible with NLP techniques.

This chapter has focused on applying NLP techniques to variable extraction from patient medical records, but other types of free-text documents contain information that may be useful for biosurveillance, including web queries, transcripts from call centers, and autopsy reports. Regardless of the type of free-text data, we suggest three questions to consider when deciding whether application of NLP techniques to textual data is feasible for disease and outbreak detection: (1) How complex is the text? The simple phrases in chief complaints are much easier to understand than complex discourses contained in ED reports. Textual data that require coreference resolution, domain modeling for inference, and other more difficult techniques required to identify values for the variables of interest will be more challenging to process and will be more prone to error. (2) What is the goal of the NLP technique? If the goal is to understand all temporal, anatomic, and diagnostic relations described in the text as well as a physician could, you may be in for a lifetime of hard but interesting work. Extraction of a single variable, such as fever, or encoding temporal, anatomic, and diagnostic relations for a finite set of findings, such as all respiratory findings, is more feasible. (3) Can the detection algorithms that will use the variables extracted with NLP handle noise? Detecting small outbreaks requires more accuracy in the input variables.

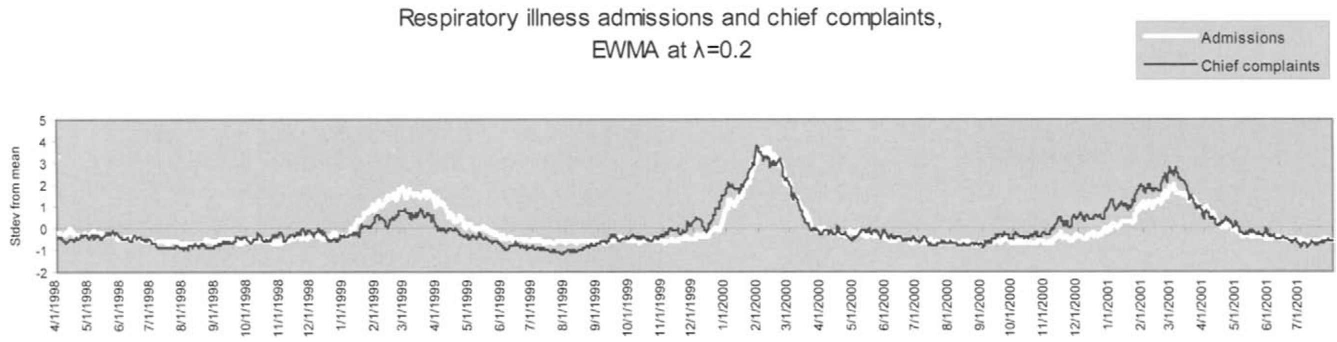


FIGURE 17.7 Time series plot of chief complaint syndromic classifications against ICD-9 discharge diagnoses for admissions of patients with respiratory illnesses including pneumonia, influenza, and bronchiolitis.

As an extreme example, some diseases such as inhalational anthrax require only a single case to be considered a threatening outbreak. If the NLP-based expert system did not correctly detect that case, then the detection system would have failed. However, in detecting an outbreak of a gastrointestinal illness, for example, if the NLP-based expert system only detected two-thirds of the true cases, there may still be enough positive patients to detect a moderate to large-sized outbreak. In addition, the consistent stream of false positive cases identified by the NLP-based expert system would comprise a noisy baseline that may not prevent the algorithm from detecting a significant increase in gastrointestinal cases but would require a larger increase to detect the outbreak. Consideration of these three questions can help determine the feasibility of using NLP for outbreak and disease surveillance.

NLP techniques can be applied to determine the values of predefined variables that may be useful in detecting outbreaks. The linguistic structure of the textual data being processed and the nature of the variables being used for surveillance determine the feasibility of applying NLP techniques to the problem. Characteristics such as linguistic variation, polysemy, negation, contextual information, finding validation, implication, and coreference must be accounted for to understand the information within patient medical reports as well as a physician does. However, because many of the variables helpful in biosurveillance do not require complete understanding of the text, NLP techniques may successfully extract variables useful for outbreak detection. In fact, evaluations of feature detection, case detection, and epidemic detection of NLP techniques have begun to demonstrate the utility of NLP techniques in this new field. More research in NLP techniques and more evaluation studies of the effectiveness of NLP will not only increase our understanding of how to extract information from text but will also help us continue to learn what types of data provide the most timely and accurate information for detecting outbreaks.

ADDITIONAL RESOURCES

Natural Language Processing Textbooks

- Allen, J. (1995). *Natural Language Understanding*. Redwood City, CA: Benjamin/Cummings Publishing Company.
- Charniak, E. (1993). *Statistical Language Learning*. Cambridge, MA: MIT Press.
- Jurafsky, D., Martin, J. H. (2000). *Speech and Language Processing*. Upper Saddle River, NJ: Prentice Hall.
- Manning, C. D., Schutze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Data Mining from Biomedical Texts

- Aronis, J. M., Cooper, G. F., Kayaalp, M., et al. (1999). Identifying patient subgroups with simple Bayes'. In: *Proceedings of American Medical Informatics Association Annual Symposium*, 658–62.
- Cooper, G. F., Buchanan, B. G., Kayaalp, M., et al. (1998). Using computer modeling to help identify patient subgroups in clinical data repositories. In: *Proceedings of American Medical Informatics Association Annual Symposium*, 180–4.
- Friedman, C., Kra, P., Yu, H., et al. (2001). GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *BioInformatics* 17 (Suppl 1): S74–82.
- Johnson, D. B., Chu, W. W., Dionisio, J. D., et al. (1999). Creating and indexing teaching files from free-text patient reports. In: *Proceedings of American Medical Informatics Association Annual Symposium*, 814–8.
- Krauthammer, M., Rzhetsky, A., Morozov, P., et al. (2000). Using BLAST for identifying gene and protein names in journal articles. *Gene* 259:1–2, 245–52.
- Liu, H., Friedman, C. (2003). Mining terminological knowledge in large biomedical corpora. *Pacific Symposium on Biocomputing*, 415–26.
- Lussier, Y. A., Shagina, L., Friedman, C. (2001). Automating SNOMED coding using medical language understanding: a feasibility study. In: *Proceedings of American Medical Informatics Association Annual Symposium*, 418–22.
- McCray, A. T. (1991). Extending a natural language parser with UMLS knowledge. In: *Proceedings of Annual Symposium on Computer Applications in Medical Care*, 194–8.
- McCray, A. T. (2000). Digital library research and application. *Stud Health Technol Inform* 76:51–62.
- McCray, A. T., Nelson, S. J. (1995). The representation of meaning in the UMLS. *Methods Inform Med* 34:1–2, 193–201.
- McCray, A. T., Razi, A. M., Bangalore, A. K., et al. (1996). The UMLS Knowledge Source Server: a versatile Internet-based research tool. In: *Proceedings of American Medical Informatics Association Fall Symposium*, 164–8.
- McCray, A. T., Soponsler, J., Brylawski, B., et al. (1987). The role of lexical knowledge in biomedical text understanding. In: *Symposium on Computer Applications in Medical Care* 87:103–7.
- McCray, A. T., Srinivasan, S., Browne, A. C. (1994). Lexical methods for managing variation in biomedical terminologies. In: *Symposium on Computer Applications in Medical Care*, 235–9.
- Mendonca, E. A., Cimino, J. J., Johnson, S. B., et al. (2001). Accessing heterogeneous sources of evidence to answer clinical questions. *J Biomed Inform* 34:85–98.
- Rzhetsky, A., Koike, T., Kalachikov, S., et al. (2000). A knowledge model for analysis and simulation of regulatory networks. *BioInformatics* 16:1120–8.
- Yu, H., Hatzivassiloglou V., Friedman, C., et al. (2002). Automatic extraction of gene and protein synonyms from MEDLINE and journal articles. In: *Proceedings of American Medical Informatics Association Symposium*, 919–23.

ACKNOWLEDGMENTS

We thank Thomas C. Rindflesch for his helpful comments.

REFERENCES

Allen, J. (1995). *Natural Language Understanding*. Redwood City, CA: Benjamin/Cummings Publishing Company.

Aronis, J. M., Cooper, G. F., Kayaalp, M., et al. (1999). Identifying patient subgroups with simple Bayes'. In: *Proceedings of American Medical Informatics Association Symposium*, 658–62.

Aronsky, D., Fiszman, M., Chapman, W. W., et al. (2001). Combining decision support methodologies to diagnose pneumonia. In: *Proceedings of American Medical Informatics Association Symposium*, 12–6.

Beitel, A. J., Olson, K. L., Reis, B. Y., et al. (2004). Use of emergency department chief complaint and diagnostic codes for identifying respiratory illness in a pediatric population. *Pediatr Emerg Care* 20:355–60.

Campbell, D. A., Johnson, S. B. (2001). Comparing syntactic complexity in medical and non-medical corpora. In: *Proceedings of American Medical Informatics Association Symposium*, 90–4.

Chapman, W. W., Bridewell, W., Hanbury, P., et al. (2001a). Evaluation of negation phrases in narrative clinical reports. In: *Proceedings of American Medical Informatics Association Symposium*, 105–9.

Chapman, W. W., Christensen, L. M., Wagner, M. M., et al. (2005a). Classifying free-text triage chief complaints into syndromic categories with natural language processing. *Artif Intell Med* 33:31–40.

Chapman, W. W., Cooper, G. F., Hanbury, P., et al. (2003). Creating a text classifier to detect radiology reports describing mediastinal findings associated with inhalational anthrax and other disorders. *J Am Med Inform Assoc* 10: 494–503.

Chapman, W. W., Dowling, J. N., Wagner, M. M. (2004). Fever detection from free-text clinical records for biosurveillance. *J Biomed Inform* 37:120–7.

Chapman, W. W., Dowling, J. N., Wagner, M. M. (2005b). Classification of emergency department chief complaints into seven syndromes: a retrospective analysis of 527,228 patients. *Ann Emerg Med.* 46(5):445–55.

Chapman, W. W., Fiszman, M., Frederick, P. R., et al. (2001b). Quantifying the characteristics of unambiguous chest radiography reports in the context of pneumonia. *Acad Radiol* 8:57–66.

Chapman, W. W., Fizman, M., Chapman, B. E., et al. (2001c). A comparison of classification algorithms to automatically identify chest X-ray reports that support pneumonia. *J Biomed Inform* 34:4–14.

Christensen, L., Haug, P. J., Fiszman, M. (2002). MPLUS: a probabilistic medical language understanding system. In: *Proceedings of Workshop on Natural Language Processing in the Biomedical Domain*, 29–36.

Chuang, J. H., Friedman, C., Hripcsak, G. (2002). A comparison of the Charlson comorbidities derived from medical language processing and administrative data. In: *Proceedings of American Medical Informatics Association Symposium*, 160–4.

Coden, A. R., Pakhomov, S. V., Ando, R. K., et al. (2005). Domain-specific language models and lexicons for tagging. *J Biomed Inform.* 38(6):422–30.

Cooper, G. F., Buchanan, B. G., Kayaalp, M., et al. (1998). Using computer modeling to help identify patient subgroups in clinical data repositories. In: *Proceedings of American Medical Informatics Association Symposium*, 180–4.

Elkins, J. S., Friedman, C., Boden-Albala, B., et al. (2000). Coding neuroradiology reports for the Northern Manhattan Stroke Study: a comparison of natural language processing and manual review. *Comput Biomed Res* 33:1–10.

Espino, J. U., Wagner, M. M. (2001). Accuracy of ICD-9-coded chief complaints and diagnoses for the detection of acute respiratory illness. In: *Proceedings of American Medical Informatics Association Symposium*, 164–8.

Fiszman, M., Chapman, W. W., Aronsky, D., et al. (2000). Automatic detection of acute bacterial pneumonia from chest X-ray reports. *J Am Med Inform Assoc* 7:593–604.

Fiszman, M., Chapman, W. W., Evans, S. R., et al. (1999). Automatic identification of pneumonia related concepts on chest x-ray reports. In: *Proceedings of American Medical Informatics Association Symposium*, 67–71.

Fiszman, M., Haug, P. J. (2000). Using medical language processing to support real-time evaluation of pneumonia guidelines. In: *Proceedings of American Medical Informatics Association Symposium*, 235–9.

Fiszman, M., Haug, P. J., Frederick, P. R. (1998). Automatic extraction of PIOPED interpretations from ventilation/perfusion lung scan reports. In: *Proceedings of American Medical Informatics Association Symposium*, 860–4.

Friedman, C. (2000). A broad-coverage natural language processing system. In: *Proceedings of American Medical Informatics Association Symposium*, 270–4.

Friedman, C., Alderson, P. O., Austin, J. H., et al. (1994). A general natural-language text processor for clinical radiology. *J Am Med Inform Assoc* 1:2, 161–74.

Friedman, C., Hripcsak, G. (1999). Natural language processing and its future in medicine. *Acad Med* 74:890–5.

Friedman, C., Hripcsak, G., Shablinsky, I. (1998). An evaluation of natural language processing methodologies. In: *Proceedings of American Medical Informatics Association Symposium*, 855–9.

Friedman, C., Hripcsak, G., Shagina, L., et al. (1999a). Representing Information in patient reports using natural language processing and the extensible markup language. *J Am Med Inform Assoc* 6:76–87.

Friedman, C., Knirsch, C., Shagina, L., et al. (1999b). Automating a severity score guideline for community-acquired pneumonia employing medical language processing of discharge summaries. In: *Proceedings of American Medical Informatics Association Symposium*, 256–60.

Friedman, C., Kra, P., Rzhetsky, A. (2002). Two biomedical sublanguages: a description based on the theories of Zellig Harris. *J Biomed Inform* 35:222–35.

Friedman, C., Shagina, L., Lussier, Y., et al. (2004). Automated encoding of clinical documents based on natural language processing. *J Am Med Inform Assoc.* 11(5):392–402.

Gesteland, P. H., Wagner, M. M., Gardner, R. M., et al. (2004). Surveillance of Syndromes during the Salt Lake 2002 Winter Olympic Games: An Evaluation of a Naive Bayes Chief Complaint Coder. In preparation.

Gundersen, M. L., Haug, P. J., Pryor, T. A., et al. (1996). Development and evaluation of a computerized admission diagnoses encoding system. *Comput Biomed Res* 29:351–72.

Hersh, W., Mailhot, M., Arnott-Smith, C., et al. (2001). Selective automated indexing of findings and diagnoses in radiology reports. *J Biomed Inform* 34:262–73.

Hripcsak, G., Austin, J. H., Alderson, P. O., et al. (2002). Use of natural language processing to translate clinical Information from a database of 889,921 chest radiographic reports. *Radiology* 224:157–63.

Hripcsak, G., Friedman, C., Alderson, P. O., et al. (1995). Unlocking clinical data from narrative reports: a study of natural language processing. *Ann Intern Med* 122:681–8.

Hripcsak, G., Knirsch, C. A., Jain, N. L., et al. (1997). Automated tuberculosis detection. *J Am Med Inform Assoc* 4:376–81.

Hripcsak, G., Knirsch, C. A., Jain, N. L., et al. (1999). A health Information network for managing innercity tuberculosis: bridging clinical care, public health, and home care. *Comput Biomed Res* 32:67–76.

Hripcsak, G., Wilcox, A. (2002). Reference standards, judges, and comparison subjects: roles for experts in evaluating system performance. *J Am Med Inform Assoc* 9:1–15.

Irvin, C. B., Nouhan, P. P., Rice, K. (2003). Syndromic analysis of computerized emergency department patients' chief complaints: an opportunity for bioterrorism and influenza surveillance. *Ann Emerg Med* 41:447–52.

Ivanov, O., Gesteland, P., Hogan, W., et al. (2003). Detection of pediatric respiratory and gastrointestinal outbreaks from free-text chief complaints. In: *Proceedings of American Medical Informatics Association Fall Symposium*, 318–22.

Ivanov, O., Wagner, M. M., Chapman, W. W., et al. (2002). Accuracy of three classifiers of acute gastrointestinal syndrome for syndromic surveillance. In: *Proceedings of American Medical Informatics Association Symposium*, 345–9.

Jain, N. L., Friedman, C. (1997). Identification of findings suspicious for breast cancer based on natural language processing of mammogram reports. In: *Proceedings of American Medical Informatics Association Fall Symposium*, 829–33.

Jain, N. L., Knirsch, C. A., Friedman, C., et al. (1996). Identification of suspected tuberculosis patients based on natural language processing of chest radiograph reports. In: *Proceedings of American Medical Informatics Association Fall Symposium*, 542–6.

Knirsch, C. A., Jain, N. L., Pablos-Mendez, A., et al. (1998). Respiratory isolation of tuberculosis patients using clinical guidelines and an automated clinical decision support system. *Infect Control Hosp Epidemiol* 19:94–100.

Liu, H., Lussier, Y. A., Friedman, C. (2001). Disambiguating ambiguous biomedical terms in biomedical narrative text: an unsupervised method. *J Biomed Inform* 34:249–61.

Lovis, C., Chapko, M. K., Martin, D. P., et al. (2001). Evaluation of a command-line parser-based order entry pathway for the Department of Veterans Affairs electronic patient record. *J Am Med Inform Assoc* 8:5, 486–98.

Manning, C. D., Schutze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Mitchell, T. M. (1997). *Machine Learning*. Boston: McGraw-Hill.

Morioka, C. A., Sinha, U., Taira, R., et al. (2002). Structured reporting in neuroradiology. *Ann N Y Acad Sci* 980:259–66.

Mutalik, P. G., Deshpande, A., Nadkarni, P. M. (2001). Use of general-purpose negation detection to augment concept indexing of medical documents: a quantitative study using the UMLS. *J Am Med Inform Assoc* 8:598–609.

Olszewski, R. T. (2003). Bayesian classification of triage diagnoses for the early detection of epidemics. In: *Proceedings of FLAIRS Conference*, 412–6.

Pakhomov, S. V., Ruggieri, A., Chute, C. G. (2002). Maximum entropy modeling for mining patient medication status from free text. In: *Proceedings of American Medical Informatics Association Symposium*, 587–91.

Roth, L., Hole, W. T. (2000). Managing name ambiguity in the UMLS metathesaurus. In: *Proceedings of American Medical Informatics Association Fall Symposium*, 1124.

Sinha, U., Dai, B., Johnson, D. B., et al. (2000). Interactive software for generation and visualization of structured findings in radiology reports. *AJR Am J Roentgenol* 175:609–12.

Sinha, U., Taira, R., Kangarloo, H. (2001a). Structure localization in brain images: application to relevant image selection. In: *Proceedings of American Medical Informatics Association Symposium*, 622–6.

Sinha, U., Ton, A., Yaghmai, A., et al. (2001b). Image content extraction: application to MR images of the brain. *Radiographics* 21:535–47.

Smith, L., Rindflesch, T., Wilbur, W. J. (2004). MedPost: a part-of-speech tagger for bioMedical text. *BioInformatics* 20: 2320–1.

Sneiderman, C. A., Rindflesch, T. C., Aronson, A. R. (1996). Finding the findings: identification of findings in medical literature using restricted natural language processing. In: *Proceedings of American Medical Informatics Association Fall Symposium*, 239–43.

Spyns, P. (1996). Natural language processing in medicine: an overview. *Methods Inform Med* 35:4–5, 285–301.

Szolovits, P. (2003). Adding a medical lexicon to an english parser. In: *Proceedings of American Medical Informatics Association Symposium*, 639–43.

Taira, R. K., Soderland, S. G. (1999). A statistical natural language processor for medical reports. In: *Proceedings of American Medical Informatics Association Symposium*, 970–4.

Travers, D. A., Haas, S. W. (2003). Using nurses' natural language entries to build a concept-oriented terminology for patients' chief complaints in the emergency department. *J Biomed Inform* 36:4–5, 260–70.

Travers, D. A., Waller, A., Haas, S. W., et al. (2003). Emergency department data for bioterrorism surveillance: electronic data availability, timeliness, sources and standards. In: *Proceedings of American Medical Informatics Association Symposium*, 664–8.

Wilcox, A. B., Narus, S. P., Bowes, W. A., 3rd. (2002). Using natural language processing to analyze physician modifications to data entry templates. In: *Proceedings of American Medical Informatics Association Symposium*, 899–903.

Wren, J. D., Garner, H. R. (2002). Heuristics for identification of acronym-definition patterns within text: towards an automated construction of comprehensive acronym-definition dictionaries. *Methods Inform Med* 41:426–34.