Chapter 15

SEMANTIC TEXT PARSING FOR PATIENT RECORDS

Carol Friedman

Department of Biomedical Informatics, Columbia University, New York, New York 10027

Chapter Overview

Accessibility to a comprehensive variety of different types of structured patient data is critical to improvement in the health care process, yet most patient information is in the form of narrative text. Semantic methods are needed to interpret and map clinical information to a structured form so that the information will be accessible to other automated applications. This chapter focuses on semantic methods that map narrative patient information to a structured coded form.

Keywords

natural language processing; data mining; electronic patient record; automated coding

"…better management of clinical information is a prerequisite for achieving patient safety and improved care … Because health care data are often narrative, natural language processing (NLP) is another important technique for mining data for quality improvement and patient safety purposes"

Institute of Medicine Report on Patient Safety, *Achieving a New Standard of Care*, 2003

1. INTRODUCTION

Advancement of health care is dependent on integration, organization, and utilization of massive amounts of genomic, pharmacological, cellular, tissular, environmental, and clinical information. The promise of the electronic health record (EHR) and the Clinical Data Architecture (CDA) (Dolin, Alschuler, et. al., 2001), an XML-based standardized exchange model for patient records, is that they will lead to substantial improvements in health care and biological research by facilitating the harnessing and accessibility of information in the EHR. The EHR is mainly expressed using natural language (i.e. narrative text), which is the primary means for communicating patient information in the patient reports. Although the content of patient reports includes a rich source of data, it is also a major bottleneck hindering widespread deployment of effective clinical applications because textual information is difficult if not impossible to reliably access by computerized processes. Natural language processing (NLP) systems are automated methods containing some linguistic knowledge that aim to improve the management of information in text. NLP systems have been shown to be successful for realistic clinical applications, such as decision support, surveillance of infectious diseases, research studies, automated encoding, quality assurance, indexing patient records, and tools for billing. Currently, NLP systems in clinical environments process patient records to: 1) index or categorize reports, 2) extract, structure, and codify clinical information in the reports so that the information can be used by other computerized applications, 3) generate text to produce patient profiles or summaries, and 4) improve interfaces to health care systems.

In a recent report, the Institute of Medicine recognized that NLP is potentially a very powerful technology for the medical domain because it enables a new level of functionality for health care applications that would not be otherwise possible (Institute of Medicine Report, 2003). NLP provides a method whereby large volumes of text reports (i.e. all textual patient reports) can be processed and the clinical information in the reports automatically encoded in a timely manner. It is not possible to have experts manually capture and encode a broad range of clinical information in the reports because it is too costly and time consuming, although limited coding is now being done manually. For example, primary and secondary diagnoses and procedures are coded manually for billing purposes. Once clinical information is encoded, it will be possible to develop a wide range of automated high throughput clinical applications, which should become invaluable tools that assist clinicians and researchers. These applications depend on structured data, and therefore are not currently feasible on a large scale. For example, an automated system could process enormous volumes

of reports to detect medical errors, whereas it would not be possible for experts to check such large volumes. Table 15-1 lists some queries that could be reliably answered using output from an NLP system. For example, it would be possible to check if a patient may have *pneumonia* based on information occurring in a radiological report of the chest. To answer this query accurately, it would be inadequate to perform only a keyword search for *pneumonia* because other contextual information in the report associated with pneumonia would also be required for a correct interpretation. For example by using a query with only the keyword *pneumonia*, false positives would be obtained if *pneumonia* 1) was in the process of being ruled out (e.g. *workup to rule out pneumonia*), 2) was ruled out (e.g. *no evidence of pneumonia*), 3) occurred in the past (e.g. *history of multiple admissions for pneumonia*), or 4) was associated with a family member (e.g. *father had pneumonia*). In addition, the query would also have to include clinical knowledge in order to associate descriptive findings, such as *consolidation*, with the condition *pneumonia*. Such expert knowledge would be necessary because descriptive findings indicative of *pneumonia* may occur in a report whereas the term *pneumonia* may not. The clinical rules that associate findings with clinical conditions typically consist of a logical combination of multiple findings. Depending on the condition and type of report, the rules could be quite complex. Because the output of an NLP system is consistent, NLP technology facilitates the development of automated alerting systems that are based on structured NLP output. An automated alerting system will consist of rules that are determined by experts, and it will be possible to incorporate the best and most current medical knowledge into the rules, promoting consistency, objectivity, and evidence-based medicine. Another significant advantage is that NLP technology can be used to standardize reports from diverse institutions and applications because the same automated system will uniformly encode clinical information that occurs in heterogeneous reports, thereby facilitating interoperability.

Table 15-1. Examples of queries that can be answered using structured output obtained by an NLP system as a result of processing textual patient reports

Does the patient have a particular clinical condition?
What medications is the patient on?
What problems does the patient currently have?
What procedures were performed on this patient?
Are there changes in a particular condition?
What problems did the patient have in the past?

This chapter will first present an overview of NLP in the clinical domain and describe challenges associated with processing clinical reports. Additionally, background information will be discussed concerning the

current state of the art in this domain. The chapter will end with a case scenario, describing the processing of a clinical narrative and the utilization of the output for an alerting clinical application.

2. OVERVIEW

2.1 Challenges of Processing Clinical Reports

NLP in the clinical domain has multiple challenges because of the health care setting, which we summarize below, and it is important for an NLP system to address such challenges if it is to be deployed in a clinical setting. A more detailed discussion can be found in (Friedman and Johnson, 2005).

2.1.1 Performance

Because the output of an NLP system will be employed by a healthcare application, it must have adequate recall, precision, and specificity for the intended clinical application, but it should be possible to adjust its performance according to the needs of the application. Different applications require varying levels of performance, which means that a clinical application involving NLP will have to undergo an evaluation before actually being deployed to ensure that the performance is appropriate.

2.1.2 Availability of Clinical Text and Confidentiality

Development of an NLP system is based on analysis (manual or automated) of samples of the text (i.e. a training set) to be processed. In the clinical domain, this means that large collections of online patient records in textual form must be available to the developers. However, patient records are confidential, and in order to make them accessible for research purposes, personal identifying information must be removed to comply with laws protecting patient confidentiality. Automated detection of identifying information within the text of the clinical reports, such as names, addresses, phone numbers, unique characteristics (i.e. mayor of New York), is an extremely difficult task that often requires manual review, although even after removal of names, addresses, etc. identification may still be possible, because as discussed by Sweeney et. al., rare characteristics may occur in the report that help identify a patient (Sweeney, 1997). Additionally, even if the data were manually checked, approval to use the records must be obtained from an Institutional Review Board and from institutional administrators.

2.1.3 Intra- and inter-operability

In order to be disseminated, an NLP system has to be able to function well in different health care facilities and for different clinical applications. It also has to be seamlessly integrated into a Clinical Information System, and generate output that is in a form usable by other components of the system. This generally means that the system will have to handle different interchange formats (i.e. XML, HL7) and heterogeneous formats that are associated with the different types of reports (i.e. formats of radiology reports, discharge summaries, echocardiograms, and pathology reports are generally different). An additional problem is that the NLP system will have to generate output that can be stored in an existing clinical repository, but the output often has complex and nested relations, and it may be impossible to map the NLP output to the schema of the clinical database without substantial loss of information. One more serious challenge is that to achieve widespread deployment, the NLP output has to be comparable so that it can be used across institutions for a variety of automated applications. This means it must be mapped to a controlled vocabulary and to a standard representation for the domain. Although different clinical vocabularies exist (i.e. UMLS, ICD-9, SNOMED-CT), none are complete and there is no single standard. An equally serious problem is that although there are standard controlled vocabularies, there is no standard representational model for medical language, and a representational model is also essential in order to interpret the underlying meaning of the clinical information in the reports and relationships among the information. Such a model would include relations between separate clinical concepts. For example, *treats*, is a relation between a medication event and a disease event in *On Asmacort for asthma*. Similarly, *suggestive of* is a relation between a diagnostic event and a disease event in *Chest x-ray suggestive of pneumonia*. In addition to relations among events there is information that modifies an individual event, such as negation (e.g. *no evidence of pneumonia*), certainty (e.g. *possible pneumonia*), severity (e.g. *mild cough*), change (e.g. *worsening cough*), and temporal information (e.g. *past history of pneumonia*). In 1994, the Canon Group (Evans, Cimino, et. al. 1994) attempted to merge different representational models to create a widely used model for medical language. That effort resulted in a common model for radiological reports of the chest (Friedman, Huff, et. al. 1995), but it has not been adopted by the community.

2.1.4 Evaluation

Evaluation of an NLP system is critical but difficult in the healthcare domain because of the difficulty of obtaining a gold standard and because it

is difficult to share the data across institutions. A fuller discussion on evaluation of NLP systems can be found in (Friedman and Hripcsak, 1998; Hripcsak, Austin et. al., 2002). Usually, there is no gold standard available that can be used to evaluate the performance of an NLP system. Therefore, for each evaluation, recruitment of subjects who are medical experts is generally required to obtain a gold standard for a test set. Obtaining a gold standard for this type of evaluation is very time consuming and costly.

2.1.5 Expressiveness

Language is extremely expressive in the sense that there are often different ways to describe the same medical concept and also numerous ways to express modifiers of the concept. For example, findings associated with cancer can be expressed using a very large number of terms, such as *neoplasm, tumor, lesion, growth, mass, infiltrate, metastasis, lymphoma, carcinoma, etc.* Similarly, a modifier, such as certainty information, is associated with more than 800 different phrases in the MedLEE lexicon, with terms such as *conceivable, definite, borderline, questionable, convincing evidence for, unlikely,* and *negative for*. Modifiers make it more complex to retrieve reports based on NLP structured output since they have to be accounted for, but then the query for retrieving the information will have a very fine granularity.

2.1.6 Heterogeneous Formats

There is no standardized structure for reports. Although sections and subsections of reports are important for many applications because they provide context, their names have not been standardized. For example, in New York Presbyterian Hospital (NYPH), there are many different section headers for reporting diagnostic tests (i.e. *Diagnostic Studies, Examination, Examination Type, Studies Performed*). Sometimes section headers are omitted or several sections are merged into one. For example, family and social history is occasionally reported in the History of Present Illness section. The Clinical Document Architecture (CDA) is an effort to address this problem because its aim is to establish standards for the structure of clinical reports (Dolin, Alschuler, et. al., 2001). Another problem occurs because the format for text within the reports is not standardized. Punctuation is often missing or is inappropriate, and a new line may be used instead of a period to signify the end of a sentence. An additional problem is that some reports contain tables with different configurations as well as text. Structured fields, such as those of a table, are easy for a human to interpret but are very problematic for a general NLP program because formatting

characteristics rather than phenomena associated with language determine the meaning of the fields and their relations.

2.1.7 Abbreviated Text

Generally, clinical reports are very compact, contain abbreviations, and often omit information that can easily be inferred by health care professionals based on their knowledge of the context and the domain. One problem with abbreviations is that they are highly ambiguous. For example, *pe,* may mean *physical examination, pleural effusion,* or *pulmonary embolism.* A clinical note, as shown in Figure 15-3, may have numerous abbreviations, which typically will occur in many reports (e.g. *yo, F, hx, HTN, COPD, CRI*). An additional problem is that a unique abbreviation may be defined in a single report. Omitted or implicit information present another challenge because an automated system, which utilizes the structured information generated from the reports, would have to automatically capture the implicit information based on knowledge of the domain, and this is a very complex task. A simple example is that the body location is frequently missing in a report of a particular diagnostic examination. For example, when *mass* occurs in a radiological report of the chest, it means *mass in lung* whereas if it occurs in a mammography report, it means *mass in breast.*

2.1.8 Interpreting Clinical Information

Clinical information in a report is important for an application, but frequently additional medical knowledge along with knowledge of the report structure is needed in order to associate findings with possible diagnoses. Interpretations of the findings also vary depending on the type of report and section. For example, retrieving information from the Admission Diagnosis section of a discharge summary is generally more straightforward than retrieving information from the Description section of a radiological report. Radiological reports typically do not have definitive diagnoses, and contain a continuum of findings that range from patterns of light (e.g. *patchy opacity*), to descriptive findings (e.g. *focal infiltrate*) to possible diagnoses (e.g. *pneumonia*). In some radiological reports, only the descriptive findings may be present and there may be no interpretation by the reporting radiologist (e.g. a finding *pneumonia* may actually be included in a report; instead, findings consistent with *pneumonia*, such as *consolidation* or *infiltrate* may occur). Therefore, in order to use an NLP system to detect pneumonia based on chest x-ray findings, the NLP system or application using the system would have to contain medical knowledge associated with findings that are suggestive of pneumonia. For example, two systems

developed by Fiszman and colleagues (Fiszman and Haug, 2000) and Hripcsak and colleagues (Hripcsak, Friedman, et. al., 1995) to detect patients with possible *pneumonia* from chest x-ray reports contain components that are used to infer *pneumonia* from the findings. However, it is also important to account for contextual information. Hripcsak and colleagues (Hripcsak, Friedman, et. al., 1995) found that the occurrence of pneumonia in the Clinical Information section was ambiguous: it could signify that 1) the patient had pneumonia based on the current examination, 2) the examination was a follow up examination for a patient with a known case, or 3) pneumonia was being ruled out. In contrast, if a finding of *pneumonia* occurs in the Impression or Description section, it is associated with findings in the current x-ray. Similarly, *history of heart disease* in the Family History section of a discharge summary does not mean the patient has heart disease. The rules needed to detect a particular condition based on output generated by an NLP system can be quite complex. In order to develop such a component, machine learning techniques can be used. This involves collecting instances of positive and negative samples, which would be used to develop rules automatically (Wilcox and Hripcsak, 1999), but this may be costly, since performance is impacted by sample size (McKnight, Wilcox, et. al., 2002) and for many conditions, a large number of instances would have to be obtained for satisfactory performance. An alternative involves having an expert manually write the rules by observing the target terms that the NLP system generates along with sample output. In that case, the rules will generally consist of combinations of Boolean operators (e.g. and, or, not) and findings. For example, a rule written by Chuang and colleagues (Chuang, Friedman, et. al., 2002), which detects a comorbidity of neoplastic disease, consisted of a Boolean combination of over 200 terms.

2.1.9 Rare Events

Natural language systems generally need a large number of training examples to train, refine, or test the system. Since some events occur rarely, it may be difficult to find a large number of reports for these events. Terminological knowledge sources, such as the UMLS (Lindberg, Humphreys, et. al., 1993) and the Specialist Lexicon (Browne, Divita et. al., 2003), may be helpful for providing lexical knowledge for rare clinical terms, but they may not include the variety of phrases that occur in natural language text.

2.2 Components of an NLP System

There are different approaches to NLP in the clinical domain. Most approaches use a combination of syntactic and semantic linguistic

knowledge as well as heuristic domain knowledge, but they vary as to the extent of each type of knowledge and as to how the different types are integrated. Some use manually developed rules, and others are more statistically oriented. Figure 15-1 shows a high level overview of a generic clinical application that utilizes NLP extraction technology.

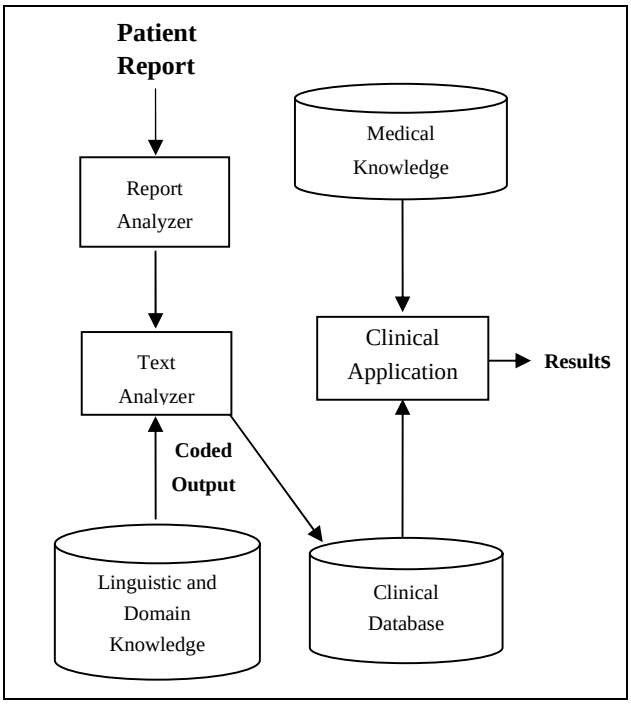

Figure 15-1. Components of a generic NLP Application in the Clinical Domain

First a **Report Analyzer** may be used to process the report in order to identify segments and to handle textual irregularities (i.e. tables, domain-specific abbreviations, missing punctuations). This process is typically straightforward to write, is heuristically-driven, and usually requires tailoring for each type of report. The next process is the **Text Analyzer**, which is the information extraction component that is the core NLP engine. It uses linguistic knowledge associated with syntactic and semantic features, and a conceptual model of the domain to structure and encode the clinical information and to generate output, which is then stored for subsequent use, generally in a structured coded clinical database. Once the data is structured, it may be used by an automated clinical application. However, use of the extracted patient information for an application typically requires additional

medical knowledge, depending on the application. For example, to detect patients who may have pneumonia based on radiological reports of the chest, the findings that were extracted will have to be interpreted based on expert knowledge.

A core NLP system will usually have a number of components, which are summarized below, but many variations are possible. A fuller discussion of syntax and semantics will follow the summary:

- **Morphological Analysis** is a process that breaks up the original words in a text so that they are in their canonical forms (i.e. *hands* → hand + -s), which are then used for lexical lookup. This process reduces the number of entries needed in a lexicon, because only the canonical forms and irregular forms need be defined, but not all the forms. One alternative approach eliminates this step and uses a lexicon containing entries for all forms; another approach uses a part of speech (POS) tagger, which identifies the syntactic part of speech of the words and possibly their canonical forms. This is discussed further below.
- **Lexical Look up** is a process where the words or phrases in the text are matched against a lexicon to determine their syntactic (i.e. noun, adjective, verb) and semantic properties (i.e. body part, disease, procedure). A lexicon requires extensive knowledge engineering and effort to develop and maintain. The Specialist Lexicon (Browne, Divita, et. al., 2003) maintained by the National Library of Medicine is comprehensive and contains syntactic information associated with both medical terms and general English terms, and is a valuable biomedical resource for NLP systems. Identification of syntactic classes using lexical lookup may not be necessary if POS tagging is used, but determining semantic properties, which are clinically relevant, is critical. A knowledge source, such as the UMLS or SNOMED may be used to identify semantic properties of words and phrases; however, since the focus of these terminologies concerns the classification of concepts, they generally do not include all the variations of the terms that may be found in text.
- **Syntactic Analysis** is a process that determines the structure of a sentence so that the relationships among the words in the sentence are established. Syntactic analysis may be complete or partial, or may not be used at all. Partial analysis usually revolves around identifying noun phrases.
- **Semantic Analysis** is a process that determines what words and phrases in the text are clinically relevant, and determines their semantic relations. An important requirement of semantic analysis is a **semantic model of the domain** or **ontology**. This model may be linguistically determined based on distributional patterns observed in the text, or it may be

developed based on deep knowledge of the domain. The semantic model may be in the form of frames (Sager, Friedman, et. al., 1987) or conceptual graphs (Baud, Rassinoux, et. al., 1995; Sowa, 1984). Semantic analysis structures and encodes the information, and may or may not depend on syntactic analysis. Structured output is generated, which is used subsequently by another automated process. Alternatively, semantic analysis may include statistical or knowledge components that classify the information in the text.

- **Encoding** is a process that maps the clinically relevant terms to well-defined concepts in a controlled vocabulary, such as the UMLS or SNOMED, or a local vocabulary. Encoding is necessary to achieve widespread use of the structured information by other automated processes, and is essential for intra-operability and inter-operability.

2.2.1 Syntax

Syntax is used to delineate the structure of the sentences in order to find the structural units of information and their relations. Some systems perform a comprehensive syntactic analysis whereas others perform a partial one. A system in the clinical domain that is based on a complete syntactic analysis is the Linguistic String Project system (Sager, 1981; Sager, Friedman, et. al., 1987), a pioneering system in the clinical domain. A complete syntactic analysis or parse of the sentence *patient experienced pain in left arm,* would determine that 1) the subject is a noun *patient*, 2) the verb is *experienced*, 3) the object is a noun phrase *pain in left arm*, 4) *in left arm* is a prepositional phrase that modifies *pain*, and 5) *left* modifies *arm*. In order to obtain a syntactic analysis, linguistic knowledge components consisting of a lexicon and grammar are typically used. The lexicon enumerates the terms that occur in clinical reports, and typically specifies the parts of speech (i.e. *discharge* is a singular noun or a verb), and the canonical forms (i.e. the canonical form *experience* for *experiences*, *experienced*, *experiencing*, and *experience*), which are used when generating the output in order to reduce variety. Grammar rules delineate the components of well-defined structures. For example, one rule would specify that one type of well-formed sentence consists of a noun phrase followed by a verb phrase followed by a noun phrase, and that one type of well-formed noun phrase consists of an adjective followed by a noun and a prepositional phrase.

Frequently, words are associated with more than one POS, and in order to obtain a correct syntactic analysis, that ambiguity must first be resolved. Prior to syntactic analysis, a process called POS tagging is generally used, which incorporates syntactic knowledge and contextual features of words surrounding a syntactically ambiguous word to determine the most likely

syntactic category. A statistically-based POS-tagging process is most frequently used. It is trained on a large corpus that has been correctly tagged (usually manually) for part of speech information. POS-tagging has been shown to be highly effective (> 95% accuracy) in the general English domain (http://www.coli.uni-sb.de/~thorsten/tnt/), but because it relies on a sample corpus, loss of accuracy occurs when moving from the general English domain to the clinical domain. Training on a clinical POS-tagged corpus would be highly desirable, but obtaining such a corpus is a very costly and time-consuming effort, and currently there are no publicly available clinical corpora because of the issue of patient confidentiality.

One difficulty with syntactic systems is the complexity of the grammar and parsing, and the prevalence of ambiguous structures. Often many analyses are obtained as a result of parsing, but most would not be correct. In contrast, humans have no trouble finding the correct parse(s) because they seamlessly use contextual and world-knowledge. For example, most people would be able to determine that *in joints* modifies *pain* in *patient experienced pain in joints*, and that *in the morning* does not modify *pain* but modifies *experienced* in *patient experienced pain in the morning.* A statistical parser can alleviate some of the ambiguity problems, but would require a large domain corpus that has been syntactically tagged so that it could be used as a training corpus to establish the appropriate parse probabilities. This has been accomplished for the general English domain (http://www.cis.upenn.edu/~treebank/home.html), but, as discussed above, no such corpus exists in the clinical domain. A partial syntactic analysis is a much simpler, more efficient, and more robust process, but has lower precision. Such an analysis could be used to determine that the sample sentence above has the noun phrases *patient, pain,* and *left arm*, and a verb phrase *experienced* but the relations between the phrases are not identified. Partial parsing could be rule-based or statistical-based. For the latter case, phrases such as noun phrases, could be detected using statistical methods that are trained on a large training corpus that has been correctly tagged to identify noun phrases in addition to POS tags. Since clinical information primarily occurs in noun phrases, this technique was used by Hersh and colleagues (Hersh, Mailhot, et. al., 2001) to index radiology reports for retrieval purposes. Another use of partial parsing would be to determine the verb and its arguments. Partial parsing results in a decrease in precision, especially when processing long complex sentences, and may not be adequate for clinical applications, such as an alerting application.

2.2.2 Semantics

Semantic analysis is used to determine the clinically relevant semantic components of a sentence, and to establish the semantic relationships among them. Thus, a semantic analysis establishes a real-world clinical interpretation of the sentence. It is generally based on a semantic knowledge base, which semantically classifies words and phrases that occur in the domain, and then establishes an interpretation of their relationships. For example, *pain* would have a semantic category **sign/symptom**, whereas *arm* would have a category **body part.** *Pain in arm* would be interpreted as a symptom occurring in a body part. Semantic knowledge is incorporated into systems in many different ways. In some systems the semantic components are separated from the syntactic components, although the processor may interleave their use.

The LSP system (Sager, 1987) used a semantic constraint component to specify well-formed semantic patterns for particular syntactic structures in order to obtain an improved syntactic analysis of the sentences. For example, an incorrect analysis of *patient experienced pain in the morning* would be eliminated if a phrase *in the morning* modified *pain.* Another separate semantic component was used by the LSP system to map a syntactic analysis that was normalized into an appropriate semantic template or format. In that system, templates form the model of the clinical domain, and different templates were designed to model different types of information. A template represents clinical information by associating slots with specific types of information and interprets the relations among the slots by associating predefined relations for the slots of the templates. For example, a patient state frame would be built for the sample sentence above, and it would contain slots that will be instantiated with patient state information, such as a **sign/symptom** slot (*pain),* a **body part** slot (*arm*), and an **evidential** slot (*certainty of sign/symptom information*) by using a process that fills slots based on the semantic classes of the words.

The MPLUS system (Christensen, Haug, et. al., 2002) represents another substantial NLP effort in the clinical domain. It is similar to the LSP system in that it has a separate syntactic parsing component, lexicon, and grammar. However, the semantic component consists of Bayesian networks (BN) that are integrated into the syntactic parsing process. The BN establishes possible values of a node based on a set of training cases, which are used to learn a probability function by considering probabilities of neighboring nodes. Parsing proceeds from the bottom up, from the word level to more complex phrase levels. When a word is parsed, a BN instance is attached to it establishing a semantic interpretation to the word, and when a phrase is parsed, the BN instances of the components of the phrase are unified and

attached to the phrase, resulting in a semantic interpretation of the relations between the components.

Another comprehensive NLP system in the clinical domain is the MedLEE system (Friedman, Alderson, et. al., 1994). It differs from the LSP and MPLUS systems in that it has an integrated syntactic and semantic component, which is realized in the form of a grammar. The MedLEE grammar consists of a specification of semantic (and sometimes syntactic) components and is used to interpret the semantic properties of the individual terms and of their relations with other terms, and to generate a target output form. Thus, one grammar rule could contain both syntactic and semantic components. For example, a rule specifies that a sentence containing sign/symptom information consists of a phrase associated with a patient (i.e. *patient*), followed by an evidential verb (*e.g. experienced*), followed by a phrase that contains a sign/symptom (e.g. *pain in arm*). The semantic grammar rules were developed based on co-occurrence patterns observed in clinical text.

2.2.3 Domain Knowledge

Some systems are based on a sound knowledge representation that models the underlying domain, and use the model to achieve a semantic analysis. For example, Baud and colleagues (Baud, Rassinoux, et. al., 1995) modeled a typology of concepts and relations for the domain of digestive surgery. Their NLP system uses a proximity parsing (Baud, Rassinoux, et. al., 1992) approach to obtain a semantic analysis of the text. In this approach words and sequences of words are grouped together if their concepts are semantically compatible according to the model under consideration. Syntax may be used if available, but the method does not rely on it. The method provides advantages for multi-lingual capabilities because it is based on domain concepts and is less dependent on characteristics of different languages. Another system that relies on a knowledge-rich infrastructure was developed by Hahn and colleagues (Hahn, Romacker, et. al., 2002). This system maps text to a knowledge base that contains a formal representation of the content of the text. The system also performs a syntactic analysis that is driven by lexical definitions and a syntactic dependency grammar. A domain-specific lexicon is used that is geared to the needs of the particular clinical subdomain.

2.3 Clinical Applications

Natural language systems have been employed and evaluated in the clinical domain for a variety of clinical applications. Table 15-2 provides a

list of papers describing use of NLP for processing clinical textual documents in different domains for a variety of different applications. Although the list is comprehensive and represents a broad collection of different clinical applications, it is not complete. In addition, methods that use NLP for medical applications that do not involve the processing of clinical documents, such as those focusing on clinical terminology, the processing of journal articles and consumer health messages, are not included.

Table 15-2. Clinical Applications Using NLP Technology

Reference	*Clinical Domain*	*Application*
Lyman, Sager, et al., 1991	Progress notes	Quality assessment
Baud, Rassinoux, et al., 1992	Surgical notes	Encoding and improved browsing; Multilingual capabilities
Moore and Berman, 1994	Pathology	Key diagnoses for indexing
Zweigenbaum, et al., 1994	Discharge summary	ICD-9-CM for indexing
		Multilingual capabilities
Hripcsak, Friedman, et al., 1995	Radiology	Detecting clinical conditions
Gunderson, Haug, et al., 1996	Admission diagnoses	ICD-9 encoding
Sager, Nhan, et al., 1996, and Lussier, Y., and Shagina, et al,. 2001	Discharge summary	SNOMED encoding for granular retrieval
Knirsch, Jain, et al., 1998	Radiology	Isolating patients with tuberculosis
Fiszman, M., and Haug, et al., 1998	Ventilations/ perfusion lung scan	Interpret findings
Blanquest and Zweigenbaum, 1999	Discharge summary	ICD-10 encoding
Aronsky and Haug, 2000	Radiology	Assessing severity of pneumonia
Friedman, Knirsch, et al., 1999	Radiology and Discharge summary	Assessing severity of pneumonia
Fiszman, M., and Haug 2000	Radiology	Pneumonia guidelines
Tuttle, M.S., Olsen, N.E., et al., 1998; Aronson, 2001; Nadkarni, Chen, et al., 2001; Leroy and Chen, 2001; Zou, Chu, et al., 2003; Friedman, Shagina, et al., 2004	Biomedical text	Mapping text to UMLS codes
Heinze, Morsch, et al., 2001b	Radiology, Emergency Medicine	Coding for billing

continued

Reference	*Clinical Domain*	*Application*
Heinze, Morsch, et al., 2001a	Variety of domains	Data mining
Hripcsak, Austin, et al., 2002	Radiology	Data mining
Hahn, Romacker, et al., 2002	Histopathology of gastro-intestinal domain	Knowledge acquisition
Mamlin, Heinze, et al., 2003	Radiology	Extract cancer-related findings
Schadow and McDonald, 2003	Surgical pathology	Extract specimens, related pathological findings
Xu, Anderson, et al., 2004	Surgical pathology	Obtaining variables for clinical study
Mitchell, Becich, et al., 2004	Surgical pathology	Detecting negated concepts
Liu and Friedman, 2004; and Meng, Taira, et al., 2004	Clinical reports	Summarize patient information

3. CASE SCENARIO

The MedLEE NLP system was developed to address a need to transform narrative text in patient reports to structured and encoded data so that the data could be stored in the Clinical Repository at NYPH and accessed by other applications, such as a monitoring system, which consists of medical rules that access data in the repository. For example, an alerting application could be used for newborn infants in a neonatal intensive care unit (ICU) to screen for hospital-acquired pneumonia. There are many different types of clinical reports in a healthcare institution, such as radiology reports, resident sign out notes, discharge summaries, and pathology reports. Although each contains text, their overall formats differ.

CLINICAL INFORMATION:
3 day old male with resp dist.
IMPRESSION:
Opacities are noted in left and right lobe of lung, which is consistent with pneumonia or atelectasis.
DESCRIPTION:
a. p. portable chest radiograph is submitted. Bibasilar opacities, right greater than left, this may represent pneumonia or atelectasis; otherwise, there is no interval change from 6/19/01.

Figure 15-2. Sample Radiology Report

76 yo F hx of HTN, COPD, CRI, DM2 from NH after noted to have fever, cough, change in MS. R/O PNA. Given IV ABX in ED, R/O Staph. CXR with bilobar involvement
ALL: NKDA
Meds: Vanco 500mg, Azithro, tylenol, atenolol, alb/atrov, NPH

Figure 15-3. Sample Resident Sign Out Note

Figure 15-2 shows a radiology report of the chest taken from an infant in the ICU with respiratory distress. Radiology reports generally have three sections, which are indicated by special section headers: **Clinical Information**, which usually contains some information related to the indication for the examination, **Impression**, which lists or interprets the most relevant findings, and **Description**, which describes the findings. All sections contain text, but the sentence structures range from noun phrases to complex sentences, and include some abbreviations. The report shown has important information for a pneumonia ICU alerting application because it contains a finding *opacities*, which may indicate *pneumonia* or *atelectasis* (e.g. a collapsed lung or lobe of lung). In order to process the report, a special preprocessor will first be used to transform the report so that the overall structure (i.e. sections followed by text associated with the section) is correct for the MedLEE core NLP processor. This will be a preprocessor tailored to the specific type of report. In the case of radiology reports at NYPH, a special report preprocessor is not needed because the report structure is in a form that is appropriate for MedLEE to use directly. However, for other types of reports, special preprocessing may be necessary. For example, the resident sign out report shown in Figure 15-3 requires a special purpose preprocessor to establish appropriate sections, to add sentence endings when necessary, and to handle abbreviations, which are prevalent in this type of report. For the sign out note example, the preprocessor will add a section header, SUMMARY, to the beginning of the report, will change *ALL* to a section ALLERGIES, and *Meds* to a section MEDICATIONS. A period will be added after NKDA and NPH to indicate the end of the two sentences. Additionally, abbreviations will be expanded based on knowledge of the abbreviations used in the domain. For example, *F* will be changed to *female*, *hx* to *history*, etc.

The MedLEE core NLP engine will be used next to process the report; it first identifies each section and then processes all the sentences within the section. Thus, in Figure 15-2, the sentence *opacities are noted in left and right lobe of lung, which is consistent with pneumonia or atelectasis*, which is in the Impression section, will be processed first. The lexicon will be used to identify the semantic and or the syntactic categories of the terms and to

specify the canonical form. For example, the words *opacities, pneumonia,* and *atelectasis* will each be associated with the semantic category **pathological finding.** In addition, *left* and *right* will be associated with the semantic category **region**, *lung* will be associated with **bodyloc**, and *consistent with* will be associated with a **relation** that connects findings. The parser will use the semantic grammar to identify the semantic relations in the sentence and to generate structured output, which for the sample sentence, will consist of four findings, each of which is represented by a tag **problem**. Several grammar rules will be used and satisfied in the course of parsing this sentence in order to match the semantic sequence of the words and phrases in the sentence to the rules. One rule that will be satisfied is a rule specifying that a finding (e.g. *opacities*) is connected (e.g. via the relation *consistent with*) to one of two findings (e.g. *pneumonia* or *atelectasis*). Another rule that will be satisfied is a rule specifying that a finding (e.g. *opacities*) is followed by a verbal phrase (e.g. *are noted*), which is associated with certainty information, and a phrase specifying body location information (e.g. *lobe*). Finally, another rule will be satisfied specifying that body location *lobe* is modified by a body location *lung*, and also by two regions *left* and *right.* Each rule in the grammar not only identifies the semantic components for that structure, but also interprets their relations and specifies an output structure, which is a composition of the output structures of each of the components. A simplified XML form of the output generated by processing the sample sentence is shown in Figure 15-4. The complete XML form has additional attributes representing the associated codes if coding is requested, and other contextual attributes, but these were omitted from the figure for simplicity. In Figure 15-4, the first problem tag has the value **opacity**, and modifiers **bodyloc** and **certainty**, which are nested within the **problem** tag. The **bodyloc** modifier with the value **lobe** has a nested body location whose value is **lung,** and a **region** modifier whose value is **left.** The value of the certainty modifier is **high certainty**, which is the target form of *are noted.* The second problem, whose value is also **opacity**, is almost identical to the first except the **region** modifier of **lobe** has the value **right**. This output is the result of expanding the conjunction relation in *left and right lobe of lung* to obtain separate findings. The remaining two findings have a **problem** tag with the value **pneumonia** and a **problem** tag with the value **atelectasis.** Each of the two findings has a **certainty** modifier, stemming from *consistent with* whose target output form is **moderate certainty**.

Once the output is in a structured form as shown below, it will be transformed to a form suitable for storage in a particular clinical database, and subsequently, a query may be used to reliably access the information. For example, a rule used to detect infants who may have hospital-acquired

pneumonia will look for findings with certain characteristics, such as one which has a **problem** tag with a value **opacity** along with certain other modifiers, such as modifiers signifying that the opacity is in both lobes of the lung. The rule will also look for a **problem** with the value **pneumonia**. However, findings that have characteristics that match the rule may also be filtered out depending on the presence of certain modifiers that negate the event, signify that it occurred in the past, or signify that it did not actually occur. For example, if the **certainty** modifier of **problem** is **no** or **rule out**, it will be filtered out because it signifies that there was no evidence for the finding or that the patient is being evaluated for the finding. Similarly, the finding will be filtered out if it has a temporal type of modifier with a value that signifies the finding is associated with a previous event. This will happen if the modifier is **status** with a value **previous** or a temporal modifier **date** with a previous date, such as **19900502** (5/2/1990).

```
<section v="impression">
   <problem v = "opacity"><bodyloc v = "lobe"><bodyloc v = "lung"/>
                                    <region v = "left"/></bodyloc>
                              <certainty v = "high certainty"/>
   </problem>
   <problem v = "opacity"><bodyloc v = "lobe"><bodyloc v = "lung"/>
                                 <region v = "right"/></bodyloc>
                                 <certainty v = "high certainty"/>
   </problem>
   <problem v = "pneumonia"><certainty v = "moderate certainty"/>
   </problem>
   <problem v = "atelectasis"><certainty v = "moderate certainty"/>
   </problem>
</section>
```

Figure 15-4. Simplified output form generated as a result of processing the sentence shown in the Impression section of Figure 15-2.

In the sample sentence above, all the words of the sentence were known to the system and the grammar rules were completely satisfied. In some cases, it may not be possible for the system to obtain a parse for the complete sentence because there may be words that are unknown to the system or the sentence may not conform to the grammar rules. For those cases, the parsing strategy is relaxed and different strategies are attempted so that a parse is always obtained if there is relevant clinical information in the sentence. The first relaxation attempt ignores unknown words and then tries to obtain a parse. For example, if the sample sentence is *opacities are noted*

in lft lobe of lung, the segment *opacities are noted in lobe of lung* will be parsed successfully. Other strategies are based on segmenting the sentence in different ways and attempting to parse the segments. However, these relaxation strategies are aimed at improving recall at the expense of some loss of precision, and may be undesirable for some clinical applications. The way this is handled is that the parse mode (i.e. the parsing strategy that was used) is saved as a special modifier **parse mode** in the structured output form and its value is the parse mode that was used to obtain the output. In this way, findings can be filtered out if the parse mode is not reliable enough for the application.

4. CONCLUSIONS AND DISCUSSION

Improved automated methods are needed to advance the quality of patient care, to reduce medical errors, and to lower costs, but these methods depend on reliable access to a comprehensive variety of patient data. However, the primary means of communication in healthcare is narrative text, and therefore natural language processing techniques are needed to structure and encode the narrative text so that the clinical information will be in a form suitable for use by automated applications. NLP is a difficult, complex, and knowledge intensive process. It involves the integration of many forms of knowledge, including syntactic, semantic, lexical, pragmatic, and domain knowledge. Successful NLP methods have been developed, and applications using NLP methods have been evaluated demonstrating their effective use in healthcare settings. However, these efforts represent individual instances, mostly at individual sites, and have not been widely disseminated. In order for NLP to become more widespread, standardization at several different levels is critical: namely a standardization of report structures, a standardization of models representing clinical information, and a standardized clinical vocabulary.

5. ACKNOWLEDGEMENTS

Work on the MedLEE NLP system was supported by grants LM07659 and LM06274 from the National Library of Medicine, by the New York State sponsored Columbia University Center for Advanced Technology, and by the Research Foundation of CUNY.

REFERENCES

Aronsky, D. and Haug, P. J. (2000). "Assessing the Quality of Clinical Data in a Computer-based Record for Calculating the Pneumonia Severity Index," *Joiurnal of the American Medical Informatics Association,* 7(1):55-65.

Aronson, A.R. (2001). "Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program," in *Proceedings of the 2001 AMIA Symposium*:17-21.

Baud, RH., Rassinoux, A.M. and Scherrer, J.R. (1992). "Natural Language in Processing and Semantical Representation of Medical Texts," *Methods of Information in Medicine,* 31(2):117-125.

Baud, RH., Rassinoux, A.M., Wagner, J.C. and Lovis, C. (1995). "Representing Clinical Narratives Using Conceptual Graphs," *Methods of Information in Medicine,* 1/2:176-186.

Blanquet, A. and Zweigenbaum, P. (1999). "A Lexical Method for Assisted Extraction and Coding of ICD-10 Diagnoses from Free Text Patient Discharge Summaries," in *Proceedings of the 1999 AMIA Symposium:*1029.

Browne, A. C., Divita, G., Aronson, A.R. and McCray, A.T. (2003). "UMLS Language and Vocabulary Tools," in *AMIA Annual Symposium Proceedings*:798.

Christensen L, Haug, P.J., and Fiszman, M. (2002). "MPLUS: A Probabilistic Medical Language Understanding System," in *Proceediongs of the ACL 2002 Workshop on Natural Language in Processing in Biomedicine:*29-36.

Chuang, J. H., Friedman, C., Hripcsak, G. (2002). "A Comparison of the Charlson Comorbidities Derived from Medical Language in Processing and Administrative Data," in *Proceedings of the AMIA.Symposium.:*160-164.

Dolin, R.H., et al. (2001). "The HL7 Clinical Document Architecture," *Journal of the American Medical Informatics Association,* 8(6):552-569.

Evans, D.A., et al. (1994). "Toward a Medical Concept Representation Language," *Journal of the American Medical Informatics Association,* 1(3):207-217.

Fiszman, M. and Haug, P.J. (2000). "Using Medical Language in Processing to Support Real-time Evaluation of Guidelines," in *Proc AMIA Symp 2000;* 235-239.

Fiszman, M, Haug, P.J., and Frederick, P.R. (1998). "Automatic Extraction of PIOPED Interpretation from Ventilation/perfusion Lung Scan Reports," in *Proc 1998 AMIA Symp:*860-864.

Fiszman, M. and Haug, P.J. (2000). "Using Medical Language in Processing to Support Real-time Evaluation of Pneumonia Guidelines," in *Proc 2001 AMIA Symp*:235-239.

Friedman, C. and Johnson, S.B. (2005). "Natural Language and Text in Processing in Biomedicine," in Shortliffe EH and JJ. Cimino, Eds. *Biomedical Informatics: Computer Applications in Health Care and Medicine.* Springer.

Friedman, C., et al. (1994). "A General Natural Language Text in Processor for Clinical Radiology," *Journal of the American Medical Informatics Association,* 1(2):161-174.

Friedman, C. and Hripcsak, G. (1998). "Evaluating Natural Language in Processors in the Clinical Domain," *Methods of Information in Medicine,* 37:334-344.

Friedman, C., et al. (1995). "The Canon Group's Effort: Working Toward a Merged Model," *Journal of the American Medical Informatics Association,* 2(1):4-18.

Friedman, C., Knirsch, C.A., Shagina, L., and Hripcsak, G. (1999). "Automating a Severity Score Guideline for Community-acquired Pneumonia Employing Medical Language in Processing of Discharge Summaries," in *Proc 1999 AMIA Symp:*256-260.

Friedman, C., Shagina, L., Lussier, Y. and Hripcsak, G. (2004). "Automated Encoding of Clinical Documents Based on Natural Language in Processing," *Journal of the American Medical Informatics Association,;* 11(5):392-402.

Gundersen, M.L., et al. (1996). "Development and Evaluation of a Computerized Admission Diagnoses Encoding System," *Computers and Biomedical Research*; 29:351-372.

Hahn, U., Romacker, M. and Schulz, S. (2002). "Creating Knowledge Repositories from Biomedical Reports: The MEDSYNDIKATE Text Mining System," in *Pac Symp Biocomput.:*338-349.

Heinze, D.T., Morsch, M.L., and Holbrook, J. (2001a). "Mining Free-text Medical Records," in *Proceedings AMIA.Symp.:*254-258.

Heinze, D.T., et al. (2001b). "LifeCode - A Deployed Application for Automated Medical Coding," *AI Magazine;* 22(2):76-88.

Hersh, W., Mailhot, M., Arnott-Smith, C., and Lowe, H. (2001). "Selective Automated Indexing of Findings and Diagnoses in Radiology Reports," *J.Biomed.Inform.;* 34(4):262-273.

Hripcsak, G., Austin, J.H., Alderson, P.O., and Friedman, C. (2002). "Use of Natural Language in Processing to Translate Clinical Information from a Database of 889,921 Chest Radiographic Reports," *Radiology*; 224(1):157-163.

Hripcsak, G., et al. (1995). "Unlocking Clinical Data from Narrative Reports," *Ann.of Int.Med.;* 122(9):681-688.

Knirsch, C.A., et al. (1998). "Respiratory Isolation of Tuberculosis Patients Using Clinical Guidelines and an Automated Decision Support System," *Infection Control and Hospital Epidemiology;* 19(2):94-100.

Leroy, G. and Chen, H. (2001). "Meeting Medical Terminology Needs--The Ontology-Enhanced Medical Concept Mapper," *IEEE Trans.Inf.Technol.Biomed.;* 5(4):261-270.

Lindberg, D.A.B., Humphreys, B. and McCray, A.T. (1993). "The Unified Medical Language System," *Methods of Information in Medicine,* 32:281-291.

Liu, H. and Friedman, C. (2004). "CliniViewer: A Tool for Viewing Electronic Medical Records Based on Natural Language in Processing and XML," *Medical Informatics,* 2004:639-643.

Lussier, Y., Shagina, L., and Friedman, C. (2001). "Automating SNOMED Coding Using Medical Language Understanding: A Feasibility Study," in *Proceedings of the 2001 AMIA Symposium,* 418-422.

Lyman, M., et al. (1991). "The Application of Natural-language in Processing to Healthcare Quality Assessment," *Medical Decision Making,* 11(suppl):S65-S68.

Mamlin, B.W., Heinze, D.T., and McDonald, C.J. (2003). "Automated Extraction and Normalization of Findings from Cancer-related Free-text Radiology Reports," *AMIA Annual Symposium Proceedings:*420-424.

McKnight, L.K., Wilcox, A. and Hripcsak, G. (2002). "The Effect of Sample Size and Disease Prevalence on Supervised Machine Learning of Narrative Data," in *Proceedings of the AMIA.Symposium:*519-522.

Meng, F., et al. (2004). "Automatic Generation of Repeated Patient Information for Tailoring Clinical Notes," *Medical Informatics,* 2004:653-657.

Mitchell, K.J., et al. (2004). "Implementation and Evaluation of a Negation Tagger in a Pipeline-based System for Information Extract from Pathology Reports," *Medical Informatics,* 2004:663-667.

Moore, G.W. and Berman, J.J. (1994). "Automatic SNOMED Coding," in *Proceedings of the Eighteenth Annual Symposium on Computer Applications in Medical Care.*

Nadkarni, P., Chen, R. and Brandt, C. (2001). "UMLS Concept Indexing for Production Databases: A Feasibility Study," *Journal of the American Medical Informatics Association,* 8(1):80-91.

National Academy of Science. (2003). "Patient Safety: Achieving a New Standard for Care," Aspden, P., Corrigan, J.M., Wolcott, J., and Erickson, S.M. Washington, D.C.

Sager, N. (1981). *Natural Language in Processing: A Computer Grammar of English and Its Applications.* Mass, Addison-Wesley.

Sager, N., Friedman, C., Lyman, M. and et al. (1987). *Medical Language in Processing: Computer Management of Narrative Data.* Reading, MA: Addison-Wesley.

Sager, N., Nhan, N.T., Lyman, M., and Tick, L.J. (1996). "Medical Language in Processing with SGML Display," in *Proceedings of the 1996 AMIA Symposium:*547-551.

Schadow, G. and McDonald, C.J. (2003). "Extracting Structured Information from Free Text Pathology Reports," in *AMIA Annuual Symposium Proceedings*:584-588.

Sowa, J. (1984). *Conceptual Structures: Information in Processing in Mind and Machine.* Reading: Addison-Wesley.

Sweeney, L. (1997). "Weaving Technology and Policy Together to Maintain Confidentiality," *Journal of Law, Medincein and Ethics*, 25(2-3):98-110, 82.

Tuttle, M.S., et al. (1998). "Metaphrase: An Aid to the Clinical Conceptualization and Formalization of Patient Problems in Healthcare Enterprises," *Methods of Information in Medicine*, 37:373-383.

Wilcox, A. and Hripcsak, G. (1999). "Classification Algorithms Applied to Narrative Reports," in *Proceedings of the AMIA.Symposium:*455-459.

Xu, H., Anderson, K., Grann, V.R., and Friedman, C. (2004). "Facilitating Cancer Research Using Natural Language in Processing of Pathology Reports," *Medical Informatics* 2004:565-572.

Zou, Q., et al. (2003). "IndexFinder: A Method of Extracting Key Concepts from Clinical Texts For Indexing," in *AMIA Annual Sympoisum Proceedings :*763-767.

Zweigenbaum, P., and et al. (1994). "MENELAS: An Access System for Medical Records Using Natural Language," *Computer Methods and Programs in Biomedicine*, 45:117-120.

SUGGESTED READINGS

Friedman, C. and Johnson S.B., 2005. *Natural Language and Text in Processing in Biomedicine.* In Shortliffe EH and Cimino JJ, editors, Biomedical Informatics, third edition. Chapter 8, New York; Springer (in press).
Provides a methodological overview of natural language processing in the biomedical domain.

Jurafsky, D. and Martin, J. H., 2000. *Speech and Language in Processing: An Introduction to Natural Language in Processing, Computational Linguistics, and Speech Recognition*, Prentice-Hall.
A comprehensive textbook describing general natural language processing techniques.

Friedman, C., ed., 2002. Sublanguage - Zellig Harris Memorial, *J Biomed Inform* **35**:213-277.
A special issue of the Journal of Biomedical Informatics devoted to Zellig Harris's theory on sublanguage processing.

Manning, C. and Sch⎫tze H. 1999. *Foundations of Statistical Natural Language in Processing.* MIT Press.
A textbook describing statistical natural language processing techniques.

Sager, N., C. Friedman, M. Lyman, and et al. 1987. *Medical Language in Processing: Computer Management of Narrative Data.* Reading, MA: Addison-Wesley.
A book describing medical language processing techniques used by the Linguistic String Project that are based on the sublanguage theory of Zellig Harris.

ONLINE RESOURCES

This site contains a summary of the capabilities and sources of a large amount of natural language processing (NLP) software available to the NLP community:
http://registry.dfki.de/

This site contains links to general NLP software tools.
http://www-a2k.is.tokushima-u.ac.jp/member/kita/NLP/nlp_tools.html

The Penn Treebank Project annotates general narrative text for linguistic structure. It produces skeletal parses showing rough syntactic and semantic information, which also includes part of speech tagging:
http://www.cis.upenn.edu/~treebank/home.html

A demonstration of the MedLEE clinical NLP System providing examples of the parsing of discharge summaries, chest radiological reports, and mammograms:
http://lucid.cpmc.columbia.edu/medlee/

A demonstration of the MedLEE clinical NLP System providing examples of the parsing of an electrocardiogram report, a discharge summary, and a radiological report of the abdomen:
http://lucid.cpmc.columbia.edu/dli2/demo/

The Unified Medical Language System is a comprehensive terminological resource for NLP in the clinical domain:
http://www.nlm.nih.gov/research/umls/

The Specialist Lexicon and software tools are linguistic resources for NLP in the biomedical domain:
http://specialist.nlm.nih.gov/

QUESTIONS FOR DISCUSSION

1. How does NLP have the potential to change the practice of medicine?

2. What are some of the hurdles that will need to be overcome?

3. What role does machine learning have in NLP? What are some of the difficulties in using machine learning techniques?

4. Describe some components of an NLP system, and provide some examples of how they would be used.

5. What are some of the potential benefits of using NLP to process clinical reports?

6. Describe some clinical applications that use NLP. What other types of applications would be useful? What impediments are there to integrating

an NLP system with a Clinical Information System?

7. There is typically a trade off between recall (number of true positives that were found by the system/the total number of true positives in the gold standard) and precision (the number of true positives found by the system/the number of all positives found by the system). What are the advantages/disadvantages of aiming for high precision or aiming for high recall in the clinical domain.

8. In what ways do abbreviations cause problems for NLP systems? What types of contextual information can be used to help NLP systems handle abbreviations and how would the information help? Would these suggestions completely solve the problem?

9. Explain the similarities and differences between controlled vocabularies and terms occurring in clinical text.

10. Modifiers change the underlying interpretation of clinical information in text. Provide examples illustrating how modifiers associated with negation, uncertainty, time, and body locations change the meaning of clinical findings.