23

Natural Language Processing in Medicine

Rui Zhang, Yan Wang, and Genevieve B. Melton

CONTENTS

23.1 Introduction

A natural language refers to any language used by people for communication, other than machine or computerized language such as C++ or Java. Natural language processing (NLP), a field of artificial intelligence and computational linguistics, is the automated analysis of natural language. Within biomedicine, the biomedical literature includes a large number of publications written in text format to which NLP techniques are applied. In the clinical domain, there has been a surge of interest in the secondary use of electronic health record (EHR) system data, including electronic clinical notes to improve health care quality through disease surveillance, decision support, and evidence-based medicine. To improve the use of textual information in EHR systems, the development of effective NLP methods for clinical texts is an important and challenging task for effectively using EHR data more reliably.

The application of NLP to process medical literatures and documents has rapidly attracted researchers, especially with the surge of EHR system adoption. However, NLP algorithms require special development for medical tasks because medical sublanguages differ largely from general English across several linguistic dimensions introduced by Harris. For instance, clinical notes are often entered by physicians who have limited time, and therefore, they frequently use domain-specific abbreviations, omit information that can be assumed by context, and have language problems such as misspellings or incorrect word usage. As a result, out-of-the-box existing NLP applications for general English usually did not perform well for medical text. Moreover, domain terminologies and local dialects are prevalent in medical documents. For example, it is not uncommon for physicians at different hospital sites to develop their own local jargon for devices, techniques, or other items. System performance is also challenged in that the outputs of medical NLP systems are frequently used in health care systems or clinical research that requires reliable, high-quality NLP performance and modular, flexible, fast systems. One other major challenge for medical NLP systems are barriers faced from data availability and confidentiality. Many medical NLP systems need to access medical documents from EHR clinic information systems. This can often be problematic because access to patient records is confidential, requires the approval of institutional review boards (IRBs), and may require data de-identification. Also, it is difficult to share data across institutions, which creates another challenge for system interoperability and interinstitutional validation of systems.

23.2 NLP Tasks in Medicine

In the section below, we enumerate some actively researched medical NLP system components from both a fundamental and a higher level, as well as issues that complicate these tasks in the medical domain.

23.2.1 Low-Level NLP Components

23.2.1.1 Tokenization

Tokenization is an initial step of automated processing of a text. It is the task of identifying boundaries that separate semantic units, which include morphemes, words, dates, and

symbols within a text. The primary indication of such semantic units, also called tokens, in general English is white space that occurs before and after a word. A token may also be separated by punctuation marks instead of a word space, such as by a period, comma, semicolon, or question mark. Some of the difficulties that occur with tokenization stem from ambiguous punctuation, such as the colon in "2:30am" or the periods in "M.D." In the medical literature, in addition to typical ambiguous punctuations often seen in general English, the biomedical literature will also have certain technical terms and heterogeneous orthographics, such as "Adams Stokes" and "Adams-Stokes," which add additional difficulties in tokenization (Arens 2004; Barrett and Weber-Jahnke 2011; Jiang and Zhai 2007; Wrenn et al. 2007). For this reason, a simple tokenizer for general English text will typically not work well in biomedical text. Therefore, tokenization algorithms often need new heuristics and domain-specific training corpora to accommodate the distinct features of medical sublanguages.

23.2.1.2 Sentence Boundary Detection

Sentence boundary detection (SBD), also called sentence boundary disambiguation or sentence breaking, can also be a challenging NLP component, particularly for clinical documents. This task aims to detect where sentences start and end. A simple SBD system can identify sentence boundaries using a small set of rules. However, the task can be complicated by the fact that punctuation marks such as question marks, semicolons, and periods are often ambiguous and need more complex logic in special cases. In addition to rule-based systems, AI methods such as decision trees, neural networks, and hidden Markov models (HMMs) are frequently used for SBD. Also, the biomedical literature and clinical documents are full of abbreviations (e.g., "q.i.d.," "p.r.n."), acronyms (e.g., "OD," "OS"), and symbolic constructions (e.g., "blood pressure: 130/67") that add difficulty to SBD (Barrows et al. 2000; Friedman 1997; Huang et al. 2005). For medical NLP systems, one frequent approach for SBD includes the use of domain lexical resources such as the National Library of Medicine's (NLM's) SPECIALIST Lexicon (McCray et al. 1994) and annotated domain corpora to ensure satisfactory SBD performance.

23.2.1.3 Part-of-Speech Tagging

Part-of-speech (POS) tagging is the process for determining the part of speech of words in a piece of text, based on both definition as well as local context. The example below shows the tagging output of the following sentence using the Penn Treebank tag set (Marcus et al. 1994): "The cystic duct was triply clipped distally and singly proximally and transected."

> "The/DT cystic/JJ duct/NN was/VBD triply/RB clipped/VBN distally/RB and/CC singly/RB proximally/RB and/CC transected/VBN."

POS tagging is an essential step of NLP systems where errors can propagate upward to the syntactic processing level and produce more errors in the syntactic output, which provides important information necessary for text understanding. Therefore, having reliable POS information is critical to successful implementation of various NLP applications. POS taggers trained merely on general English do not usually achieve state-of-the-art performance on medical text. A number of POS taggers (Fan et al. 2011; Pakhomov et al. 2006; Smith et al. 2004) have been developed specifically for the medical domain, such as the adapted Trigrams'n'Tags (TnT) tagger (Pakhomov et al. 2006), which is a TnT tagger

(http://www.coli.uni-saarland.de/~thorsten/tnt/) trained on a relatively small set of clinical notes, and the MedPost tagger (Smith et al. 2004), a POS tagger based on an HMM and trained on manually tagged sentences in medical text.

23.2.1.4 Shallow Parsing

Shallow parsing, also called chunking, is the process of identifying constituents (syntactically correlated parts of words like noun groups, verb groups, etc.) in a sentence. As an intermediate step toward deep parsing, shallow parsing produces a limited amount of syntactic information from sentences and does not specify internal structures or roles of each constituent in the main sentence. The sentence below exemplifies shallow parsing output:

> "[NP The cystic duct] [VP was triply clipped] [ADVP distally and singly] [ADVP proximally] and [UCP transected]"

In the medical domain, shallow parsing is used in a wide range of tasks such as drug–drug interaction (DDI) detection, medical problem assertion detection, biological entity relation extraction, and medical information extraction (IE). Several shallow parsers have been built for medical text processing, such as the SPECIALIST minimal commitment parser (McCray et al. 1993), which produces high-level syntactic information rather than the traditional full syntactic information for better noun phrase discovery in medical text.

23.2.1.5 Deep Parsing

Deep parsing is the process to produce an ordered, rooted tree that represents the syntactic structure of a string according to some formal grammar such as constituency grammars (Sipser 1996) and dependency grammars (Mel'Čuk 1988). Figure 23.1 shows the constituency parse tree of the sentence "The dressing was removed from it."

Full syntactic parsing of text can provide a large amount of deep linguistic information such as sentence voice, phrase type, and POS tags, which are shown to perform considerably better than surface-oriented features (e.g., pattern matching) for many NLP tasks. Because of the special features of medical sublanguage (e.g., domain vocabulary, telegraphic text,

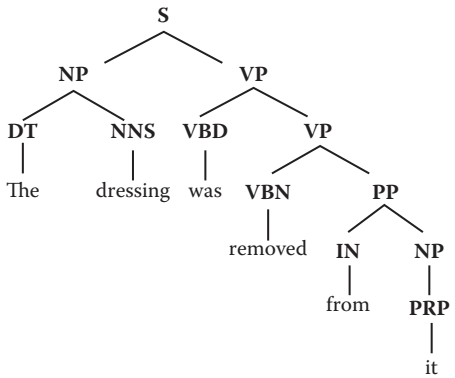

FIGURE 23.1
A constituent (phrase structure) tree for "The dressing was removed from it."

special grammar), parsers trained on general English corpus like the *Wall Street Journal* (http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2000T43) only have limited performance on medical text. NLP experts have investigated several methods to adapt parsers trained on general English to new target domains. New entries can be imported from domain resources to existing parser lexicons using morphological clues, heuristic mapping, and direct expansion (Szolovits 2003). POS tag information of domain-specific lexical elements can also be provided to a parser to avoid inconsistencies between domain POS tags and parser lexicon POS tags (Rimell and Clark 2009). Moreover, better parsing performance can also be acquired by adjusting the syntactical category statistics for important domain lexical elements like verbs and other lexical elements that have unusual usage in a particular domain (Huang et al. 2005).

23.2.2 High-Level NLP Components

23.2.2.1 Negation Detection

Many medical documents such as discharge summaries and radiology reports contain large amounts of important information of patients, like conditions, findings, and diseases, that can be used for a wide range of secondary applications. In these reports, about half of the described findings and diseases are actually absent in a given patient. For example: "They have not noticed any abnormal behaviors, movements, or rash anywhere else on his body" or "no significant complications of bleeding." As a result, negation detection is a critical component in medical NLP systems. In medical text, negation detection is not an easy task as negation can be explicit (e.g., "Patient denies any fevers, emesis, or diarrhea") or implied (e.g., "Chest x-ray is clear upon my read").

The scope of negation is another challenge for the negation detection process. Consider two sentences: "The child is not tired" and "The child is not very tired." In the first sentence, the word "not" scopes over "tired," while in the second sentence, the word "very" redirects the scope of "not" to itself and away from "tired."

Negation detection systems can detect negation solely by rules developed on a handcrafted list of negation phrases that appear before or after a term of interest or through the use of an ontology of negative medical concepts generated from a standard medical dictionary, such as the Systematized Nomenclature of Medicine—Clinical Terms (SNOMED CT), to decide if a term is negated (Chapman et al. 2001; Elkin et al. 2005; Mutalik et al. 2001). In more complex systems, machine learning techniques such as naïve Bayes and decision trees are applied with more clues, such as deeper syntactic structure and lexical cues (Goryachev et al. 2006; Sarafraz and Nenadic 2010; Yang et al. 2009).

23.2.2.2 Relationship Extraction

Relation extraction aims to determine or discover relationships between entities (e.g., drugs, diseases, findings, genes) in medical texts. Relations among these entities, in their simplest form, are binary, involving only two entities. Other relations can involve more than two entities. A large variety of relations have been investigated, such as interactions between drugs, genes, associations between diseases and symptoms, and relations between patient problems and treatments. Co-occurrence statistics are effective methods that are frequently used for identifying relations between medical entities by collecting instances where the entities co-occur (Cao et al. 2005; Chen et al. 2008; Wang et al. 2009). The hypothesis behind this approach is that an entity and its related entities are more

likely to appear together than random combinations of entities. Thus, if entities are repeatedly mentioned together, then there is a good chance that they may be related. However, the nature of the relationship between these associated entities usually cannot be determined by the method alone.

Rule-based approaches for relation extraction work by exploiting the particular linguistic patterns exhibited by relations. Rules used can be manually defined by domain experts or derived from annotated corpora. Machine learning-based systems rely on machine learning techniques along with a variety of features based on the nature of the relationship, such as lexical, syntactic, semantic, and dependency features (Barnickel et al. 2009; Katrenko and Adriaans 2007).

Several important challenges are associated with relation extraction in the medical domain. First, in the medical domain, annotation of relations can be complicated because relations are often expressed across discontinuous spans of text. Secondly, there can be a lack of consensus on how to best annotate a particular type of relation. As a result, annotation resources between research groups can be largely incompatible and the quality of systems constructed based upon these resources can be difficult to evaluate.

23.2.2.3 Named Entity Recognition

Named entity recognition (NER) aims to identify and classify elements into named entities, which are predefined categories such as names (e.g., drugs, genes, person), findings, diseases, and medications. The biomedical literature is full of terms particular to the biomedical domain that are typically not detected by conventional general English NLP systems. In order to extract relations between entities, it is crucial for the system to be able to detect unknown nouns or named entities. Some named entities can be effectively identified solely through surface patterns (e.g., phone number: xxx-xxx-xxxx, person: Carole Green MD). However, rule-based systems require a significant manual effort, and these rules may not easily extend to a new domain. Machine learning is another effective approach for NER. Compared with rule-based systems, it requires less human intuition with rules, and this approach can easily be adapted to new domains. The disadvantage of machine learning approaches is the large annotation corpus required for model training.

23.2.2.4 Word Sense Disambiguation

Ambiguity is a problem inherent to natural language, where a term can have more than one meaning depending upon the context or use of the term in a particular text. Word sense disambiguation (WSD) is the process of understanding which sense of a term, including single words, abbreviations, or acronyms, is being used in a particular context among a list of predefined sense candidates. In the medical domain, researchers have suggest that the problem might be more restricted compared to general English based upon the idea that since medicine is scientific, it might be more specific than general English. Instead, the problem is more extensive in the medical domain due to the high use of abbreviations and acronyms in medical documents and the biomedical literature.

Approaches to WSD (McInnes et al. 2011; Pakhomov et al. 2005; Stevenson et al. 2012; Xu et al. 2007) generally rely on a particular domain knowledge source, such as the Unified Medical Language System (UMLS) (Humphreys et al. 1998), MEDLINE (http://www.nlm.nih.gov/bsd/pmresources.html), and Entrez Gene Database (http://jura.wi.mit.edu/entrez_gene/), and domain corpora such as GENIA (Kim et al. 2003) or the Biosemantics test collection (Weeber et al. 2001) for sense collection, sample collection, and model training.

23.2.2.5 Semantic Role Labeling

Semantic role labeling (SRL) is the task of detecting semantic roles associated with predicates, which are mainly verbs, such as "hit" and "move," in a sentence. For example, in the sentence "He placed the ball beside the couch," the predicate is the verb "place." The semantic roles associated with "place" include "placer"—who placed; "thing placed"—what is placed; and "location"—where it (the ball) is placed. Labeling semantic roles like above for predicates in a given text answers questions such as "who," "when," "what," "where," and "why." SRL can be used for IE, question answering (QA), text summarization, and other NLP tasks that require some kind of semantic interpretation.

Example semantic roles include agent, patient, instrument, and adjunctive arguments indicating other meanings such as locative and temporal. In general, SRL can be accomplished using supervised machine learning approaches. Given a predicate and each constituent in a syntactic parsed output, an SRL system assigns a semantic role from a predefined set of roles for the predicate. A typical supervised SRL system can be designed by extracting machine learning features for each constituent, training a machine learning classifier on an annotated, training set and then labeling each unlabeled constituent in a new set of text with a given set of features. In order to build an SRL system to process medical text, domain resources such as semantic annotated corpus and semantic frames (a semantic frame is a collection of semantic roles of a predicate) often must be created (Dahlmeier and Ng 2010; Kogan et al. 2005). Also, domain-specific features often boost the SRL performance in medical domain (Tsai et al. 2006).

23.2.2.6 IE

IE is a task that involves extracting problem-specific information from the text of interest and then transforming this information into structured form. For example, vaccination reactions can be extracted from medical reports, and relationships between genes and diseases from the biomedical literature are all cases of IE. Most early and straightforward IE systems were built mostly using pattern matching techniques such as regular expressions over features such as text strings, syntactic structure, semantic type, and dictionary entries.

Recent systems are mostly based on machine learning methods. State-of-the-art lower-level components and high-level components introduced before such as deep parsing, NER, and WSD are often part of an IE system. In the medical domain, a variety of IE systems have been built for various tasks (Dang et al. 2008; Denecke and Bernauer 2007; Hripcsak et al. 1998; Lakhani and Langlotz 2010; Long 2005) as well as many NLP tools for IE, such the Medical Language Extraction and Encoding System (MedLEE) (Hripcsak et al. 1998) and the clinical Text Analysis and Knowledge Extraction System (cTAKES) (Savova et al. 2010).

23.3 NLP Methods

NLP methods include symbolic (linguistics-based) methods, statistics-focused methods, and machine learning methods. Symbolic methods are built based on linguistic rules, while statistical methods and machine learning methods require training to build models. In this section, we cover a few of these methods briefly.

23.3.1 Support Vector Machine

A support vector machine (SVM) is a discriminative machine learning approach that belongs to the family of margin-based classifiers. It works by looking for an optimal hyperplane that maximizes the distance between the hyperplane and the nearest samples from each of the two classes. In the simplest two-feature case, a straight line would separate samples in an X–Y plot. In a general *N*-feature case, the separator will be a hyperplane with *N*–1 dimensions. For an *N*-feature case, the input may be transformed mathematically using a kernel function (e.g., Gaussian), to allow linear separation of the data points. During the learning process, the separation process selects a subset of the training data, the "support vectors," that best differentiates the classes. The resulting hyperplane maximizes the distance between each class and the support vectors, as shown in Figure 23.2.

As a binary classification approach, SVM is only directly applicable for two-class tasks. However, it can be easily extended to multiclass classification problems by the one-versus-all (OVA) approach or pairwise approach. The prediction accuracy of SVM is generally high because of the sound mathematical theory behind it and the robustness of the method. It generally works well when training examples contain errors, as well, because of its use of a separation process. On the downside, SVM is computationally expensive. Its training process is a convex optimization problem that requires at least quadratic time with respect to the number of training examples. In the medical domain, SVM has been shown to perform well on many classification tasks such as smoking status classification (Cohen 2008), SRL for protein transport predicates (Bethard et al. 2008), and disease comorbidity status classification (Ambert and Cohen 2009).

23.3.2 Maximum Entropy Modeling

The principal of maximum entropy modeling (MEM) is simple and realistic. It is a statistical learning method that models all that is known and assumes nothing about that which is unknown. The modeling contains a set of predefined features or constraints as shown below.

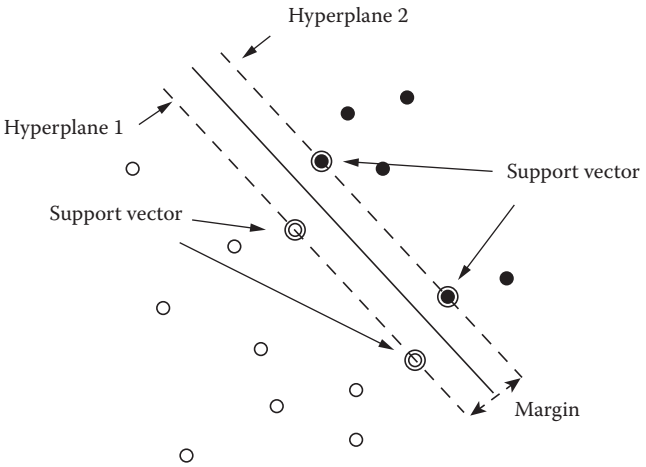

FIGURE 23.2
Linear separating hyperplanes.

$$\sum_{x,y} \tilde{p}(x)p(y|x)f(x,y) = \sum_{x,y} \tilde{p}(x,y)f(x,y)$$

Here, x is a random variable representing some context information, y is the output, and $f(x, y)$ is the feature function, as shown below.

$$f(x,y) = \begin{cases} 1 & \text{if } x \text{ denotes a context and } y \text{ indicates an output} \\ 0 & \text{otherwise} \end{cases}$$

$\tilde{p}(x,y)$ is the joint empirical distribution that is derived from the training data expressing some relationship between features and outcome, as shown below.

$$\tilde{p}(x,y) = \frac{1}{N} \times \text{number of } (x,y)$$

$p(x, y)$ is the conditional probability model for predicting the output y given a context x. Among many conditional probability models, the best model p^* is the one that maximizing the conditional entropy $H(p)$, which is shown below, on $p(x, y)$, as it has a more uniform probability distribution on unseen x in the training set, consequently allowing less bias for unseen contexts.

$$H(p) = -\sum_{x,y} \tilde{p}(x)p(y|x)\log p(y|x)$$

For details of how MEM works and why it works for NLP, readers may refer to an article by Berger (1996).

One advantage of MEM is that heterogeneous information sources such as lexical, syntactical information and bigrams can be modeled easily as features in an integrated model. Another advantage of MEM is that it handles overlapping features very well. It is sometimes more effective to use a combined feature together with its component features, compared with using simple features alone. Because of its ability to incorporate heterogeneous features, MEM has been used in the medical domain for a diverse set of NLP tasks, such as patient medication status mining (Pakhomov et al. 2002), SRL for biomedical verbs (Tsai et al. 2006), and noun phrase identification in radiology reports (Huang et al. 2005).

23.3.3 *n*-Gram Model

Statistical language modeling (SLM) is widely used for many NLP tasks, such as POS tagging, parsing, information retrieval, and machine translation. Generally speaking, SLM assigns a probability to a set of n words based on a probability distribution.

An n-gram model is a typical language method used in the field of computational linguistics and NLP (Manning and Schütze 2003). The word "n-gram" means consecutive items, such as words or terms. n-gram models (n = 2, 3, 4 …) are used to estimate the probability of the existence of the n-gram. To simplify the calculation of the probability of the word, the Markov assumption states that the probability of the word is only based on the

prior few words instead of all previous words. The probability of a word is then simplified as follows:

$$P(w_1^n) \approx \prod_{k=1}^{n} P(w_k \mid w_{k-1})$$

An *n*-gram model (which checks the *n*-1 previous words) is an (*n*-1)th order Markov model.

Probability of a given word can be estimated by its relative frequency. One commonly used estimate is called maximum likelihood estimate (MLE).

$$P_{MLE}(w_1 \ldots w_n) = \frac{C(w_1 \ldots w_n)}{N}$$

$$P_{MLE}(w_n \mid w_1 \ldots w_{n-1}) = \frac{C(w_1 \ldots w_n)}{C(w_1 \ldots w_{n-1})}$$

where N is the number of training instances and C is the count.

In practice, MLE is not an ideal estimator due to sparseness, which occurs in many data sets. MLE assigns a value of zero to the probability of unseen events (in the training set), which will propagate to the whole sentence. To avoid issues related to the sparseness of data sets, which always exist if the data set is large, discounting methods are commonly used, such as the Good–Turing (GT) method and that by Ney and Essen.

GT discounting is based on the assumption that the probability of items follows a binomial distribution. It is suitable for a large data set.

$$\text{If } C(w_1 \ldots w_n) = r > 0,\ P_{GT}(w_1 \ldots w_n) = \frac{r^*}{N}$$

where

$$r^* = \frac{(r+1)S(r+1)}{S(r)}$$

$$\text{If } C(w_1 \ldots w_n) = 0,\ P_{GT}(w_1 \ldots w_n) = \frac{1 - \sum_{r=1}^{\infty} N_r \frac{r^*}{N}}{N_0} \approx \frac{N_1}{N_0 N}$$

where s is the function that fits the observed values of (r, Nr), and $S(r)$ is the expectation of the frequency.

Ney and Essen proposed two discounting models for estimating frequencies of *n*-grams. One is absolute discounting:

$$\text{If } C(w_1 \ldots w_n) = r,\ P_{abs}(w_1 \ldots w_n) = \begin{cases} (r - \delta/N) & \text{if } r > 0 \\ (B - N_0\delta/N_0 N) & \text{otherwise} \end{cases}$$

where δ is a small constant number for all nonzero MLE frequencies and *B* is the number of target feature values.

Another is linear discounting:

$$\text{If } C(w_1 \ldots w_n) = r,\ P_{abs}(w_1 \ldots w_n) = \begin{cases} (1-\alpha)r/N & \text{if } r > 0 \\ \alpha/N_0 & \text{otherwise} \end{cases}$$

where α is a constant slightly less than 1.

These discounting methods make the probability of unseen events a small number instead of zero and rescale the other probabilities. The absolute discounting approach is very successful, while the linear discounting is less so since it does not approximate even higher frequencies.

n-Grams have been used for low-level NLP tasks such as spelling correction, speech recognition, and word disambiguation. They have been used in a longitudinal analysis of clinical notes to identify new information (Zhang et al. 2012). *n*-Grams have also been used to improve the efficiency of systematic reviews (SRs) for evidence-based medicine (Munshaw and Kepler 2010).

23.3.4 HMM

HMM (Figure 23.3) is another statistical NLP method. Markov models are built on the Markov assumption that the current state occurs based upon on the previous state(s). For the simplest first-order Markov model, there are M^2 transitions between *M* states. Unlike deterministic models, where each state is dependent on another state, Markov models assign probability to each transition between two states. In a visible Markov model, the state is visible, and state transition probabilities are the only parameters to calculate. In an HMM, hidden states have a probability contribution to the outputs. For example, in a speech recognition system, the sound we hear is the output of hidden states, such as vocal chords, the size of the person's throat, the position of the person's tongue, and many other factors. Each sound of a word is generated from changes of these hidden factors.

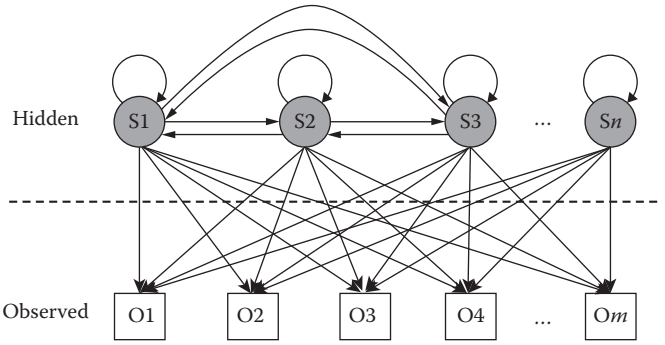

FIGURE 23.3
Hidden Markov models. S1, S2, S3…S*n* are hidden states; O1, O2, O3, O4…O*m* are outputs. Each state can transit to other states or itself, shown in the lines between states. Transitions between the state S*n* and other states are not shown. Each observed output is generated from hidden states with probabilities, indicated as darker lines.

HMMs are defined by the following quintuple:

$$\lambda = (N, M, A, B, \pi)$$

where N is the number of states for the model; M is the number of distinct observation symbols per state; A is the $N \times N$ state transition probability distribution given in the form of a matrix $A = \{a_{ij}\}$; B is the $N \times M$ observation symbol probability distribution given in the form of a matrix $B = \{b_j(k)\}$; and π is the initial state distribution vector $\pi = \{\pi_i\}$.

Three canonical problems are associated with HMM:

1. Evaluation: computing the probability of a particular output sequence based on the given model parameters. This is typically implemented using Viterbi–forward or forward algorithms (Viterbi 1967).
2. Decoding: finding the state candidates that can generate a particular output sequence based on the given model parameters, which is typically computed by using the Viterbi algorithm.
3. Learning: finding the set of state transition and output probabilities that fit the given output sequence(s) the best.

HMM has been widely used in speech recognition and bioinformatics (Drawid et al. 2009; Munshaw and Kepler 2010). It has been used in medicine to describe the effect of alcoholism treatment on the likelihood of healthy/unhealthy populations (Wall and Li 2009), to estimate the transition probabilities between states of liver cirrhosis (Bartolomeo et al. 2011), and for disease surveillance with public health data (Watkins et al. 2009).

23.4 Clinical NLP Resources and Tools

In this section, we mainly focus on several NLP resources and tools available in the biomedical domain. Specifically, we introduce the resources and tools supported by the US NLM at the National Institutes of Health (NIH).

23.4.1 UMLS

UMLS (http://www.nlm.nih.gov/research/umls/) was developed by and is maintained by the NLM to provide health care professionals and researchers with a biomedical domain knowledge resource (Humphreys et al. 1998). UMLS is a structured knowledge base that connects different biomedical sources and enables biomedical research application development. UMLS contains three knowledge sources: Metathesaurus, Semantic Network (McCray 2003), and SPECIALIST Lexicon (McCray et al. 1994) and lexical tools.

Metathesaurus is created based on over 100 vocabularies, code sets, and thesauri. It covers several major categories, including comprehensive vocabularies [e.g., SNOMED CT, http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html, Medical Subject Headings (MeSH, http://www.nlm.nih.gov/mesh/)]; laboratory and observational data [e.g., Logical Observation Identifier Names and Codes (LOINC, http://loinc.org/) (Forrey et al. 1996; McDonald et al. 2003)]; diseases [e.g., International Classification of Diseases

and Related Health Problems (ICD, http://www.who.int/classifications/icd/en/)]; and procedures and supplies [e.g., Current Procedural Terminology (CPT, http://www.ama-assn.org/ama/pub/physician-resources/solutions-managing-your-practice/coding-billing-insurance/cpt.page)]. While 60% of the Metathesaurus terms are in English, it also contains terms in 17 other languages, including Spanish, French, Dutch, Italian, Japanese, and Portuguese. Each equivalent biomedical term (same meaning, various names) from different sources is assigned a concept unique identifier (CUI).

The Semantic Network is an upper-level ontology of biomedical knowledge. It contains 135 semantic types and 54 relationships between semantic types. Each concept is assigned at least 1 of 135 defined semantic types. All semantic types are hierarchically organized under two main topics: Entity and Event. The top level of semantic types in the UMLS Semantic Network is depicted in Figure 23.4.

The relationship "ISA" is used to link most concepts. For example, "Carbohydrate" ISA "Chemical." There are also five major, nonhierarchical relationships: physical (e.g., PART_OF, BRANCH_OF); spatial (e.g., LOCATION_OF, ADJACENT_TO); temporal (e.g., CO-OCCURS_TO, PRECEDES); functional (e.g., TREATS, CAUSES); and conceptual (e.g., EVALUATION_OF, DIAGNOSES).

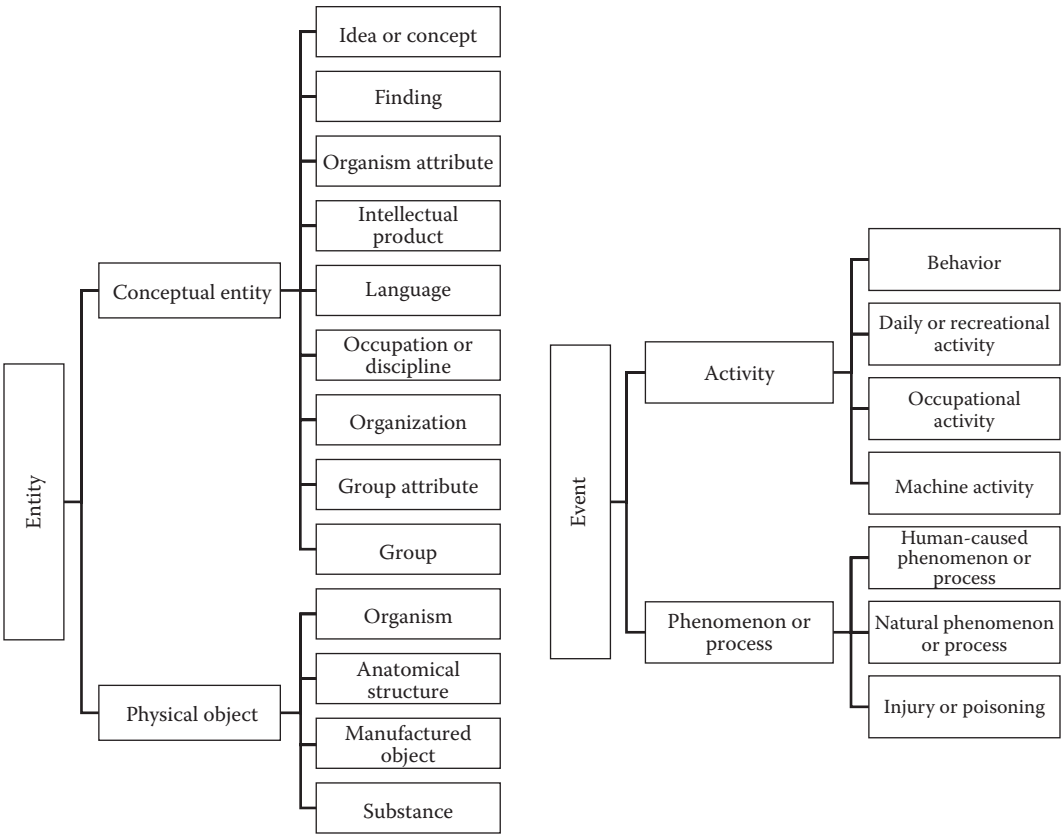

FIGURE 23.4
Hierarchy structure of UMLS semantic types.

The SPECIALIST Lexicon is an English dictionary including over 200,000 biomedical terms as well as common English words. It also contains syntactic, morphological, and orthographic information of each term or word. For example, each lexical record contains base forms of the term; the part of speech; a unified identifier; spelling variants; and inflection for nouns, verbs, and adjectives.

23.4.2 Corpora

Development of NLP systems requires large volumes of biomedical and clinical texts. The MEDLINE database is a collection of biomedical abstracts. It is maintained by the NLM and contains over 21 million reference from 1946 to the present. The GENIA corpus (http://www.nactem.ac.uk/genia/genia-corpus) collects 1999 MEDLINE abstracts, selected from a PubMed query for MeSH terms "human," "blood cells," and "transcription factors." The corpus has been annotated with various levels of linguistic and semantic information covering POS, syntactic, term, event, relation, and coreference annotation (Kim et al. 2003). In the clinical domain, there are a few collections of clinical texts, including the Pittsburgh collection of clinical reports (http://www.dbmi.pitt.edu/nlpfront), Multiparameter Intelligent Monitoring in Intensive Care (MIMIC II) database (Saeed et al. 2011), and Informatics for Integrating Biology & the Bedside (i2b2) NLP research data sets (https://http://www.i2b2.org/NLP/DataSets/Main.php). Most research groups created their own clinical text corpus and annotations for specific NLP tasks locally.

23.4.3 SPECIALIST NLP Tools

SPECIALIST NLP tools (http://lexsrv3.nlm.nih.gov/Specialist/Home/index.html) are computer programs developed by the NLM to aid in dealing with different biomedical NLP tasks. Tools include lexical tools such as lexical variant generator (LVG), normalized string generator (Norm), word index generator (WordInd), dTagger POS tagger, subterm mapping tools (STMTs), and others.

LVG contains a series of commands to perform lexical transformation of text. Norm provides a normalization process for those terms included in the SPECIALIST Lexicon. Norm can help to find similar terms and to map terms to UMLS concepts. WordInd creates a sequence of alphanumeric characters by reading the text, helping UMLS to produce the word index for the Metathesaurus. dTagger is a POS tagger built specifically with SPCIALIST Lexicon. dTagger was trained on MedPost corpus, a set of annotated MEDLINE abstracts, and tokenizes text into multiword terms. STMT was built to provide comprehensive subterm-related features, including all subterms, the longest prefix subterm, and synonymous subterm substitutions.

23.4.4 MetaMap

MetaMap (http://metamap.nlm.nih.gov/) is a program developed by the NLM to map biomedical text to the UMLS Metathesaurus (Aronson 2001; Aronson and Lang 2010). MetaMap provides various options, including data option (choose specific vocabularies and data model); processing options (such as author-defined acronyms/abbreviations, negation detection, WSD) and output options (human readable, machine output, and XML). Released application programming interfaces (APIs) provide options to integrate MetaMap into other programs. MetaMap was originally developed for information retrieval from bibliographic data such as MEDLINE citations. As it is an effective tool

to map biomedical terms, MetaMap has been widely used in applications of the clinical domain, such as detection of clinical findings.

23.4.5 SemRep

SemRep is a rule-based, symbolic NLP program developed by NLM for semantic knowledge representation from biomedical literatures, mainly from titles and abstracts in MEDLINE (Fiszman et al. 2003; Rindflesch and Aronson 1993; Rindflesch and Fiszman 2003; Srinivasan and Rindflesch 2002). SemRep uses underspecified syntactic analysis and structured domain knowledge from UMLS. SemRep relies on syntactic analysis based on the SPECIALIST Lexicon and the MedPost POS tagger (Smith et al. 2004). MetaMap helps to map noun phrases in the sentences to UMLS Metathesaurus concepts. SemRep interpreted the semantic relationships (syntactic indicators in the sentence, such as verbs, nominalizations, prepositions, etc.) between two concepts in the sentences based on dependency grammar rules and ontology (i.e., an extended version of the UMLS Semantic Network). SemRep represents semantic knowledge from each sentence in citations as the format of semantic predications (a subject–predicate–object triplet). Both subjects and objects are Metathesaurus concepts and predicates that correspond to a relation type in SemRep ontology. For example, SemRep interprets sentence 1 as semantic predications in sentence 2.

1. We used hemofiltration to treat a patient with digoxin overdose that was complicated by refractory hyperkalemia
2. Hemofiltration-TREATS-Patients
   Digoxin overdose-PROCESS_OF-Patients
   Hyperkalemia-COMPLICATES-Digoxin overdose
   Hemofiltration-TREATS-Digoxin overdose

23.5 Current Clinical NLP Systems

In this section, we introduce several existing clinical NLP systems. We summarize some current clinical NLP systems in Table 23.1 and then discuss a few of them.

23.5.1 MedLEE

Medical language extraction and encoding system (MedLEE) is an NLP system that extracts information from clinical texts into a structured format and translates the information to terms in a controlled dictionary. MedLEE has been used to process various types of clinical records, including radiology reports, discharge summaries, sign-out notes, pathology reports, electrocardiogram reports, and echocardiogram reports (Cao et al. 2004; Chen et al. 2008; Chun et al. 2005; Friedman et al. 1994, 2004; Xu et al. 2004). The MedLEE preprocessor first transforms reports into a structure for the core NLP engine to process, for example, adding or changing section headers and expanding abbreviations (e.g., "hx" to "history"). The core MedLEE engine maps medical terms to semantic types and uses grammar rules to extract their semantic relationships. A structured output in

TABLE 23.1

Current Clinical NLP Systems

System	Description	Institution (Principle Investigator)	References
BioMedICUS[a]	A UIMA pipeline system designed for researchers for extracting and summarizing information from unstructured text of clinical reports	University of Minnesota (Pakhomov)	http://code.google.com/p/biomedicus/
cTAKES[a]	A UIMA pipeline built around OpenNLP, Lucene, and LVG for extracting disorder, drug, anatomical site, and procedure information from clinical notes	Mayo Clinic (Chute)	Savova et al. 2010
HITEx[a]	An NLP system distributed through i2b2	Harvard (Zeng)	Goryachev et al. 2006
MedEx[a]	A semantic-based medication extraction system designed to extract medication names and prescription information	Vanderbilt (Xu)	Xu et al. 2010 Doan et al. 2010
MedLEE	An expert-based NLP system for unlocking clinical information from narratives	Columbia (Friedman)	Friedman and Hripcsak 1998 Friedman 2000
MedTagger[a]	A machine learning–based name entity detection system utilizing existing terminologies	Mayo Clinic (Liu)	Torii et al. 2011
MetaMap[a]	An expert-based system for mapping text to the UMLS	NLM (Aronson)	Aronson and Lang 2010
SecTag[a]	A system to tag clinical note section headers	Vanderbilt (Denny)	Denny et al. 2009 Denny et al. 2008

Note: Systems are listed alphabetically.

[a] Publicly available systems.

XML format is then generated for each sentence. The data are finally transformed and stored in a clinical repository.

23.5.2 cTAKES

Clinical text analysis and knowledge extraction system (cTAKES) is an NLP system developed at Mayo Clinic for IE (specifically disorders, drugs, anatomical sites, and procedures) from free texts in clinical notes (Savova et al. 2010). cTAKES was built on a pipeline framework called the Unstructured Information Management Architecture (UIMA, IBM), which allows components in the system to be implemented sequentially. UIMA enables NLP systems to be decomposed into components, each of which is responsible for different tasks in analyzing the unstructured information. In cTAKES, components include basic NLP tasks such as sentence boundary detector, tokenizer, morphologic normalizer, part-of-speech tagger, dependency parser, NER annotator, and negation detector. It also contains clinical-specific tasks including the patient's smoking status identifier and drug mention annotator.

23.5.3 HITEx

Health Information Text Extraction (HITEx) is an open-source NLP system developed at the National Center for Biomedical Computing, i2b2 (Goryachev et al. 2006). HITEx was built on General Architecture for Text Engineering (GATE) framework and assembles GATE pipeline application and standard NLP components (such as POS tagger, parser). Each pipeline was developed to extract different clinical information, including diagnoses,

discharge medications, smoking status, negation finding, and so forth. For example, to find principal diagnoses, the pipeline searches UMLS concepts in specific note sections and filters semantic types of the concepts that are either findings or symptoms (Zeng et al. 2006). To assign various diagnoses to the correct patient family member, from discharge summary to outpatient notes, HITEx mapped the family member concept and eight diagnosis semantic types from notes and associated diagnosis with the most relevant family member by using a set of rules (Goryachev et al. 2008).

23.6 Medical Applications of NLP

In a time-constrained clinical practice environment, clinicians may have limited time to review and synthesize all clinical notes. Thus, successful processing of large volumes of clinical narratives is the key component for improving health care. In this section, we will provide some research studies to discuss the role of NLP in health care.

23.6.1 NLP for Surveillance

Surveillance is a fundamental and important task in health care, especially surveillance of adverse events (AEs) based on the clinical texts. Hripcsak et al. (2003) developed a framework to discover AEs from clinical notes. They used MedLEE to parse the clinical narratives and generate a coded database, followed by query generation to detect and classify events. Penz and colleagues (2007) also used MedLEE to detect AEs related to the placement of central venous catheters (CVCs) from clinical texts in the Veterans Administration database. They used both an NLP program and a phrase-matching algorithm to achieve a sensitivity of 72% and specificity of 80.1%. Melton and Hripcsak (2005) constructed an AE detection system using MedLEE to identify 45 AE types from discharge summaries and obtained better results than traditional methods, with sensitivity of 28% and specificity of 98.5%. Recently, Friedman et al. reported adverse drug reactions (ADRs) using EHR and an automated method that combines MedLEE with an expert-generated disease identifier. They applied the method to identify two serious ADRs from almost 0.2 million records and reached a good performance (sensitivity of 93.8% and specificity of 91.8%) (Haerian et al. 2012).

23.6.2 NLP for Clinical Decision Support

An NLP system can transfer clinical texts to encoded information, which meets the needs of clinical decision support (CDS) (Demner-Fushman et al. 2009). For example, NLP systems can help to find patients who match certain criteria based on the information extracted from clinical texts. Jain et al. (1996) used MedLEE to encode the information in chest radiograph and mammogram reports and identified patients at risk of having tuberculosis (TB). Fiszman et al. (2000) found that an NLP system for automatic detection of acute bacterial pneumonia from chest x-ray reports performed similarly to physicians and better than lay persons and keyword searching. Day et al. (2007) have developed a daily program using the MPLUS NLP system and decision support technologies to automatically identify trauma patients. Compared with results with clinicians' judgments, the system performed well, with sensitivity of 71% and specificity of 99%.

References

Ambert, K. H., and Cohen, A. M. (2009). A system for classifying disease comorbidity status from medical discharge summaries using automated hotspot and negated concept detection. *J Am Med Inform Assoc*, 16(4), 590–595.

Arens, R. (2004). A preliminary look into the use of named entity information for bioscience text tokenization. In *Proceedings of the Student Research Workshop at HLT-NAACL*, 37–42.

Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *AMIA Annual Symposium Proceedings*, Washington, DC, 17–21.

Aronson, A. R., and Lang, F. M. (2010). An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc*, 17(3), 229–236.

Barnickel, T., Weston, J., Collobert, R., Mewes, H.-W., and Stümpflen, V. (2009). Large scale application of neural network based semantic role labeling for automated relation extraction from biomedical texts. *PLoS One* 2009, 4(7), e6393.

Barrett, N., and Weber-Jahnke, J. (2011). Building a biomedical tokenizer using the token lattice design pattern and the adapted Viterbi algorithm. *BMC Bioinformatics*, 12(Suppl 3), S1.

Barrows, R. C., Busuioc, M., and Friedman, C. (2000). Limited parsing of notational text visit notes: ad-hoc vs. NLP approaches. In *AMIA Annual Symposium Proceedings*, Los Angeles, 51–55.

Bartolomeo, N., Trerotoli, P., and Serio, G. (2011). Progression of liver cirrhosis to HCC: an application of hidden Markov model. *BMC Med Res Methodol*, 11, 38.

Berger, A. L., Pietra, V. J. D., and Pietra, S. A. D. (1996). A maximum entropy approach to natural language processing. *Comput Linguist*, 22(1), 39–71.

Bethard, S., Lu, Z., Martin, J., and Hunter, L. (2008). Semantic role labeling for protein transport predicates. *BMC Bioinformatics*, 9(1), 277.

Cao, H., Chiang, M. F., Cimino, J. J., Friedman, C., and Hripcsak, G. (2004). Automatic summarization of patient discharge summaries to create problem lists using medical language processing. *Stud Health Technol Inform*, 107, 1540.

Cao, H., Markatou, M., Melton, G. B., Chiang, M. F., and Hripcsak, G. (2005). Mining a clinical data warehouse to discover disease-finding associations using co-occurrence statistics. In *AMIA Annual Symposium Proceedings*, Washington, DC, 106–110.

Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F., and Buchanan, B. G. (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform*, 34(5), 301–310.

Chen, E. S., Hripcsak, G., Xu, H., Markatou, M., and Friedman, C. (2008). Automated acquisition of disease drug knowledge from biomedical and clinical documents: an initial study. *J Am Med Inform Assoc*, 15(1), 87–98.

Chen, H., Fuller, S. S., Friedman, C., and Hersh, W. (2005). *Medical Informatics: Knowledge Management and Data Mining in Biomedicine*. Springer, New York.

Cohen, A. (2008). Five-way smoking status classification using text hot-spot identification and error-correcting output codes. *J Am Med Inform Assoc*, 15(1), 32–35.

Dahlmeier, D., and Ng, H. T. (2010). Domain adaptation for semantic role labeling in the biomedical domain. *Bioinformatics*, 26(8), 1098–1104.

Dang, P. A., Kalra, M. K., Blake, M. A., Schultz, T. J., Halpern, E. F., and Dreyer, K. J. (2008). Extraction of recommendation features in radiology with natural language processing: exploratory study. *Am J Roentgenol*, 191(2), 313–320.

Day, S., Christensen, L. M., Dalto, J., and Haug, P. (2007). Identification of trauma patients at a level 1 trauma center utilizing natural language processing. *J Trauma Nurs*, 14(2), 79–83.

Demner-Fushman, D., Chapman, W. W., and McDonald, C. J. (2009). What can natural language processing do for clinical decision support. *J Biomed Inform*, 42(5), 760–772.

Denecke, K., and Bernauer, J. (2007). Extracting specific medical data using semantic structure. *Artif Intel Med*, 4594, 257–264.

Denny, J. C., Miller, R. A., Johnson, K. B., and Spickard, A. 3rd. (2008). Development and evaluation of a clinical note section header terminology. In *AMIA Annual Symposium Proceedings*, Washington, DC, 156–160.

Denny, J. C., Spickard, A. 3rd, Johnson, K. B., Peterson, N. B., Peterson, J. F., and Miller, R. A. (2009). Evaluation of a method to identify and categorize section headers in clinical documents. *J Am Med Inform Assoc*, 16(6), 806–815.

Doan, S., Bastarache, L., Kilmkowski, S., Denny, J. C., and Xu, H. (2010) Integrating existing natural language processing tools for medication extraction from discharge summaries. *J Am Med Inform Assoc*, 17, 528–531.

Drawid, A., Gupta, N., Nagaraj, V. H., Gelinas, C., and Sengupta, A. M. (2009). OHMM: a Hidden Markov Model accurately predicting the occupancy of a transcription factor with a self-overlapping binding motif. *BMC Bioinformatics*, 10, 208.

Elkin, P. L., Brown, S. H., Bauer, B. A., Husser, C. S., Carruth, W., Bergstrom, L. R., and Wahner-Roedler, D. L. (2005). A controlled trial of automated classification of negation from clinical notes. *BMC Med Inform Decis Mak*, 5, 13.

Fan, J.-W., Prasad, R., Yabut, R. M., Loomis, R. M., Zisook, D. S., Mattison, J. E., and Huang, Y. (2011). Part-of-speech tagging for clinical text: wall or bridge between institutions. In *AMIA Annual Symposium Proceedings*, Washington, DC, 2011, 382–391.

Fiszman, M., Chapman, W. W., Aronsky, D., Evans, R. S., and Haug, P. J. (2000). Automatic detection of acute bacterial pneumonia from chest x-ray reports. *J Am Med Inform Assoc*, 7(6), 593–604.

Fiszman, M., Rindflesch, T. C., and Kilicoglu, H. (2003). Integrating a hypernymic proposition interpreter into a semantic processor for biomedical texts. In *AMIA Annual Symposium Proceedings*, Washington, DC, 239–243.

Forrey, A. W., McDonald, C. J., DeMoor, G., Huff, S. M., Leavelle, D., Leland, D., Fiers, T., Charles, L., Griffin, B., Stalling, F., Tullis, A., Hutchins, K., and Baenziger, J. (1996). Logical Observation Identifier Names and Codes (LOINC) database: a public use set of codes and names for electronic reporting of clinical laboratory test results. *Clin Chem*, 42(1), 81–90.

Friedman, C. (1997). Towards a comprehensive medical language processing system: methods and issues. In *AMIA Annual Symposium Proceedings*, 595–599.

Friedman, C. (2000). A broad-coverage natural language processing system. In *AMIA Annual Symposium Proceedings*, Los Angeles, 270–274.

Friedman, C., Alderson, P. O., Austin, J. H., Cimino, J. J., and Johnson, S. B. (1994). A general natural-language text processor for clinical radiology. *J Am Med Inform Assoc*, 1(2), 161–174.

Friedman, C., and Hripcsak, G. (1998). Evaluating natural language processors in the clinical domain. *Methods Inf Med*, 37, 334–344.

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

Goryachev, S., Kim, H., and Zeng-Treitler, Q. (2008). Identification and extraction of family history information from clinical reports. In *AMIA Annual Symposium Proceedings*, Washington, DC, 247–251.

Goryachev, S., Sordo, M., and Zeng, Q. T. (2006). A suite of natural language processing tools developed for the I2B2 project. In *AMIA Annual Symposium Proceedings*, Washington, DC, 931.

Haerian, K., Varn, D., Vaidya, S., Ena, L., Chase, H. S., and Friedman, C. (2012). Detection of pharmacovigilance-related adverse events using electronic health records and automated methods. *Clin Pharmacol Ther*, 92(2), 228–234.

Hripcsak, G., Bakken, S., Stetson, P. D., and Patel, V. L. (2003). Mining complex clinical data for patient safety research: a framework for event discovery. *J Biomed Inform*, 36(1–2), 120–130.

Hripcsak, G., Kuperman, G. J., and Friedman, C. (1998). Extracting findings from narrative reports: software transferability and sources of physician disagreement. *Methods Inform Med*, 37, 1–7.

Huang, Y., Lowe, H. J., Klein, D., and Cucina, R. J. (2005). Improved identification of noun phrases in clinical radiology reports using a high-performance statistical natural language parser augmented with the UMLS SPECIALIST Lexicon. *J Am Med Inform Assoc*, 12(3), 275–285.