Chapter 16
Natural Language Processing, Electronic Health Records, and Clinical Research

Feifan Liu, Chunhua Weng, and Hong Yu

Abstract Electronic health records (EHR) capture "real-world" disease and care processes and hence offer richer and more generalizable data for comparative effectiveness research than traditional randomized clinical trial studies. With the increasingly broadening adoption of EHR worldwide, there is a growing need to widen the use of EHR data to support clinical research. A big barrier to this goal is that much of the information in EHR is still narrative. This chapter describes the foundation of biomedical natural language processing and its common uses for extracting and transforming narrative information in EHR to support clinical research.

Keywords Electronic health records • Biomedical natural language processing • Sublanguage approach • Machine-learning approach • Decision tree • Rule-based approach

Electronic health records (EHR) capture "real-world" disease and care processes and hence offer richer and more generalizable data for comparative effectiveness research [1] than traditional randomized clinical trial studies. With the increasingly broadening adoption of EHR worldwide, there is a growing need to widen the use of EHR data to support clinical research [2]. A big barrier to this goal is that much of the information in EHR is still narrative. This chapter describes the foundation of biomedical language processing and its common uses for extracting and transforming narrative information in EHR to support clinical research.

F. Liu, Ph.D. (✉) • H.Yu, Ph.D.
Department of Electric Engineering and Computer Science,
University of Wisconsin Milwaukee,
3200 N Cramer Street, Milwaukee, WI 53211, USA
e-mail: liuf@uwm.edu

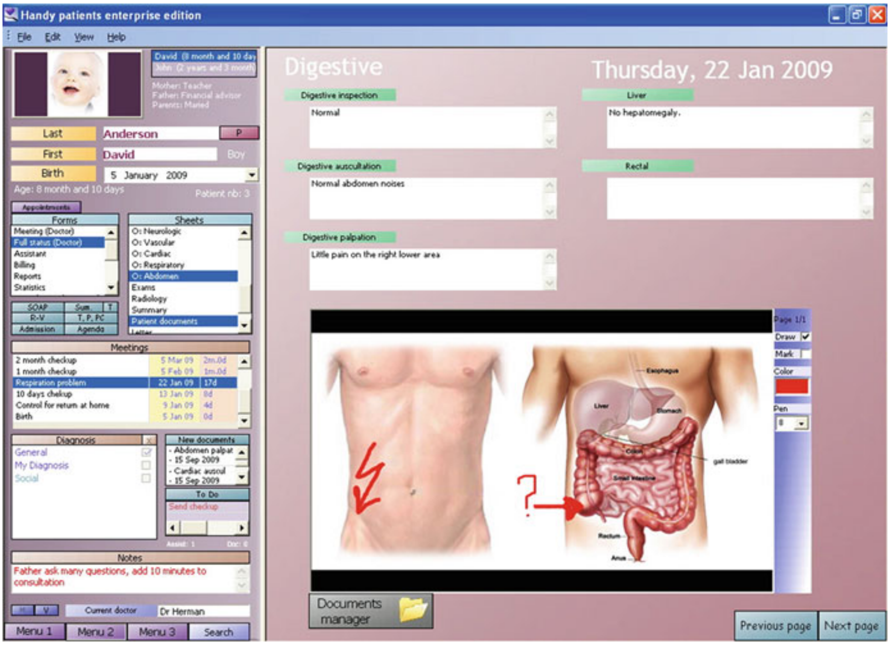
C. Weng, Ph.D.
Department of Biomedical Informatics, Columbia University,
622 W 168th Street, VC-5, New York, NY 10332, USA

R.L. Richesson, J.E. Andrews (eds.), *Clinical Research Informatics*, Health Informatics,
DOI 10.1007/978-1-84882-448-5_16, © Springer-Verlag London Limited 2012

Accelerating Clinical Research Using EHR: Opportunities and Challenges

The NIH defines clinical research as *patient-oriented research*, *epidemiological and behavioral studies*, *or outcomes and health services research* [3]. Patient-oriented research involves a particular person or group of people or uses materials from humans. In recent years, national clinical research enterprises have been under increased jeopardy [4] in part due to the rising costs associated with participant screening and recruitment, as well as issues surrounding data collection. Only 13% of clinicians are involved in clinical research [5]. To integrate research with clinical care, and to speed the application of research findings to clinical practice, the National Institute of Health (NIH) has created the Clinical and Translational Science Awards (CTSA) program to reengineer the clinical research enterprise [6]. A potential powerful accelerator to clinical research is electronic health records.

An EHR is a legal computerized medical record for documenting patient information captured at every patient encounter [7, 8]. Figure 16.1 shows a sample EHR [9]. As of 2008, more than 40% of physicians in the USA were using EHRs, more than double the percentage at the start of the decade [10]. The resident population of the USA as of 2009 was 307 million [11]. During that same year, it was reported that 83% adults and 90% children had contact with a health-care professional, there were



David (8 month and 10 day
John (2 years and 3 month)

Mother: Teacher
Father: Financial advisor
Parents: Maried

Last: Anderson
First: David Boy
Birth: 5 January 2009
Age: 8 month and 10 days
Patient nb: 3

Forms: Meeting (Doctor), Full status (Doctor), Assistant, Billing, Reports, Statistics

Sheets: O: Neurologic, O: Vascular, O: Cardiac, O: Respiratory, O: Abdomen, Exams, Radiology, Summary, Patient documents, Letter

Meetings		
2 month checkup	5 Mar 09	2m.0d
1 month checkup	5 Feb 09	1m.0d
Respiration problem	22 Jan 09	17d
10 days chekup	13 Jan 09	8d
Control for return at home	9 Jan 09	4d
Birth	5 Jan 09	0d

Diagnosis: General ☑, My Diagnosis □, Social □

New documents:
- Abdomen palpat
- 15 Sep 2009
- Cardiac auscul
- 15 Sep 2009

To Do: Send checkup

Assist: 1 Doc: 0

Notes: Father ask many questions, add 10 minutes to consultation

Current doctor: Dr Herman

Digestive

Thursday, 22 Jan 2009

Digestive inspection: Normal

Liver: No hepatomegaly.

Digestive auscultation: Normal abdomen noises

Rectal:

Digestive palpation: Little pain on the right lower area

Fig. 16.1 Illustration of a sample electronic health record (EHR)

1.1 billion ambulatory care visits (to physician offices, hospital outpatient, and emergency departments), and the number of physician office visits was 902 million. In other words, there were possibly over 800 million record entries in EHRs in 2009.

EHRs offer great potential to improve the efficiency and reduce the cost for clinical research, but this potential has not yet been fully realized. EHR includes standards-based structured laboratory test results and narrative interpretations by care providers. Unstructured narrative information can be provided for admission notes, discharge summaries, radiology images, and all sorts of ancillary notes, etc. Unlocking discrete data elements from such narrative information is a big challenge for reusing EHR data for clinical research.

Many studies and demonstration projects have explored the use of EHR data for clinical research, including detecting possible vaccination reactions in clinical notes [12], identifying heart failure [13], classifying whether a patient has rheumatoid arthritis [14], identifying associations between diabetes medications and myocardial infarction [15], and predicting disease outcomes [16]. EHR data has also been used for computerized pharmacovigilance [17] (see Chap. 19). Below, we elaborate two common use cases as examples of applying information extraction and retrieval techniques in EHR to support clinical research.

Use Case 1: Eligibility Screening or Phenotype Retrieval

The foremost, albeit costly, information retrieval task in clinical research is eligibility screening, which is to determine whether a person may or may not be eligible to enter a clinical research study. Chute has described this as essentially “patient phenotype retrieval” since it is meant to identify patients who manifest certain characteristics, which include diagnosis, signs, symptoms, interventions, functional status, or clinical outcomes [18]. Such characteristics are generally described in the eligibility criteria section for a research protocol. In recent years, the increasing volume of genome-wide association studies also raised the demand for clinical phenotype retrieval in discovering the genetics underlying many medical conditions. Traditional methods of participants search through manual chart review cannot scale to meet this need. In the study of rare diseases, there are usually only a small number of patients available, so it is feasible to have research personnel carefully collect, record, and organize the phenotypic information of each study participant. Diseases like diabetes mellitus, hypertension, and obesity, however, are complex, multifactorial, and chronic, and it is likely that a large number of patients will need to be followed over an extended period to ascertain important phenotypic traits. Large-scale studies involving many participants, or even smaller studies in which participants are selected from a larger population, will require innovative means to extract reliable, useful phenotype information from EHR data.

In recent years, several academic institutions have used EHR data to electronically screen (E-Screen) eligible patients for clinical studies [19]. Manually screening charts is time-consuming for research personnel, who must search for information in patient records to determine whether a patient meets the eligibility criteria for a clinical trial. E-Screening, however, can exclude ineligible patients and establish a much smaller patient pool for manual chart review. Thus, E-Screening helps clinical research personnel's transition from random and burdensome browsing of patient records to a focused and facilitated review. Consistent with concerns for patient safety and trial integrity, clinical research personnel should review all patients classified as "potentially eligible" by E-screening to confirm their eligibility. E-screening systems essentially perform "prescreening" for clinical research staff and should not fully replace manual review.

Use Case 2: Secondary Use of Clinical Data for Research

The national movement toward the broad adoption of EHRs obviously means that more clinical data will be captured and stored electronically. Secondary use of data for clinical research is a competitive requirement for a clinical and research enterprise [20]. In late 2009, the National Center for Research Resources called for "widening the use of electronic health records for research" to strengthen our capacity for using clinical care data for research. The nation's transition from traditional clinical trials to comparative effectiveness research [21] led by the US government has further emphasized the need for effective tools to extract research variables from preexisting clinical data. As an example, i2b2 (Informatics for Integrating Biology and the Bedside) is an NIH-funded National Center for Biomedical Computing based at Partners HealthCare System. The i2b2 Center is developing a scalable informatics framework that will enable clinical researchers to use existing clinical data for discovery research. In addition to that, the US Office of the National Coordinator for Health Information Technology (ONC) recently awarded $60 million in research grants through the Strategic Health IT Advanced Research Projects (SHARP) Program to the Mayo Clinic College of Medicine for secondary use of EHR data research.

A major challenge of using EHRs to facilitate clinical research is that much EHR data are presented as clinical narratives, which is largely unstructured and poses machine readability problems. Clinical natural language processing has been an active field since the inception of EHR in the 1960s and is an area that explores tools that can effectively extract, mine, and retrieve clinically relevant structured data from narrative EHRs. Clinical natural language processing has been influenced by the theory of sublanguage, which is characterized by distinctive specializations of syntax and the occurrence of domain-specific word subclasses in particular syntactic combinations. More recently, clinical NLP has been experiencing a shift from rule-based approach to machine-learning methods, as discussed later.

The rest of this chapter is organized as follows: we will first introduce the foundations of clinical NLP research in terms of sublanguage analysis and machine-learning models, including cutting-edge information extraction and retrieval techniques that can be applied to EHRs-based clinical research. Then, a few existing clinical NLP systems will be reviewed, followed by discussions on the challenges and future directions in this field.

Foundations of Biomedical Natural Language Processing

Natural language processing (NLP) is a research field dedicated to enable computers with the right knowledge for understanding natural language text, ultimately to facilitate the different types of natural language interaction between humans and computers. Biomedical NLP is a subfield specified for biomedical texts from biology, medicine, and chemistry. There exists great variability in the language in each of these areas, as reflected in their respective literature, guidelines, etc. In addition, the same type of biomedical text, such as narrative in an EHR, as discussed earlier, could differ greatly due to the expression variances and some organization-specific variance (i.e., among different medical centers). Sublanguage and machine-learning theory and approaches lay strong foundations for developing efficient clinical NLP systems in many real-world applications. Although some approaches and models are described below in the context of biomedical NLP applications, all of them can be adapted on electronic health records (EHRs) for clinical research informatics.

Sublanguage Approach

A sublanguage is defined by Grishman [22] as a specialized form of natural language used to describe a limited subject matter, generally employed by a group of specialists dealing with a particular subject. Zellig Harris [23] was one of the first linguists to apply the term sublanguage to natural language, using algebra as the underlying formalism. He defines a sublanguage as a subset of the language that is closed under some or all of the operations of the language.

Sublanguage theory laid a foundation for NLP in specific contexts such as clinical narratives. Many NLP applications are developed by exploiting the sublanguage characteristics, that is, restricted domain syntax and semantics. For example, an electronic health record (EHR) is limited to discussions of patient care and is unlikely to cover gene annotations or cell-line issues as in the biomedical literature. Sublanguages have many unique properties in comparison to more everyday language, resulting in a specialized vocabulary, structural patterns, as well as specialized entities and relationships among them.

Vocabulary Level

A sublanguage tends to have a specialized vocabulary which is quite different from standard language. For example, “cell line” is unlikely to be mentioned in nonbiological documents. In particular, the development of scientific and technological advancements in the biomedical domain has led to the discovery of new biological objects, functions, and events, which can only be acquired by analyzing sublanguage in the corresponding corpus.

Syntax Level

A sublanguage is not merely an arbitrary subset of sentences and may differ in syntax structure as well as vocabulary. For example, in medicine, telegraphic sentences such as "patient improved" are grammatical due to operations that permit dropping articles and auxiliaries. In addition, there are certain patterns of expression in sublanguage consisting of predicate words and ordered arguments, as in "<antibody> <appeared in> <tissue>"; "appeared in" is predicate words, and "<antibody>" and "<tissue>" are two arguments which can have semantically related terms filled in.

Semantics and Discourse Level

In addition to differences on the vocabulary and syntax levels, a sublanguage may also have specialized ways of interpreting language and organizing larger units of discourse. For example, "secondary to" has a specialized meaning that indicates a causal relationship, which is different from its use in standard language. In discharge summaries, the structural format often includes history of present illness, medications on admission, social history, physical examination, etc.

These properties of sublanguages allow the use of methods of analysis and processing that would not be possible when processing the language of newspaper articles or novels. Sublanguage analysis also provides a way of integrating domain-specific knowledge with existing systems. For example, a biomedical information retrieval system can be developed by indexing medical articles on only terms from a list of terminology known to be of interest to researchers; controlled medical vocabulary can be derived using sublanguage analysis based on terms combining regularly with particular other words; biological information extraction system can be adapted by sublanguage analysis of specialized expression patterns; a system that analyzes clinical reports can look for predictable semantic patterns that are characteristics of the clinical domain [24–27].

Machine-Learning Approach

Sublanguage patterns (rules) and manually specified models often lack the quality of generalization and also are time-consuming to keep well maintained and updated. With the ever-growing availability of electronic biomedical resource data and advanced computational power, machine-learning models have been arousing intense interests for many biomedical NLP tasks, which can be mainly divided into five categories:

- Classification: assign documents predefined labels
- Ranking: order objects by preference
- Regression: obtain real-value output as prediction

- Structured prediction: sequence labeling and segmentation to recognize entities or other semantic units
- Clustering: discover the underlying structure of unlabeled data to form natural groups

Many clinical research informatics applications can be formulated into the abovementioned tasks, such as entity (medications, diseases, doses). Extraction from EHRs can be realized using structured prediction models; adverse events detection from EHRs is an example of classification tasks. For these tasks, the goal of machine learning is to enable correct predictions for target variables given observation variables (attributes or features) from corresponding instances. Different learning models have been applied in recent years. In terms of their modeling approaches, they can be grouped as generative models and discriminative models. The generative approach models a joint probability distribution over both input and output variables (observation and label sequences), such as Naive Bayes, Bayesian network, hidden Markov model, and Markov random field, while the discriminative approach directly models the dependence of the output variables (label to be predicted) on the input variables (observation) by conditional probability, such as decision tree, logistic regression, support vector machine, *k* nearest neighbor, artificial neural network, and conditional random fields. This section will cover the introductory descriptions of those algorithms, but we encourage interested readers to explore these in more detail through further readings [28–32].

Generative Model

The generative model is a full probability model on all variables, which can simulate the generation of values for any variables in the model. By using Bayes' rule, it can be formed as a conditional distribution to be used for classification. When there is little annotated data, the generative model is advantageous for making use of a large quantity of raw data for better performance. The generative model reduces the variance of parameter estimation by modeling the input, but at the expense of possibly introducing model bias.

Naive Bayes Classifier. The Naive Bayes classifier is based on Bayesian theorem [33] and is a very simple probabilistic generative model that can be used to compute the probability of each candidate class label given observed features, under the assumption that all the features are independent given class label. It requires only a small size of training data with faster parameter estimation, but the strong independence assumption is violated in numerous occasions for real applications, which can lead to a large bias.

Bayesian network. Bayesian network [34], also belief network, is a probabilistic graphical model, whose nodes are a set of random variables connected by a directed acyclic graph (DAG) to represent the conditional dependences among those variables. This model does not require the independence assumption as in Naive Bayes, providing stronger representational power in real-world applications and making the parameter estimation more complex as well. It models the dependency between variables providing a good ability to handle missing values and is widely used in causal relationship reasoning applications, such as clinical decision support [35] and gene expression data analysis [36].

Hidden Markov Model. The hidden Markov model (HMM) [37] is a probabilistic generative model of a Markov process (Markov chain), where the model passes different state sequences, which are unobserved, producing a sequence of observations. Each hidden state has a probability distribution over the possible output observations, and there are transition probabilities among those states.

HMM is widely used in temporal pattern recognition (e.g., medical dictation system) and other sequence-labeling tasks (e.g., gene/protein recognition [38] and biosequence alignment [39]). Although this type of statistical model has worked extremely well in many situations, it does have limitations. A major limitation is the assumption that successive observations are independent, which cannot take into account the contextual dependency in the observation sequence. Another limitation is the Markov assumption itself, that is, the current state only depends on the immediate preceding state, which is also inappropriate for some problems.

Markov Random Field. Markov random field (MRF), also a Markov network or undirected graphic model [40], is a graphic model on the joint probability over a set of random variables each corresponding to a node in the graph. Markov properties exist among those variables to provide conditional independence for graph factorization.

MRF is similar to Bayesian network in terms of modeling dependency relationships among variables. Bayesian network is a direct graphic model, and it represents probability distributions that can be factorized into products of conditional distributions, which is desirable to capture causal relationships among variables, while MRF is an undirected graphic model, where there is no directionality on each edge connecting a pair of nodes, and the probability distribution it represents will be factorized into products of potential functions of conditionally independent cliques[28], which makes MRF better suited to expressing soft constrains between random variables. In addition, MRF can represent certain dependencies that a Bayesian network cannot, such as cyclic dependencies, wherein it cannot represent certain dependencies that a Bayesian network can such as induced dependencies. MRF model has been successfully applied in biomedical image analysis for computer-aided diagnosis as shown in [41, 42].

Discriminative Model

Compared with the generative model, the discriminative model is designed to only involve a target variable(s) conditional on the observed variables, directly computing the input to output mappings (posterior) and eschewing the underlying distributions of the input. As there are fewer independence assumptions, the discriminative model often provides more robust generalization performance when enough annotated data are available. However, it usually lacks flexible modeling methods for prior knowledge, structure, uncertainty, etc. In addition, the relationships between variables are not as explicit or visualizable as in the generative model.

Decision Tree. A decision tree (DT) [43] is a logical model represented as a tree structure that shows how the value of a target variable can be predicted by using the values of a set of observation variables (attributes). Each branch node represents a split between a number of alternatives based on a specific attribute, and each leaf node represents a decision. The induction of a decision tree is a top-down process to reduce information content by mapping them to fewer outputs but seek a trade-off between accuracy and simplicity.

Decision trees provide a way to easily understand the derived decision rules and interpret the predicted results and have been used for diagnosis of aortic stenosis [44] and folding mechanism prediction of protein homodimer [45]. One of the disadvantages of DT models is that DT split the training set into smaller and smaller subsets, which makes correct generalization harder and incorrect generalization easier because smaller sets have accidental regularities that do not generalize. Pruning can address this problem to some extent though.

Logistic Regression. Logistic function was first discovered by Peral and Reed [46] in 1920, and logistic regression is a generalized linear model used to calculate the probability of the occurrence of an event by fitting the data to a logit function through maximum likelihood. It is a discriminative counterpart of naive Bayes model as they represent the same set of conditional probability distributions. It has been extensively used for prediction and diagnosis in medicine [47, 48] due to its robustness, flexibility, and ability to handle nonlinear effects. But generally, it requires more data to achieve stable and meaningful results than standard regression.

Support Vector Machines. Support vector machines (SVMs) [49] are also linear models that are trained to separate the data points (instances) based on both empirical and structural risk minimum principles; that is, they not only classify objects into categories but also construct a hyperplane or set of hyperplanes in a high dimension space with a maximum margin among different categories. New instances are then mapped into the same space and classified into a category based on which side of hyperplanes they fall on.

The SVM model has been used for many biomedical tasks, such as microarray data analysis [50], classification [51], information extraction [52], and image segmentation [53]. SVM model can leverage an arbitrary set of features to produce accurate and robust results on a sound theoretical basis, with powerful generalization ability due to optimizing margins. However, from a practical point of view, the most serious problem with the SVM model is the high level of computational complexity and extensive memory requirements for large-scale tasks.

K Nearest Neighbor. The *k* nearest neighbor (*k*-NN) rule [54] is a type of instance-based learning, or lazy learning, where generalization beyond the training data is delayed. The goal is to assign a new instance a value or category that is averaged (for regression) or voted (for classification) based on examining the *k* closest labeled training instances.

The *k*-NN method has been used in gene expression analysis [55], screening data analysis [56], protein-protein interaction [57], biomedical image interpretation [58], etc. The main advantage of this method is that the target function will be approximated locally for each new instance so that it can deal well with changes in the problem domain. A practical problem is that it tends to be slower especially for large training sets as the entire training set would be traversed for each new instance.

Artificial Neural Network. Artificial neural networks (ANNs) [59] are a mathematical model of human intellectual abilities that seek to simulate the structure and functional aspects of biological neural networks. In an ANN model, the artificial neutrons (processing units) are connected together via unidirectional signal channels in different layers to mimic the biological neural network. Usually, only neutrons in two consecutive layers are connected.

In the biomedical domain, ANNs have been used for many diagnostic [60, 61] and prognostic [62, 63] tasks. Neural networks have the ability to implicitly detect complex nonlinear relationships between dependent and independent variables, as well as possible interactions among predictor variables. On the other hand, ANNs are computationally expensive, prone to overfitting, and lack a sound theoretical foundation.

Conditional Random Fields. Conditional random fields (CRFs) [64] consist of a probabilistic framework for labeling and segmenting structured data, such as sequences, trees, and lattices. The underlying idea is that of defining a conditional probability distribution over label sequences given a particular observation sequence, rather than a joint distribution over both label and observation sequences.

Much like MRF, a CRF is an undirected graphic model, but they have different characteristics. CRFs would have better predictive power due to direct modeling on posterior, have flexible to use features from different aspects, and relax the strong assumption of conditional independence of the observed data. On the other hand, MRFs can handle incomplete data problems, augment small labeled data with larger amounts of cheap unlabelled data. Similarly, the primary advantage of CRFs over hidden Markov models (HMM) [37] is also their conditional nature, resulting in the relaxation of the independence assumptions required by HMMs in order to ensure tractable inference. Additionally, CRFs avoid the label bias problem, a weakness exhibited by maximum entropy Markov models (MEMMs) [65] and other conditional Markov models based on directed graphical models. CRF model is very popular in biomedical entity recognition [66, 67], relation extraction [68], and event detection [69].

Unsupervised Clustering

The learning models discussed above are mostly for supervised learning, which requires labeled data for model training. Clustering is a commonly used unsupervised learning method which automatically discovers the underlying structure or pattern in a collection of unlabelled data. The goal is to partition a set of objects into subsets whose members are similar in some way as well as dissimilar to members from a different subset. Determining how similarity (or dissimilarity) between objects is defined and measured is very crucial for the clustering task. Examples of distance metrics are Mahalabobis, Euclidean, Minkowski, and Jeffreys-Matusita. There are three main types of clustering approaches: partition clustering [70], hierarchical clustering [71], and a mixed model [72].

The most typical example of clustering in bioinformatics is microarray analysis [55, 73–76], where genes with expressional similarities are grouped together, assuming that they have regulatory or functional similarity.

An Overview of Existing Clinical Natural Language Processing Systems

In electronic health records (EHRs), the central challenge of extracting detailed medical information is dealing with the heterogeneity of clinical data, which involves both structured descriptions and narratives. Over the last two decades, there have been great efforts to develop biomedical NLP systems for clinical narrative text mining. There are mainly two types of approaches that have been explored. Rule-based approaches focus on making use of sublanguage analysis and pattern matching rules, while machine learning–based approaches investigate various useful features and appropriate algorithms. For both approaches, a domain knowledge resource is generally used.

Rule-Based Approach

One of the earliest clinical NLP systems developed, which emerged from the Linguistic String Project [77, 78], used comprehensive syntactic and semantic knowledge rules to extract encoded information from clinical narratives. But systems containing syntactic knowledge are very time-consuming to build and maintain because syntax is so complex.

Later, MedLEE (Medical Language Extraction and Encoding system) system [79] was developed to process clinical information expressed in natural language. This system incorporates a semantically based (simple syntax rules are also included) parser for determining the structure of text. The parser is driven by a grammar that consists of well-defined semantic patterns, their interpretations, and the underlying target structures. By integrating the pattern matching with semantic techniques, the MedLEE system is expected to reduce the ambiguity within the language of domain because of the underlying semantics.

Gold et al. [80] developed a rule-based system called MERKI to extract medication names and the corresponding attributes from structured and narrative clinical texts. Recently, Xu et al. [81] built an automatic medication extraction system (MedEx) on discharge summaries by leveraging semantic rules and a chart parser, achieving promising results for extracting medication and related fields, for example, strength, route, frequency, form, dose, duration, etc. This information was defined by a simple semantic representation model for prescription type of medication findings, into which medication texts were mapped.

Learning-Based Approach

SymText (Symbolic Text Processor) [82] is a learning-based NLP system which integrates a syntactic parser based on augmented transition networks and transformational grammars with a semantics model based on the Bayesian network [34] statistical formalism, which has been used for various applications such as extracting pneumonia-related findings from chest radiograph reports [83].

Agarwal and Yu developed two biomedical NLP systems named NegScope [84] and HedgeScope [85], which were able to detect negation and hedge cues as well as their scopes in both the biomedical literature and clinical notes. Both systems were built on the conditional random fields (CRFs) [64] learning model trained on the publicly available BioScope [86] corpus.

Lancet [87] is a supervised machine-learning system that automatically extracts medication events consisting of medication names and information pertaining to their prescribed use (dosage, mode, frequency, duration, and reason) from clinical discharge summaries. Lancet employs the CRFs model [64] for tagging individual medication names and associated fields, and the AdaBoost model with decision stump algorithm [88] for determining which medication names and fields belong to a single medication event. During the third i2b2 shared task for challenges in natural language processing for clinical data, medication extraction challenge, Lancet achieved the highest precision among top ten systems.

In order to help health-care providers quickly and efficiently answer the questions that arise during their meetings with patients, Cao et al. [89] built a clinical question answering system called AskHERMES, a computational system that automatically analyzes the input clinical questions, retrieves and mines large sets of literature documents and clinical notes pertaining to specific questions, and generates short text summary as the output answer. For question analysis [90], support vector machines (SVMs) learning algorithm [49] and CRFs model [64] were used for the question classification and keyword identification, respectively. The system is designed to enable health-care providers to efficiently seek information in clinical settings.

Liu et al. [91] developed speech recognition system for clinical setting, ClinicalASR, to provide the speech interface to clinical NLP applications for more efficient information access, such as clinical question answering system. ClinicalASR explored language model (LM)–based adaptation on the SRI Decipher system [92] using clinical questions.

Challenges and Future Directions

Although remarkable progress has been made for clinical NLP, there are many challenges and open questions to be investigated in the future. One obstacle to clinical NLP is access to EHRs. In the USA, the Health Insurance Portability and Accountability Act of 1996 (HIPAA) has required that the use of protected health information (PHI) in research studies is not permitted except with the explicit consent of the patient, which prevents gathering data for NLP applications if the data are not de-identified. But HIPAA does allow for the creation of de-identified health information. De-identification tools have been developed, and commercial tools are also available. De-ID [93] information has been used by affiliated hospitals at the University of Pittsburgh, which made available a whole year of EHR data for NLP use. Currently, de-identification tools are still not widely used by hospitals, hampering the NLP applications which are highly based on available EHR data.

Although the sublanguage analysis works well in many subdomains, it is very time-consuming to compile rules syntactically and semantically and needs a lot of efforts to keep them well maintained, especially as ever-increasing amount of EHR data become available. But sublanguage analysis does provide more information that could be helpful in the design of learning-based systems. Therefore, how to effectively and systematically integrate the sublanguage analysis as feature into the learning framework and how to employ the learning methods for automatically extracting sublanguage specific patterns have great potential to facilitate the advancement of EHR-based clinical research informatics.

Currently, most clinical NLP systems are still in an experimental stage rather than deployed and regularly used in clinical setting. The difficulties in translation of clinical NLP research into clinical practice and obstacles in determining the level of practical engagement of NLP systems provide more challenging research opportunities in this field. In addition, to assist clinical decision support, NLP system needs to deal with time series information extraction, reasoning, and integration, for example, linking clinical findings to patient profile, linking different records of same patient, and integrating factual information from multiple sources. However, all those tasks are not trivial in the clinical setting. Last but not least, effectively mining EHRs for clinical research presents the following two challenges:

1. Data Quality Issues. EHR data hold the promise for secondary use for research and quality improvement; however, such uses remain extremely challenging because EHR data can be inaccurate, incomplete, fragmented, and inconsistent in semantic representations for common concepts. For example, patient data such as glomerular filtration rate (GFR) or body mass index are often unavailable in EHR but are important research variables. In addition, for a study looking for hypertension patients, the determination of hypertension should account for the use of hypertensive drugs, the ICD-9 diagnosis codes for hypertension, or the blood pressure values out of the normal range in certain measurements contexts. Blood pressure values captured in an emergency room are found to be generally elevated compared with the blood pressure values documented during physical exams; therefore, the former value may not represent the patient's real value.

Moreover, the saying “absence of evidence is not evidence of absence” is very true for using EHR data. If a clinical research investigator is looking for patients with cardiovascular diseases but cannot find corresponding diagnoses in a patient, the investigator cannot jump to the conclusion that the patient has no cardiovascular disease until further confirmation can be obtained. Typical reasons can be that the patient’s medical history is not completely captured by the hospital where the EHR is used or the patient has not been diagnosed. Moreover, much data are not amenable for computer processing, especially those in free-text notes. Whenever it is free-text, there is a challenge for identifying semantic equivalence of multiple linguistic forms of the shared concepts. For example, among hypertensive patients, the medical records can store values such as “HTN,” “hypertension,” or “401.9” as an ICD-9 code to indicate hypertension.

2. Challenges for Converting Clinical Data to Research Variables. Many people are still skeptical about reusing clinical data for clinical research because they believe clinical data are “garbage in, garbage out.” Although this statement is a little exaggerated, there are dramatic differences between a clinical database and a clinical research database developed following a rigorous clinical research protocol. A research protocol will specify what data will be collected at what time and how. A clinical research database is often designed as a relational database with a tabular format, organized by patient and variables over time. There is a strict quality assurance procedure to ensure the completeness and accuracy of research data. Furthermore, clinical research databases are optimized for statistical analysis. In contrast, a clinical database is organized by clinical events, not by patients. Moreover, clinical data are collected for administrative uses or personal interpretations of medical doctors. Copy and paste, as well as creative abbreviations that only doctors themselves can interpret in certain contexts are very common in clinical databases. Therefore, ad hoc extraction of research variables from a clinical database is not a trivial task.

In conclusion, natural language processing (NLP) offers an effective way to unlock disease knowledge from unstructured clinical narratives. Although standards are emerging and EHR data is becoming better encoded with clinical terminology standards, there will likely always be a narrative aspect (at least for the foreseeable future), which makes clinical NLP technologies indispensible for clinical researchers and informatics professionals. Different approaches and models have been widely applied for biomedical literature, and all those NLP techniques are crucial and can be adapted for effectively mining electronic health records (EHRs) to support important clinical research activities.

References

1. Sox HC, Greenfield S. Comparative effectiveness research: a report from the institute of medicine. Ann Intern Med. 2009;151:203–5.
2. National Center for Research Resources (U.S.). Widening the use of electronic health record data for research [electronic resource]. Bethesda, MD. : National Institutes of Health, 2009.

3. National Institute of Child Health & Human Development. Clinical Research and Clinical Trials. http://www.nichd.nih.gov/health/clinicalresearch/. Accessed Aug 2011.
4. Sung NS, Crowley WF, Genel M, Salber P, Sandy L, Sherwood LM, Johnson SB, Catanese V, Tilson H, Getz K, Larson EL, Scheinberg D, Reece EA, Slavkin H, Dobs A, Grebb J, Martinez RA, Korn A, Rimoin D. Central challenges facing the national clinical research enterprise. JAMA. 2003;289:1278–87.
5. Harris Interactive. Most physicians do not participate in clinical trials because of lack of opportunity, time, personnel support and resources. Rochester, NY, June 11, 2004. http://www.harrisinteractive.com/news/allnewsbydate.asp?NewsID=811. Accessed Aug 2011.
6. National Center for Research Resources. Clinical and Translational Science Awards. http://www.ncrr.nih.gov/clinical%5Fresearch%5Fresources/clinical%5Fand%5Ftranslational%5Fscience%5Fawards/. Accessed Aug 2011.
7. Garets D, Davis M. Electronic medical records vs. electronic health records: yes, there is a difference. A HIMSS Analytics White Paper, HIMSS Analytics, Chicago; 2005.
8. Garets D, Davis M. Electronic patient records, EMRs and EHRs: concepts as different as apples and oranges at least deserve separate names. Healthc Inform. 2005;22(10):53–4.
9. Wikipedia. File: Electronic medical record.jpg. http://en.wikipedia.org/wiki/File:Electronic_medical_record.jpg. Accessed Dec 2011.
10. Walker EP. More doctors are using electronic medical records 2010. http://www.medpagetoday.com/PracticeManagement/InformationTechnology/17862. Accessed Aug 2011.
11. U.S. Census Bureau. National Totals: Vintage 2009. http://www.census.gov/popest/data/national/totals/2009/index.html. Accessed Dec 2011.
12. Hazlehurst B, Mullooly J, Naleway A, Crane B. Detecting possible vaccination reactions in clinical notes. AMIA Annu Symp Proc. 2005; 2005:306–10.
13. Pakhomov S, Weston SA, Jacobsen SJ, Chute CG, Meverden R, Roger VL. Electronic medical records for clinical research: application to the identification of heart failure. Am J Manag Care. 2007;13:281–8.
14. Liao KP, Cai T, Gainer V, Goryachev S, Zeng-treitler Q, Raychaudhuri S, Szolovits P, Churchill S, Murphy S, Kohane I, Karlson EW, Plenge RM. Electronic medical records for discovery research in rheumatoid arthritis. Arthritis Care Res (Hoboken). 2010;62:1120–7.
15. Brownstein JS, Murphy SN, Goldfine AB, Grant RW, Sordo M, Gainer V, Colecchi JA, Dubey A, Nathan DM, Glaser JP, Kohane IS. Rapid identification of myocardial infarction risk associated with diabetes medications using electronic medical records. Diabetes Care. 2010;33:526–31.
16. Reis BY, Kohane IS, Mandl KD. Longitudinal histories as predictors of future diagnoses of domestic abuse: modelling study. BMJ. 2009;339:b3677.
17. Wang X, Hripcsak G, Markatou M, Friedman C. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. J Am Med Inform Assoc. 2009;16:328–37.
18. Chute CG. The horizontal and vertical nature of patient phenotype retrieval: new directions for clinical text processing. Proc AMIA Symp. 2002:165–9.
19. Thadani SR, Weng C, Bigger JT, Ennever JF, Wajngurt D. Electronic screening improves efficiency in clinical trial recruitment. J Am Med Inform Assoc. 2009;16:869–73.
20. Embi PJ, Payne PRO. Clinical research informatics: challenges, opportunities and definition for an emerging domain. J Am Med Inform Assoc. 2009;16:316–27.
21. Kuehn BM. Institute of Medicine outlines priorities for comparative effectiveness research. JAMA. 2009;302:936–7.
22. Grishman R, Hirschman L, Nhan NT. Discovery procedures for sublanguage selectional patterns: initial experiments. Comput Linguist. 1986;12:205–15.
23. Harris Z. Mathematical structures of language. New York: Wiley; 1968.
24. Grishman R, Kittredge R. Analyzing language in restricted domains: sublanguage description and processing. New York: Routledge; 1986.
25. Johnson SB, Gottfried M. Sublanguage analysis as a basis for a controlled medical vocabulary. Proc Annu Symp Comput Appl Med Care. 1989:519–23.
26. Bronzino JD. The biomedical engineering handbook. New York: Springer; 2000.

27. Friedman C, Kra P, Rzhetsky A. Two biomedical sublanguages: a description based on the theories of Zellig Harris. J Biomed Inform. 2002;35:222–35.
28. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: data mining, inference, and prediction. New York: Springer; 2001.
29. Bishop C. Pattern recognition and machine learning (information science and statistics). New York: Springer; 2007.
30. Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference. San Francisco: Morgan Kaufmann Publishers Inc; 1988.
31. Michalski RS, Carbonell JG, Mitchell TM. Machine learning: an artificial intelligence approach. Los Altos: Morgan Kaufmann Pub; 1986.
32. Manning CD, Schütze H. Foundations of statistical natural language processing. Cambridge: MIT Press; 2000.
33. Bayes M, Price M. An essay towards solving a problem in the Doctrine of chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S. Philos Transact (1683–1775). 1763;53:370–418.
34. Pearl J. Bayesian networks: a model of self-activated memory for evidential reasoning. In: Proceedings of the 7th cConference of the Cognitive Science Society, University of California, Irvine; 1985, p. 334, 329.
35. Verduijn M, Peek N, Rosseel PMJ, de Jonge E, de Mol BAJM. Prognostic Bayesian networks: I: rationale, learning procedure, and clinical use. J Biomed Inform. 2007;40:609–18.
36. Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. J Comput Biol. 2000;7:601–20.
37. Baum LE, Petrie T. Statistical inference for probabilistic functions of finite state Markov chains. Ann Math Stat. 1966;37:1554–63.
38. Lukashin AV, Borodovsky M. GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res. 1998;26:1107–15.
39. Yu L, Smith TF. Positional statistical significance in sequence alignment. J Comput Biol. 1999;6:253–9.
40. Kindermann R. Markov random fields and their applications, Contemporary mathematics, vol. 1. Providence: American Mathematical Society; 1950.
41. Komodakis N, Besbes A, Glocker B, Paragios N. Biomedical image analysis using Markov random fields & efficient linear programing. Conf Proc IEEE Eng Med Biol Soc. 2009;2009:6628–31.
42. Lee N, Laine AF, Smith RT. Bayesian transductive Markov random fields for interactive segmentation in retinal disorders. In: World congress on medical physics and biomedical engineering, Munich; 7–12 Sept 2009. p. 227–30.
43. Quinlan JR. Induction of decision trees. Mach Learn. 1986;1:81–106.
44. Pavlopoulos S, Stasis A, Loukis E. A decision tree – based method for the differential diagnosis of aortic stenosis from mitral regurgitation using heart sounds. Biomed Eng Online. 2004;3:21.
45. Suresh A, Karthikraja V, Lulu S, Kangueane U, Kangueane P. A decision tree model for the prediction of homodimer folding mechanism. Bioinformation. 2009;4:197–205.
46. Pearl R, Reed LJ. A further note on the mathematical theory of population growth. Proc Natl Acad Sci USA. 1922;8:365–8.
47. Bagley SC, White H, Golomb BA. Logistic regression in the medical literature: standards for use and reporting, with particular attention to one medical domain. J Clin Epidemiol. 2001;54:979–85.
48. Gareen IF, Gatsonis C. Primer on multiple regression models for diagnostic imaging research. Radiology. 2003;229:305–10.
49. Vapnik VN. The nature of statistical learning theory. New York: Springer; 1995.
50. Brown MPS, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M, Haussler D. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci USA. 2000;97:262–7.

51. Polavarapu N, Navathe SB, Ramnarayanan R, ul Haque A, Sahay S, Liu Y. Investigation into biomedical literature classification using support vector machines. Proc IEEE Comput Syst Bioinform Conf. 2005:366–74.
52. Takeuchi K, Collier N. Bio-medical entity extraction using support vector machines. In: Proceedings of the ACL 2003 workshop on natural language processing in biomedicine – Volume 13, Association for Computational Linguistics, Sapporo; 2003. p. 57–64.
53. Pan C, Yan X, Zheng C. Hard Margin SVM for biomedical image segmentation. In: Wang J, Liao X-F, Yi Z, editors. Advances in neural networks – ISNN 2005. Heidelberg: Springer; 2005. p. 754–9.
54. Fix E, Jr. Discriminatory analysis: nonparametric discrimination: consistency properties. Technical Report, No. Project 21-49-004, Report Number 4, 1951:261–279.
55. Pan F, Wang B, Hu X, Perrizo W. Comprehensive vertical sample-based KNN/LSVM classification for gene expression analysis. J Biomed Inform. 2004;37:240–8.
56. Shanmugasundaram V, Maggiora GM, Lajiness MS. Hit-directed nearest-neighbor searching. J Med Chem. 2005;48:240–8.
57. Qi Y, Klein-Seetharaman J, Bar-Joseph Z. Random forest similarity for protein-protein interaction prediction from multiple sources. Pac Symp Biocomput. 2005:531–42.
58. Barbini P, Cevenini G, Massai MR. Nearest-neighbor analysis of spatial point patterns: application to biomedical image interpretation. Comput Biomed Res. 1996;29:482–93.
59. McCulloch W, Pitts W. A logical calculus of the ideas immanent in nervous activity. Bull Math Biol. 1990;52:99–115.
60. Xue Q. Reddy BRS: late potential recognition by artificial neural networks. IEEE Trans Biomed Eng. 1997;44:132–43.
61. Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C, Meltzer PS. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med. 2001;7:673–9.
62. Jerez-Aragonés JM, Gómez-Ruiz JA, Ramos-Jiménez G, Muñoz-Pérez J, Alba-Conejo E. A combined neural network and decision trees model for prognosis of breast cancer relapse. Artif Intell Med. 2003;27:45–63.
63. Burke HB, Goodman PH, Rosen DB, Henson DE, Weinstein JN Jr FEH, Marks JR, Winchester DP, Bostwick DG. Artificial neural networks improve the accuracy of cancer survival prediction. Cancer. 1997;79:857–62.
64. Lafferty J, McCallum A, Pereira F. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proc of International Conference on Machine Learning (ICML), San Francisco, CA, 2001:282–89.
65. McCallum A, Freitag D, Pereira FCN. Maximum entropy Markov models for information extraction and segmentation. In: Proceedings of the seventeenth international conference on machine learning. San Francisco: Morgan Kaufmann Publishers Inc.; 2000. p. 591–8.
66. Settles B. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. Bioinformatics. 2005;21:3191–2.
67. Leaman R, Gonzalez G. Banner: an executable survey of advances in biomedical named entity recognition. Pac Symp Biocomput. 2008;13:652–63.
68. Bundschus M, Dejori M, Stetter M, Tresp V, Kriegel H-P. Extraction of semantic biomedical relations from text using conditional random fields. BMC Bioinformatics. 2008;9:207.
69. Sarafraz F, Eales J, Mohammadi R, Dickerson J, Robertson D, Nenadic G. Biomedical event detection using rules, conditional random fields and parse tree distances. In: Proceedings of the workshop on BioNLP: shared task, Association for Computational Linguistics, Colorado; 2009. p. 115–8.
70. Forgy E. Cluster analysis of multivariate data: efficiency vs. interpretability of classifications. Biometrics. 1965;21:768.
71. Jardine N, Sibson R. Mathematical taxonomy. New York: Wiley; 1971.
72. McLachlan GJ, Basford KE. Mixture models. Inference and applications to clustering. New York: Marcel Dekker; 1988.

73. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci USA. 1999;96:2907.
74. De Smet F, Mathys J, Marchal K, Thijs G, De Moor B, Moreau Y. Adaptive quality-based clustering of gene expression profiles. Bioinformatics. 2002;18:735.
75. Sheng Q, Moreau Y, De Moor B. Biclustering microarray data by Gibbs sampling. Bioinformatics. 2003;19 Suppl 2:ii196–205.
76. Schafer J, Strimmer K. An empirical Bayes approach to inferring large-scale gene association networks. Bioinformatics. 2005;21:754.
77. Sager N, Friedman C, Chi E. The analysis and processing of clinical narrative. Medinfo. 1986;1986:86.
78. Sager N, Friedman C, Lyman MS, others. Medical language processing: computer management of narrative data. Reading: Addison-Wesley; 1987.
79. Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. J Am Med Inform Assoc. 1994;1:161–74.
80. Gold S, Elhadad N, Zhu X, Cimino JJ, Hripcsak G. Extracting structured medication event information from discharge summaries. AMIA Annu Symp Proc. 2008;2008:237–41.
81. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives. J Am Med Inform Assoc. 2010;17:19–24.
82. Haug PJ, Koehler S, Lau LM, Wang P, Rocha R, Huff SM. Experience with a mixed semantic/ syntactic parser. Proc Annu Symp Comput Appl Med Care. 1995;1995:284–8.
83. Fiszman M, Chapman WW, Aronsky D, Evans RS, Haug PJ. Automatic detection of acute bacterial pneumonia from chest X-ray reports. J Am Med Inform Assoc. 2000;7:593–604.
84. Agarwal S, Hong Yu. Biomedical negation scope detection with conditional random fields. J Am Med Inform Assoc. 2010;17(6):696–701.
85. Agarwal S, Yu H. Detecting hedge cues and their scope in biomedical literature with conditional random fields. J Biomed Inform. 2010;43(6):953–61. Epub 2010 Aug 13.
86. Vincze V, Szarvas G, Farkas R, Mora G, Csirik J. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. BMC Bioinformatics. 2008;9:S9.
87. Li Z, Liu F, Antieau L, Cao Y, Yu H. Lancet: a high precision medication event extraction system for clinical text. J Am Med Inform Assoc. 2010;17(5):563–7.
88. Rennie J. Boosting with decision stumps and binary features. Relation. 2003;10:1666.
89. Cao Y, Liu F, Simpson P, Antieau L, Bennett A, Cimino JJ, Ely J, Yu H. AskHERMES: an online question answering system for complex clinical questions. J Biomed Inform. 2011;44(2):277–88. Epub 2011 Jan 21.
90. Cao Y-gang, Cimino JJ, Ely J, Yu H. Automatically extracting information needs from complex clinical questions. J Biomed Inform. 2010;43(6):962–71. Epub 2010 Jul 27.
91. Liu F, Kruse AM, Tur G, Hakkani-Tür D. Towards spoken clinical question answering: evaluating automatic speech recognition systems for clinical spoken questions. J Am Med Inform Assoc. 2011;18(5):625–30.
92. Stolcke A, Anguera X, Boakye K, Çetin Ö, Janin A, Peskin B, Wooters C, Zheng J. Further progress in meeting recognition: the ICSI-SRI Spring 2005 speech-to-text evaluation system. Vol. 3869, LNCS, MLMI workshop 2005;78:463–75.
93. Gupta D, Saul M, Gilbertson J. Evaluation of a deidentification (De-Id) software engine to share pathology reports and clinical documents for research. Am J Clin Pathol. 2004;121: 176–86.