Extracting Information from Textual Documents in the Electronic Health Record: A Review of Recent Research

S. M. Meystre[1], G. K. Savova[2], K. C. Kipper-Schuler[2], J. F. Hurdle[1]

[1] Department of Biomedical Informatics, University of Utah School of Medicine, Salt Lake City, Utah, USA
[2] Biomedical Informatics Research, Mayo Clinic College of Medicine, Rochester, Minnesota, USA

Summary

Objectives: We examine recent published research on the extraction of information from textual documents in the Electronic Health Record (EHR).

Methods: Literature review of the research published after 1995, based on PubMed, conference proceedings, and the ACM Digital Library, as well as on relevant publications referenced in papers already included.

Results: 174 publications were selected and are discussed in this review in terms of methods used, pre-processing of textual documents, contextual features detection and analysis, extraction of information in general, extraction of codes and of information for decision-support and enrichment of the EHR, information extraction for surveillance, research, automated terminology management, and data mining, and de-identification of clinical text.

Conclusions: Performance of information extraction systems with clinical text has improved since the last systematic review in 1995, but they are still rarely applied outside of the laboratory they have been developed in. Competitive challenges for information extraction from clinical text, along with the availability of annotated clinical text corpora, and further improvements in system performance are important factors to stimulate advances in this field and to increase the acceptance and usage of these systems in concrete clinical and biomedical research contexts.

Keywords

Electronic health record, natural language processing, information extraction, text mining, state-of-the-art review

Introduction

In the biomedical domain, the rapid adoption of Electronic Health Records (EHR) with the parallel growth of narrative data in electronic form, along with the needs for improved quality of care and reduced medical errors are both strong incentives for the development of Natural Language Processing (NLP) (sometimes called Medical Language Processing in this domain). Much of the available clinical data are in narrative form as a result of transcription of dictations, direct entry by providers, or use of speech recognition applications. This free-text form is convenient to express concepts and events, but is difficult for searching, summarization, decision-support, or statistical analysis. To reduce errors and improve quality control, coded data are required; this is where NLP, and more precisely Information Extraction (IE), is needed as explained below.

IE typically requires some "pre-processing" such as spell checking, document structure analysis, sentence splitting, tokenization, word sense disambiguation, part-of-speech tagging, and some form of parsing. Contextual features like negation, temporality, and event subject identification are crucial for accurate interpretation of the extracted information. Several different techniques can be used to extract information, from simple pattern matching to complete processing methods based on symbolic information and rules or based on statistical methods and machine learning. The information extracted can then be linked to concepts in standard terminologies and used for coding. The information can also be used for decision support and to enrich the EHR itself. Biosurveillance, biomedical research, text mining, and automatic terminology management can also benefit from information extraction. Finally, automatic de-identification of textual documents also uses the extraction of personal information before its removal or replacement. We review all these uses of information extraction in this paper.

This review focuses on research about information extraction from narrative documents stored in the EHR and published after 1995, with an emphasis on recent publications. Previous research on this topic is described in a review by Spyns [1]. Research on information extraction from the biomedical literature is not discussed in this paper, but is well described in reviews by Cohen et al. [2] and by Zweigenbaum et al. [3].

What Is Information Extraction?

IE involves extracting predefined types of information from text [4]. In contrast, information retrieval (IR) is focused on finding documents and has some very popular examples such as the Google [5] or PubMed [6] search engines. IR returns documents whereas IE

returns information or facts. IE is a specialized sub-domain of Natural Language Processing. As cited in the Encyclopedia of Artificial Intelligence, "Natural Language Processing is the formulation and investigation of computationally effective mechanisms for communication through natural language." [7] NLP research focuses on building computational models for understanding natural language. "Natural language" is used to describe any language used by human beings, to distinguish it from programming languages and data representation languages used by computers and described as "artificial." Some important domains of research are closely related to information extraction (and sometimes are confused with it), and these are explained below. Named Entity Recognition (NER) is a sub-field of information extraction and refers to the task of recognizing expressions denoting entities (i.e., Named Entities), such as diseases, drugs, or people's names, in free text documents [8]. Some entities can be identified solely through surface structure patterns (e.g., Social Security Numbers: XXX-XX-XXXX), but most of them require rules like [TITLE][PERSON] (for "Mr. Doe"), or [LOCATION], [LOCATION] (for "Salt Lake City, Utah"). Rule-based NER systems can be very effective, but require some manual effort. Machine learning approaches can successfully extract named entities but require large annotated training corpora. Advantages of machine learning approaches are that they do not require human intuition and can be retrained without reprogramming for any domain.

Text mining uses information extraction and is defined by Hearst [9] as the process of discovering and extracting knowledge from unstructured data. Text mining typically comprises two or three steps: information retrieval (to gather relevant texts; this step is not always necessary), information extraction (to extract specific types of information from texts of interest), and data mining (to find associations among the extracted pieces of information).

Clinical versus Biomedical Text

Much of what has been written on the biomedical uses of NLP can be broken down into two categories: those that focus on biomedical text and those that focus on clinical text. For our purposes here, we define biomedical text to be the kind of text that appears in books, articles, literature abstracts, posters, and so forth. Clinical texts, on the other hand, are texts written by clinicians in the clinical setting. These texts describe patients, their pathologies, their personal, social, and medical histories, findings made during interviews or during procedures, and so forth. Indeed, the term "clinical text" covers the entire gamut of narratives appearing in the patient record. These can be surprisingly short (e.g., a chief complaint) or quite long (a medical student history and physical). There is an important class of texts that arise in the clinical research setting that are rarely described in the literature. Some of these resemble biomedical texts (e.g., internal research reports) while others resemble classic clinical texts (e.g., patient notes made during a clinical trial). Since these narratives are rarely made available outside the corporate setting that generated them, formal studies of them are sparse. While we do not address these texts further here, we note that there is no a priori reason to think that the techniques tailored to either biomedical or to clinical texts would not be useful (perhaps with modification) in the realm of clinical research narratives.

What makes clinical text different from biomedical text, and why does it pose a special challenge to NLP? First, some clinical texts are ungrammatical and composed of short, telegraphic phrases. Other texts, including discharge summaries and consult reports such as radiographic readings, are often dictated and are composed deliberately for clear communication, while texts like progress notes are written mainly for documentation purposes. Second, clinical narratives are pregnant with shorthand (abbreviations, acronyms, and local dialectal shorthand phrases). These shorthand lexical units are often overloaded (i.e., the same set of letters has multiple renderings); Liu et al. estimate that acronyms are overloaded about 33% of the time and are often highly ambiguous even in context [10]. Third, misspellings abound in clinical texts, especially in notes without rich-text or spelling support. For example, the US Veterans Administration's (VA) EHR system is the largest in the world, but offers essentially only simple text support. It is not uncommon in the VA corpus to find abbreviations or acronyms themselves misspelled. Fourth, clinical narratives can contain any characters that can be typed or pasted. A common example in the VA corpus is long, pasted sets of laboratory values or vital signs. Such embedded non-text strings complicate otherwise straightforward NLP tasks like sentence segmentation, since they are usually filled with periods. Fifth, in an attempt to bring some structure and consistency to otherwise unstructured clinical narratives, templates and pseudo-tables (e.g., plain text made to look tabular by the use of white space) are common. Implicit templates, like the normative structures for a history-and-physical or a discharge summary that are commonly used across care settings can be quite useful to NLP. Explicit templates, though, are pre-formatted, highly idiosyncratic, and institution-specific with fields to be filled in by the user.

All of these issues complicate NLP on clinical text, making it especially challenging. In spite of the challenges, excellent research has been described in the literature on extracting information from clinical text. We review this work below.

A Short History of Information Extraction in the Biomedical Domain

IE has been developed mostly outside of the biomedical domain, in the Message Understanding Conferences (MUC) organized between 1987 and 1998 and sponsored by the U.S government. The MUC conferences fostered much of the work in the IE domain and consisted in competitive evaluations of systems developed for extraction of specific information such as named entities (people, organizations, locations), events, and relations (e.g. employee_of, location_of, manufacture_of). The MUC evaluation methods have been widely adopted and adapted to other domains such as biomedicine.

In the biomedical domain, IE was initially evaluated with complete NLP systems (i.e. large systems featuring all functions required to fully analyze free-text). The first of these large-scale projects was the Linguistic String Project-Medical Language Processor (LSP-MLP) [11], at New York University, enabling extraction and summarization of signs/symptoms and drug information, and identification of possible medication side effects. Inspired by this work, Friedman at al. [12] developed the MedLEE (Medical Language Extraction and Encoding system) system. This system currently is in production at the New York Presbyterian Hospital and at Columbia University. MedLEE is mainly semantically driven and is used to extract information from clinical narrative reports, to participate in an automated decision-support system, and

to allow natural language queries. MedLEE was the first biomedical NLP system to be applied to an institution different than the one where it was developed. This resulted in a small drop in performance. However, after some adjustments, MedLEE performed as well as in the original institution [13]. SPRUS (Special Purpose Radiology Understanding System) [14] was the first NLP application developed by the Medical Informatics group at the University of Utah (Salt Lake City), and was only semantically driven. Later came SymText (Symbolic Text processor) [15], with syntactic and probabilistic semantic analysis. SymText relied on Bayesian networks for semantic analysis.

The U.S. National Library of Medicine has developed a set of NLP applications called the SPECIALIST system [16], as part of the Unified Medical Language System (UMLS®) project [17]. It includes the SPECIALIST Lexicon, the Semantic Network, and the UMLS Metathesaurus®. The NLM also developed several applications that use the UMLS, such as the Lexical Tools and MetaMap [18], with many other applications described in the following sections.

The examples of complete NLP systems cited above required significant resources to develop and implement. Considering this issue, several authors progressively experimented with more simple systems focused on specific IE tasks and on a limited number of different types of information to extract. These more focused systems demonstrated good performance and now constitute the majority of the systems used for IE. This review includes all systems used for IE, complete or more focused.

Methods

This paper presents a review of recent work in information extraction from

textual clinical documents in the EHR. As mentioned previously, we only included research published after 1995. Our focus is on recent published research. We selected interesting publications from bibliographic queries in PubMed (for "information extraction", "text mining" without "information retrieval", "natural language processing" and "record" without "literature" or "Medline", "medical language processing", and "natural language understanding"), conference proceedings, and the ACM Digital Library (for "information extraction" with "medical" or "medicine" or "biomedical" or "clinical" without "literature" or "Medline"). We also added relevant publications referenced in papers that were already included.

State of the Art in Information Extraction from the EHR

Methods Used for Information Extraction

A variety of methods have been employed in the general and biomedical literature domains to extract facts from free text and fill out template slots. McNaught et al. [19] describe a detailed review of IE techniques in the biomedical domain; however their review does not include the clinical field. Here, we adopt their classification scheme with references to the clinical subdomain.

A typical IE system consists of a combination of the following components described by Hobbs [20,21]: tokenizer, sentence boundary detector, part-of-speech tagger, morphological analyzer, shallow parser, deep parser (optional), gazetteer, named entity recognizer, discourse module, template extractor, and template combiner. The performance of the higher-level components (discourse module, template extractor and template com-

biner) is highly dependent on the performance of the lower level components. The state-of-the-art of the lower-level components is discussed in a following section. Higher-level components, e.g., templates, require careful modeling for relevant attributes; with any template change, the IE system needs to be rerun to populate the modified template.

One approach to IE is pattern-matching, which exploits basic patterns over a variety of structures - text strings, part-of-speech tags, semantic pairs, and dictionary entries [22]. The main disadvantage of pattern-matching approaches is their lack of generalizability, which limits their extension to new domains. Another set of approaches is the use of shallow and full syntactic parsing. However, non-robust parser performance is an outstanding issue since medical/clinical language has different characteristics than general English. The difference between general and medical English has led to the development of sublanguage-driven approaches, which formulate and exploit a sublanguage's particular set of constraints [23-27]. The disadvantage of sublanguage approaches lies in their poor transferability to new domains. Ontology-driven IE aims at using an ontology to guide the free-text processing [28]. Syntactic and semantic parsing approaches combine the two in one processing step. Machine learning techniques have demonstrated remarkable results in the general domain and hold promise for clinical IE, but they require large, annotated corpora for training, which are both expensive and time-consuming to generate.

The general and biomedical IE communities have pushed the field towards the development of sophisticated methods for deeper, comprehensive extractions from text. Clinical - and medical - domain IE has lagged behind mainly because of limited access to shareable clinical data (e.g., constraints that pro-

tect patient confidentiality). A major challenge is the creation of a large and vibrant community around shared data, tasks, annotation guidelines, annotations, and evaluation techniques. So far, there have been three clinical shared tasks competitive evaluations on clinical texts:

(1) Automatic assignment of ICD-9-CM codes to clinical free text [29]. The shared task involved mapping ICD-9-CM codes to radiology reports. Pestian et al. [30] describe the task, its organization, and results.

(2) De-identification of discharge summaries within the i2b2 [31] initiative held in November 2006. The task is described in Uzuner et al. [32]. Top systems achieved F-measure results in the high 90's. More details are provided in the "De-identification of clinical text" section below.

(3) Patient smoking status discovery from discharge summaries within the i2b2 initiative held in November 2006 [33]. The participating systems applied a variety of techniques to assign the final patient smoking status with the top micro-averaged F-measure results in the 80's. More details are provided in the "Extraction of information in general" section below.

"Pre-processing" of Textual Documents

The vast amount of medical and clinical data available are only useful in as much as the information contained in them can be properly extracted and understood. Much work has been done recently in developing and adapting natural language tools for cleaning and processing this data for the subsequent tasks of information extraction and text- and data-mining.

One useful pre-processing task is spell-checking. Ruch et al. [34] note that the incidence of misspellings in medical records is around 10%, which is significantly higher than the misspelling incidence for other types of texts. Ruch

et al. use morpho-syntactic disambiguation tools in addition to a classical spell-checker to rank and select the best candidate for word correction. Tolentino et al. [35] create a UMLS-based spelling error correction tool. Their method performs spelling correction by detecting errors and suggesting corrections against a dictionary. They use the UMLS Specialist Lexicon as the primary source of dictionary terms and WordNet [36,37] as a secondary source. Tomanek et al. [38] examine the question of whether sentence and token splitting tools trained on general annotated corpora are adequate for medical texts. They compiled and annotated a corpus (Julie) according to a schema developed for sentence and token-splitting, which then served as a training set for a machine learning algorithm using Conditional Random Fields. The results indicate that for sentence splitting, the training corpus is not very critical; for tokenization, however, performance is significantly improved when training on a domain-specific corpus.

Word Sense Disambiguation (WSD) is the process of understanding which sense of a word (from a set of candidates) is being used in a particular context. WSD is a crucial task for applications that aim to extract information from text. Weeber et al. [39], at the National Library of Medicine, derived a corpus of MEDLINE abstracts and manually sense-tagged 5,000 instances of 50 ambiguous words using the UMLS as sense inventory. Liu et al. [10] present a very good background section on general English WSD, biomedical WSD and supervised approaches to the task. They avoid the use of manually annotated sense-tagged data by using a two-step unsupervised approach. They automatically derive a sense-tagged corpus from MEDLINE abstracts using the knowledge in the UMLS, and use the derived sense-

tagged corpus as a training set for a classifier for ambiguous words. Liu et al. [40] implement four machine learning algorithms and use three datasets in a study showing that supervised WSD is suitable when there is enough sense-tagged data. All supervised WSD classifiers performed with a precision of less than 80% for biomedical terms while most classifiers achieved around 90% for general English. There was no single combination of feature representation, window size, or algorithm that performed best for all ambiguous words. Xu et al. [41] also investigated the effects of sample size, sense distribution, and degree of difficulty on the performance of WSD classifiers. Pakhomov et al. [42] focus on WSD in the clinical domain and experiment with abbreviation and acronym disambiguation by applying a combination of supervised and unsupervised methods. Coden et al. [43] use a supervised method to train a classifier for the top 50 ambiguities from a clinical corpus compiled from Mayo Clinic notes.

Part-of-speech tag assignment has a major impact on the natural language tasks that follow. According to Campbell et al. [44] an error of 4% in part-of-speech-tag assignment can translate to 10% in error rate at the sentence level. Part-of-speech taggers for general English achieve very high scores in the task. Coden et al. [45] suggest two ways to adapt a part-of-speech tagger (POS-tagger) trained in general English texts to the clinical language: by adding a 500-word domain-specific lexicon; and by creating manual annotations on domain-specific documents and adding these documents to the English corpus. The addition of the annotated documents increased the tagger's performance by 6% to 10%, whereas the addition of the lexicon increased its performance by about 2%. The authors noted however that the creation of the

lexicon required much less effort than manual annotations. Liu et al. [46] developed a manually annotated corpus of pathology reports and a domain specific lexicon to evaluate the performance of a maximum-entropy POS-tagger trained on general English. The POS-tagger re-trained with the annotated corpus performed better than with the lexicon addition. The study also showed that more than 30% of the words in the pathology reports were unknown to the general English trained tagger. The addition of an 800-word domain-specific lexicon revealed a performance increase of 5% and selecting sentences that contained the most frequent unknown words proved to be most helpful. Hahn et al. [47] investigated the use of a rule-based POS-tagger (Brill tagger) and a statistical tagger (TnT) on clinical data. The statistical tagger TnT trained on general texts performed close to the state of the art in the medical domain. They claim that the model (statistical vs. rule-based) is more important than the sublanguage. Nonetheless, the statistical tagger improved its performance substantially when trained on medical data.

Parsers generate a constituent tree that provides a syntactic representation of the sentence structure with its dependencies. Medical language is especially challenging because of its ungrammatical and fragmented constructions. Campbell et al. [48] argue in favor of dependency grammars (DG) for biomedical text exactly because of the ungrammaticality of many sentences. In DG, each word has one attachment only and a tree structure with the dependencies is the sentence representation. The authors applied a Transformational Based Learning algorithm to learn a dependency grammar for medical texts. Clegg et al. [49] present a method for evaluating parsers' performance using an intermediate representation based on dependency graphs.

They evaluated Bikel, Collins, Stanford, Charniak, and Charniak-Lease parsers and mapped the constituent parsed trees to dependency graphs. Bikel and Charniak-Lease parser performed well on parsing sentences from the Genia Treebank (also mapped to dependency graphs). Pyysalo et al. [50] investigated the adaptation of a Link Grammar parser to the biomedical language with a focus on unknown words.

Contextual Feature Detection and Analysis

When extracting information from narrative text documents, the context of the concepts extracted plays a critical role. Important contextual information includes negation (e.g. "denies any chest pain"), temporality (e.g. "...fracture of the tibia 2 years ago..."), and the event subject identification (e.g. "his mother has diabetes").

NLP systems such as the LSP [11] or MedLEE [12] include negation analysis in their processing, but research focused explicitly on negation detection started only a few years ago with NegExpander [51], a program detecting negation terms and then expanding (NegExpanding) the related concepts. This program had a precision of 93% and was used by a mammography reports classification algorithm. More recently, a negation detection algorithm called NegEx was developed using regular expressions [52] and achieved 94.5% specificity and 77.8% sensitivity. Several systems later implemented NegEx, such as the system developed by Mitchell et al. [53] to extract information from pathology reports in the Shared Pathology Informatics Network (SPIN). When only evaluating negation detection, they measured a precision of 77% and a recall of 83%. In the process of developing NLP tools

for the i2b2 (Informatics for Integrating Biology and the Bedside) project, Goryachev et al. [54] compared NegEx, NegExpander, and two classification-based algorithms, and measured the best performance with NegEx (94.5% sensitivity and 94.3% specificity).

A more complex system, called Negfinder [55], also used indexed concepts using the UMLS and regular expressions, but added a parser using a LALR (Look-Ahead Left-Recursive) grammar to identify negations, and achieved 97.7% specificity and 95.3% sensitivity when analyzing surgical notes and discharge summaries.

A system extracting SNOMED-CT concepts from History and Physical Examination reports at the Mayo Clinic implemented a negation detection algorithm based on an ontology for negation. They measured a 97.2% sensitivity and 98.8% specificity [56]. The most recently published negation detection algorithm used a hybrid approach based on regular expressions and grammatical parsing [57]. Negation terms were detected using regular expressions to achieve high sensitivity, and the part-of-speech parse tree was then traversed to locate negated phrases with high specificity. When evaluating negation detection on radiology reports, a 92.6% sensitivity and a 99.8% specificity were measured.

Temporality analysis in clinical narrative text can be significantly more complex than negation analysis, and has been investigated by Zhou, Hripcsak and colleagues, starting by proposing a model for temporal information based on a simple temporal constraint satisfaction problem [58]. Discharge summaries were analyzed for temporal structures and a temporal constraint structure for historical events was developed and then applied to other discharge summaries. The temporal constraint structure successfully modeled 97% of the temporal expressions [59]. The authors then pro-posed a system for automated temporal information extraction based on a temporal tagger, an NLP system (MedLEE), some post-processing based on medical and linguistic knowledge to treat implicit temporal information and uncertainty, and the simple temporal constraint satisfaction problem for temporal reasoning [60]. This system, called TimeText, has recently been evaluated with discharge summaries [61]. TimeText detected clinically important temporal relations with 93.2% recall and 96.9% correctness. It also answered clinically plausible temporal queries with 83.7% accuracy.

Harkema et al. have developed temporal analysis in the context of the CLEF (Clinical eScience Framework) IE component [62]. The information extracted is used to build the patient chronicle, an overview of the significant events in the patient's medical history. Events extracted from narrative reports are associated with structured data from the EHR. The system still includes some manual steps, but the authors are working on a fully automatic system. Focusing on discharge summaries, Bramsen et al. analyzed temporal segments (i.e., a fragment of text that does not exhibit abrupt changes in temporal focus), a coarser level of analysis than Zhou et al., and their ordering to characterize the temporal flow of discourse [63]. The authors use machine learning techniques for automatic temporal segmentation and segment ordering. For temporal segmentation, they use lexical, topical, positional, and syntactic features and measured 78% recall and 89% precision. The best results for segments ordering were obtained with the Integer Linear Programming framework (84.3% accuracy) [64]. Finally, algorithms combining the analysis of the subject of the text (e.g., the patient) and other contextual features have recently been developed and evaluated. As a first step towards automated extraction of contextual features,

Chu et al. [65] have manually annotated four contextual features for 56 clinical conditions detected in ED reports. These features - Validity (valid/invalid), Certainty (absolute, high, moderate, low), Directionality (affirmed, negated, resolved), and Temporality (recent, during visit, historical) - were then evaluated in terms of their contribution to the classification of the detected conditions as acute, chronic or resolved. Directionality (i.e. negation) was the most important contextual feature. Chapman et al. [66] propose an algorithm for contextual features identification. This algorithm, called ConText, is an extension of NegEx cited above. ConText determines the values of three contextual features: Negation (negated, affirmed), Temporality (historical, recent, hypothetical), and Experiencer (patient, other). Like NegEx, this algorithm uses regular expressions to detect trigger terms, pseudo-trigger terms, and scope termination terms, and then attributes the detected context to concepts between the trigger terms and the end of the sentence or a scope termination term. The evaluation of ConText used an NLP-assisted review methodology described by Meystre et al. [67] and measured a 97% recall and precision for negation, 50% recall and 100% precision for experiencer, and 67.4% to 82.5% recall and 74.2% to 94.3% precision for temporality (when assigning historical or hypothetical values).

Some conclusions that can be drawn from this research are that separate algorithms (i.e., specialized in contextual features analysis) are easier to implement, and one of the best performing negation detection algorithms - NegEx - is a good example of this. Most of these algorithms are based on lexical information, even if some algorithms add part-of-speech information like ChartIndex cited below.

Extraction of Information in General

In this section, we present published research involving extraction of information from textual documents in the EHR and emphasis the different methods used and types of documents analyzed. The application of IE for specific purposes such as coding, surveillance, terminology management, or research are described in the subsequent sections.

Useful IE has been attempted with basic pattern matching techniques such as regular expressions. Dictionaries of variable size were also typically used. Long [68] has extracted diagnoses from discharge summaries using regular expressions and a punctuation marks dictionary as well as the UMLS Metathesaurus [69]. To extract blood pressure and antihypertensive treatment intensification information, Turchin [70] also used regular expressions. The REgenstrief eXtraction tool [71] uses pattern matching to extract some diagnoses and radiological findings related to congestive heart failure. Finally, a module extracting UMLS concepts and based on pattern matching was developed by Bashyam [72]. It was faster than MMTx [73] and was proposed as a new option in MMTx.

Systems based on full or partial parsing were based on morpho-semantems (i.e., elementary meaningful units that compose words and are prefixes like oto- or suffixes like -itis) [74], on lexical and/or syntactic information such as the LifeCode® system [75], or the Dutch Medical Language Processor developed by Spyns et al. [76], or on semantic information like the application developed for the ChartIndex project to convert documents to the HL7 CDA [77] format and extract UMLS concepts [78].

Approaches combining syntactic and semantic analysis constitute the majority of the systems. A famous system that has been adapted and used for several

different tasks is MedLEE [79]. Besides being progressively extended to most of the documents present in the EHR [24] and tested for its transferability in another institution [13], it has been used to detect findings evocative of breast cancer [80], to analyze modifications to data entry templates [81], and even combined with machine translation to detect abnormal findings and devices in Portuguese radiology reports [82]. MetaMap and its Java™ version called MMTx (MetaMap Transfer) were also often used to extract information from clinical documents, even it they have been developed for MEDLINE abstracts and lack negation detection. Some examples are Schadow et al. [83] who used it to extract information from pathology reports, Chung et al. [84] who used it with echocardiography reports, and Meystre et al. [67] who used it to extract medical problems.

SymText [15] and its successor, MPLUS [85], make extensive use of semantic networks for semantic analysis. These networks are implemented as Bayesian networks (also called belief networks), trained to infer probabilistic relationships between extracted terms and their meaning. They have been used to extract interpretations of lung scans [86], to detect pneumonia [87], and to detect mentions of central venous catheters [88]. Other systems combining syntactic and semantic analysis have recently been developed and evaluated. The Pittsburgh SPIN information extraction system [27] was a project of the Shared Pathology Informatics Network (SPIN) [89] based on GATE (General Architecture for Text Engineering) [90] and evaluated to extract specific information from pathology reports. A very similar application - caTIES (cancer Text Information Extraction System) [23] - was later developed by the same team as a caBIG-compliant [91] application. It is based on the NCI Enterprise Vocabulary Sys-

tem [92] instead of the UMLS utilized by the SPIN system. HITEx (Health Information Text Extraction) was also based on GATE and was developed to extract diagnoses and smoking status [93]. Finally, the KnowledgeMap Concept Identifier (KMCI) was adapted to extract UMLS concepts from echocardiography reports [94] and to detect QT interval prolongations [95].

Recent systems are almost always based on some machine learning methods, for limited tasks or for most of their functions. An example is a system developed by Taira et al. [96] that used Maximum Entropy classifiers for parsing and semantic analysis, and later also a vector space model to extract UMLS concepts [97]. Another example is the semantic category classifier developed by Sibanda et al. [98]. It employs support vector machines to attribute semantic categories to each word in discharge summaries.

Systems developed to extract information from textual documents in the EHR have mostly been focused on chest radiography reports [13,71,75,82, 88,93,96,99,100]. They have also been developed to analyze other types of radiology reports [72,78,80,85,86,97], echocardiogram reports [84,94,95], and other types of documents that have more diversity and larger vocabularies such as discharge summaries [68, 93,98], pathology reports [26,27,83], and other notes [70,81,101].

Some systems have been developed to analyze several different types of documents [15,24,67], and the effort required to port an NLP application from only chest radiography report, to other radiology reports, discharge summaries, and pathology reports is well described by Friedman [24,102]. The largest efforts to develop and evaluate information extraction from clinical text have been achieved in the context of the i2b2 smoking status identification challenge in 2006 and the Medical NLP challenge

[30] in 2007 and described in the next section. For the i2b2 "smoking challenge", a corpus of 502 de-identified and "re-identified" (with realistic surrogates) discharge summaries was first created by Uzuner et al. [33]. Eleven teams participated. Their task was to use discharge summaries to classify each patient as a smoker, current smoker, past smoker, non-smoker, or unknown. The best performing system was developed by Clark et al. [103] and first filtered out documents with unknown smoking status, and then used SVMs (Support Vector Machines) to classify the smoking status. They also added 1200 documents to improve the training of their system. The overall accuracy of their system reached 93.6%. Some other good performing systems are described in Cohen [104], Heinz et al. [105], Savova et al. [106], and Wicentowski and Sydes [107].

Extracting Codes from Clinical Text

A popular approach in the literature over the last several years has been to use NLP to extract codes mapped to controlled sources from text. The most common theme was to extract codes dealing with diagnoses, such as International Classification of Diseases (ICD) versions 9 and 10 codes. In addition to a focus on systematic coding schemes like ICD-9, institutions often also have local coding schemes they wish to extract.

2007 was a particularly interesting year for this because it was the year of the Medical NLP challenge, a shared task exercise that provided a moderately large test/training corpus of radiological reports and their corresponding human-coded ICD-9-CM codes. The project is described in Pestian et al. [30] and was very well conceived - especially the evaluation metrics. Ultimately, 44 different research teams par-

ticipated, an astounding number. We found two papers that reported techniques and results. Aronson et al. [108] leveraged several existing technologies (e.g., NLM's Medical Text Indexer, a support vector machine classifier, a k-NN classifier, etc.) and arranged them in a stack-like architecture to evaluate their fused performance. They placed 11th in the challenge with a mean F-measure considerably higher than the average score for all participants (F-measure = 0.85; the best score was 0.89; mean score was 0.77). Crammer et al. [109] also described a multi-component coding system; it used machine learning, a rule-based system, and an automatic coding system based on human coding policies. They judged these to be loosely orthogonal so they combined the results in a cascade that gave priority to the human coding policy approach. They placed fourth in the challenge and in this paper describe the same technology's performance against a local corpus of radiology reports. ICD-10, the newer ICD standard, is more popular overseas than in the US, so it is not surprising that the literature describing automatic extraction of these codes comes mainly from Europe and Japan. Baud et al. [110] detail an interesting overview of the problems inherent in the task of ICD-10 encoding. And in a vein similar to the ICD-9 approaches above, Aramaki et al. [111] use a multi-component approach using three different extraction algorithms followed by a polling technique at the end to determine the winner. A consistent theme with all these recent NLP-based code extractors for ICDs is the use of multiple, parallel components followed by some sort of adjudication module.

The past decade has seen the ascendancy of a remarkable general-purpose information extraction tool for clinical texts. As noted above, it is called MedLEE and its use as a code extractor is well

summarized in Friedman et al. [25] MedLEE has seen use in code extraction in many contexts. Friedman herself describes an automated pneumonia severity score coding system using it [112]. Elkins et al. [113] describe an adaptation of its use for neuroradiology standard concept extraction; Kukafka et al. [114] used it to code to a standard for health, and health-related states, called the International Classification of Functioning, Disability, And Health (ICF; also a WHO standard). Lussier et al. [115] have applied MedLEE to extract SNOMED codes. SNOMED was also the driver for work done by Hasman et al. [116] They have exploited SNOMED coding in clinical text NLP, primarily to assist pathologists during the coding process.

In addition to extracting codes that conform to a standard coding scheme like ICD-9/10 or SNOMED, there is considerable interest in extracting codes from text that conform to a local institutional standard like a problem list. Pakhomov et al. [117] and Haug et al. [118] describe examples of problem-list extraction at the Mayo Clinic and Intermountain Healthcare, respectively. These are two mature centers for clinical informatics. The Pakhomov system uses a multi-pass, certainty-based approach; while Haug's efforts use a Bayesian belief network technology. That team's work built on the work presented in Gundersen et al., [119] which makes a convincing case for the superiority of just-in-time automated coding over static, pre-coded systems.

Extracting Information to Enrich the EHR and for Decision Support

The past dozen years has seen an increase of interest in using NLP for enriching the content and utility of the EHR, especially to support computerized decision-making. We have catego-

rized this work into four broad groups, but the boundaries between them are fluid, as is often the case in NLP.

In contrast to work done in the early 1990s, recent work on the automatic structuring of documents using NLP has been on the wane. Kashyap et al. [120] used a commercial product called A-Life© to automatically structure standard admission notes such as the history and physical. They reached the same conclusion common to similar efforts in the past, namely that NLP technology is not ready yet to completely structure these texts. They argue that, given the volume of admission data, even partial support is a worthy goal. The VA CPRS system that was noted in the Introduction provides an interesting and large-scale platform for research; Lovis et al. [121] used a handcrafted parser to assist in the structuring of computerized provider order entry fields. While the parser itself is limited to use within CPRS, the study is important because it was the first to show that NLP could be used successfully within the CPRS environment. See also the section below on research uses of NLP, which describes a note structuring system in use at the Mayo Clinic.

As clinical text systems have grown in popularity, a problem revealed itself: the sheer number of notes from so many diverse disciplines being integrated into one spot makes navigation through them all quite difficult. Two interesting papers reported on the use of NLP to make navigation easier through visualization of notes. Cimino et al. [122] recently described work that successfully abstracted and summarized medication data in an effort to improve patient safety. Liu and Friedman [123] demonstrated that a tool they call CliniViewer, built using MedLEE and an XML engine, can be used to summarize and navigate clinical text. As clinical text modules in the EHR become more popu-

lar, it is likely that we will see an increase in research in this area.

Another way to enrich the value of the EHR using information extraction is case finding. In this setting the goal is to find patients that match certain criteria based on either text alone or text in conjunction with other EHR data. Day et al. [124] used the MPLUS NLP system to classify trauma patients, and the system did well enough that it is in use daily at a Level 1 trauma center. Mendonça's team [125] used MedLEE to identify pneumonia in newborns with a very reasonable F-measure. Community acquired pneumonia (CAP) is a very common problem in healthcare today and it has been the focus of several NLP efforts. Fiszman et al. [126] showed how SymText could be used to find cases of CAP by comparing clinical notes to the CAP clinical guidelines. Using a similar technical approach, Aronsky et al. combined the same NLP system with a Bayesian network to identify general pneumonia. Finally, again using MedLEE, Jain et al. [127] demonstrated a very impressive F-measure and finding cases of tuberculosis in radiographic reports.

Beyond the three categories above, the use of NLP to enrich the EHR and to support decision-making is quite diverse. Representative of that work are examples such as: Meystre and Haug using MMTx, combined with Chapman's NegEx algorithm, to enrich the problem list [67,128,129]. In 2005 Hazelhurst et al. [130] described their MediClass NLP system which is an interesting combination of knowledge-based and NLP-based techniques. They demonstrate its utility in the automatic discovery of vaccination reactions from clinical notes and in assessing adherence to tobacco cessation guidelines [131,132].

In 1996, Johnson and Friedman [133] noted a caution: the performance of any

NLP system is constrained by the quality of the human-composed text. They showed that even the most basic of information, demographics, are often inconsistently entered by humans. They compared the demographic data in discharge summaries as extracted by an early prototype of MedLEE to the original data input by humans at admission. The NLP system performed quite well at extracting the demographics, while the demographics input by humans was quite inconsistent. As clinical text repositories grow, they note, the repositories will increasingly be filled with conflicting data, posing a challenge to any NLP system.

Information Extraction for Surveillance

One of the great benefits of computing in general is the ability of a computer to do mundane, repetitive tasks where humans have a hard time maintaining vigilance. Surveillance based on clinical texts is precisely such a task, at the same time both very important and profoundly tedious. Adverse events surveillance based on clinical texts is a good example. Penz et al. [134] used MedLEE to test the feasibility of mounting surveillance for adverse events related to central venous catheters, using surgical operation reports from the VA's CPRS. Their specificity was about 0.80 and their sensitivity was about 0.72. Error analysis showed that errors were due to the difficulties of processing raw clinical text using a standard parser (coupled with inadequate provider documentation). Interestingly, they found that the corresponding administrative data for detected catheter placements (e.g., ICD-9 codes) only captured about 11% of the use of these devices, showing that the text was a far better place to look for catheter placement information than billing data. Melton and Hripcsak [135] used

MedLEE to mount a surveillance for a broad range of adverse events. While the sensitivity for their technique was low at 0.28, the specificity was quite high at 0.99. Cao et al. [136] used a straightforward keyword-based NLP approach for the surveillance of adverse drug events, but found only modest positive predictive value. Both of Melton and Cao studies used discharge summaries, which are fairly clean clinical documents.

Syndromic surveillance has become a popular area of research, especially with growing concerns about national security and pandemic issues. Chapman et al. [137] reported on a system using MPLUS to conduct biosurveillance of chief complaint text fragments. The system's performance was good enough that it was used in the Winter Olympic Games in 2002. Pneumonia outbreaks are an important clinical surveillance issue, so Haas et al. [138] extracted information from neonatal chest x-ray reports using MedLEE. The positive predictive value of the system was 0.79 but the negative predictive value was greater than 0.99. The NLM's MetaMap tool was used by Chapman et al. [139] for the biosurveillance of general respiratory findings in the emergency department. The results were moderately low with an F-measure in the mid-60s. Their error analysis allowed the research team to identify areas that would improve MetaMap's performance, and these are very likely to be applicable to any concept extractor using emergency department clinical text (e.g., temporal discrimination, anatomic location discrimination, finding-disease pair discrimination, and in contextual inference). In another study, Chapman [140] did show that surveillance for fever, as a biosurveillance indicator, could be readily accomplished using keywords and a probabilistic algorithm called CoCo to infer fever from chief com-

plaint, as well as keyword searching in dictated ED histories and physicals. There are several efforts underway within the VA system to use NLP in quality surveillance. A representative study of this work is that by Brown et al. [141]. They used pattern-matching techniques to extract information from an electronic quality (eQuality) assessment form used within the VA system. They reported a sensitivity of 0.87 and a specificity of 0.71, and the note that for sensitivity human performance was only 4 to 6% better.

Information Extraction Supporting Research

Under the stewardship of the NIH Roadmap project called the Clinical and Translational Science Award (CTSA) process, translational research is booming, along with translational informatics research. The first CTSA awards were made in 2007 and we anticipate that research-oriented NLP studies will soon be appearing. For now, the application of NLP to information extraction from clinical texts to support research is a comparatively small body of work.

By far, the most common use of NLP in this context is in subject recruitment, where textual data is used to identify patients who may benefit from being in a study. Pakhomov et al. adapted the text analysis system created at the Mayo Clinic for the structuring of semi-structured notes for use in identifying patients with angina [142,143] and heart failure [144]. In both domains, the NLP system improved ICD-9 based subject searching. Once the texts of interest were structured, they used keyword searches on now-mapped conceptual entities to identify the patients of interest. His two papers that appeared in 2007 are especially interesting because they appeared in medical journals, not informatics journals. This re-

flects a growing acceptance of informatics research into the mainstream medical literature. Medical journals often require a more rigorous clinical evaluation of informatics tools such as Pakhomov's, and it is refreshing to see informatics tools compared in a rigorous statistical way to quantitative and qualitative health services research techniques. Xu et al., [26] using MedLEE, extracted subject eligibility data from surgical pathology reports. These reports often present structural processing barriers for MedLEE, so the team designed a preprocessor that was tailored to emphasize eligibility data.

An interesting use of statistical NLP to support research is presented by Niu et al. [145]. They used classic n-gram techniques coupled with machine learning and negation to try to discern the "polarity" of sentences in the journal Clinical Evidence, which summarizes recent findings in the clinical literature. In this sense, the polarity refers to whether the outcome was "positive," "negative," "neutral," or "no outcome reported." Their average F-measure for each was in the high 80s, with the best performance on positive outcomes. This approach could be used to assist clinicians in automatic question answering or to locate studies that are pertinent to their research.

De-identification of Clinical Text

In the United States, the HIPAA (Health Insurance Portability and Accountability Act, codified as 45 CFR §160 and 164) protects the confidentiality of patient data, and the Common Rule (codified as 45 CFR §46) protects the confidentiality of research subjects. The European Union Data Protection Directive provides similar confidentiality protection. These laws typically require the informed consent of the patient and approval of the Institutional Review

Board (IRB) to use data for research purposes, but these requirements are waived if data are de-identified. Anonymization and de-identification are often used interchangeably, but de-identification only means that explicit identifiers are hidden or removed, when anonymization implies that the data cannot be linked to identify the patient (i.e., de-identified is often far from anonymous). Scrubbing is also sometimes used as a synonym of de-identification.

For a narrative text document to be considered de-identified, the HIPAA "Safe Harbor" technique requires 18 data elements (called PHI: Protected Health Information) to be removed, such as names, telephone numbers, addresses, dates, and identifying numbers. Dorr et al. [146] have evaluated the time cost to manually de-identify narrative text notes (average of 87.2 ± 61 seconds per note), and concluded that it was time-consuming and difficult to exclude all PHI required by HIPAA.

Already well aware of these issues, several authors have investigated automated de-identification of narrative text documents from the EHR. Sweeney developed the Scrub system [147] to hide personally identifying information (names, contact information, identifying numbers, age, etc.). Each specific entity was detected by a specific algorithm using a list of all possible values (e.g., an algorithm detected first names and used a list of all commonly known first names). This system found 99-100% of identifying information.

Ruch et al. [148] adapted a system build for disambiguation, the MEDTAG system, to detect and replace all instances of titles and names. They used the MEDTAG lexicon to tag semantic types, and manually written disambiguation rules. The system was evaluated with mostly French surgery reports, laboratory results, and dis-

charge summaries, and successfully removed about 99% of the identifiers. To detect proper names only, two different approaches have been reported. Taira et al. [149] trained a system with a corpus of annotated reports from pediatric patients. A lexical analyzer attributed syntactic and semantic tags to each token, and obvious non-patient names (drug names, institutions, devices, etc.) were removed. A maximum entropy model was then used to determine the probability that a token can take the PATIENT role. With a decision threshold of 0.55, a 99.2% precision and a 93.9% recall were measured. Thomas et al. [150] used the property of names to usually occur in pairs or be preceded or followed by affixes (e.g. Dr, MD) to detect and replace them in the narrative section of pathology reports. With a list of clinical and common usage words, and a list of proper names, they correctly identified 98.7% of the proper names.

The Concept-Match scrubbing algorithm was developed by Berman [151] and took a radical approach to de-identify pathology reports: all phrases that could be matched with UMLS concepts were replaced by the corresponding code (CUI) and another synonym mapping to the same code, and all other words (except stop words) were replaced by asterisks. The algorithm was fast but was not formally evaluated. Fielstein et al. [152] have evaluated an algorithm using regular expressions and a city list to remove PHI as defined by HIPAA (except photographic images) and achieved a 92% sensitivity and a 99.9% specificity. The De-Id system was developed to remove all PHI from narrative clinical reports [153]. It used rules and dictionaries that were incrementally improved to finally miss some identifiers in only 3.4% of the reports. Unlike all other system described, De-Id keeps an encrypted linkage file ty-

ing the de-identified document to the suppressed identifiers.

Beckwith at al. [154] have developed an open source system removing PHI from pathology reports and called HMS Scrubber. This system first removed all identifying information from the header of the reports that were also found in the body of the report. It then used 50 regular expressions to detect and remove dates, addresses, accession numbers, and names cited with markers such as Dr, MD, PhD, etc. Finally, it used two freely available lists of names (90,000 unique first and last names from the 1990 US census) and of locations (16,000 unique cities, towns, etc. from the US Census Bureau). When evaluated, this system removed 98.3% of the PHI present in 1800 pathology reports from the SPIN (Shared Pathology Informatics Network).

The largest effort to develop and evaluate automated de-identification has been achieved in the context of the i2b2 de-identification challenge in 2006. Uzuner et al. [32] have first created a corpus of 889 de-identified and "re-identified" (with realistic surrogates) discharge summaries. Identifying information was first tagged using statistical Named Entity Recognition techniques. This system was based on SVMs using local context (mostly lexical features and part-of-speech) and a few dictionaries (names, locations, hospitals and months). It was compared to other systems and achieved the best performance with 95% recall and 97.5% precision [155]. A manual verification of the de-identified documents was then executed, followed by the replacement of this information with realistic surrogates and the addition of some ambiguity and randomly generated surrogate PHIs. This corpus, with tagged PHI, was then made available to the seven teams who participated in the challenge. About 3/4 of the corpus

was made available for training, and then the remaining 1/4 was used for testing. The systems developed and submitted for testing by the teams had to remove names of patients, doctors, hospitals, and locations, as well as identification numbers, dates, phone numbers and ages above 90. The best systems were developed by Wellner et al. [156], and by Szarvas et al. [157]. The best system developed by Wellner et al. was based on Carafe, a toolkit implementing Conditional Random Fields developed at the MITRE Corporation (Bedford, MA). This system tagged each token as part of a PHI phrase or not, and also included some regular expressions to detect phone numbers, zip codes, addresses, etc. and a lexicon of US state names, months, and English words. It reached a 97.5% recall and 99.22% precision (F-measure of 98.35%). The system developed by Szarvas et al. used local context, regular expressions (for ages, dates, identification numbers, and phone numbers), and dictionaries (first names, US locations, names of countries, and names of diseases). They then used decision tree algorithms (C4.5 and Boosting) to classify each word as PHI or non-PHI. Their system reached a 96.4% recall and a 98.9% precision (F-measure of 97.6%). In general, methods based on dictionaries performed better with PHI that is rarely mentioned in clinical text, but are difficult to generalize. Methods based on machine learning tend to perform better but require annotated corpora for training.

Automatic Terminology Management

Terminologies, lists of vetted terms for a given domain, and ontologies, the relational organization of the vetted terms, are critical for a number of clinical domain applications - concept-based information retrieval, decision-support systems, autocoding among many - to

ensure system interoperability. The traditional method for building them relies on experts to identify the terms and create the hierarchy, a process which is time-consuming and which requires the collaborative effort of domain specialists. Here, we focus on summarizing the field as applied to the clinical domain. For a comprehensive review of the topic as related to the entire field of biomedicine, its methods and terminological resources, consult [158]. They outline the general steps for automatic terminology management: (1) automatic term recognition, (2) term variants augmentation, (3) automatic term structuring.

The most recent advances in automatic terminology management in the clinical domain are represented by systems that employ the combination of NLP techniques for term discovery and lexico-syntactic patterns for semantic relation discovery along with visualization tools. Baneyx et al. [159,160] investigation focuses on building an ontology of pulmonary diseases. Zhou et al. [161] experiment with surgical pathology reports, while Charlet et al. [162] work is in the surgical intensive care domain. Kolesa and Preckova [163] tackle an additional complexity - that of a semi-automated, NLP-based localization of international biomedical ontologies, in their case a Czech drug ontology seeded with terms discovered from drug information leaflets. All of them successfully demonstrate the use of NLP and IE techniques in the full-circle process of terminology discovery and ontology building.

A number of other efforts describe approaches to the subtasks in the process of automatic terminology management. Hersh et al. [164] is one of the first investigations combining NLP techniques for the task of candidate term discovery and terminology expansion, which they test on all EHR narrative reports at the Oregon Health Sciences

University and the Portland Veterans Administration Medical Center through February 1995. Harris et al. [165] and Savova et al. [166] investigate a method for term candidate discovery for the domain of patient functioning, disability and health and later apply lexico-syntactic patterns and latent semantic analysis to induce structure for the candidate terms [167]. Do Amaral et al. [168] use radiology reports to apply NLP techniques to abstract the reports' general framework and discover the reports' semantic template. Friedman et al. [169] describe their controlled vocabulary development tool, which displays candidate terms along with usage statistics obtained from a corpus, their compositional structure, and suggested ontology mappings.

A number of vocabulary servers are available for the biomedical domain to support terminology management - UMLS knowledge source server [170], LexGrid [171], Metaphrase [172], Medical Entities Dictionary (MED) [173]. All of them are Web-based interfaces that take as input a user specified term and return ontological mappings.

Clinical Text Corpora and their Annotation

The use of automatic information extraction and retrieval tools depends heavily on the quality of the annotated corpora available for their training and testing. Currently, much work is being done on developing guidelines for corpus annotation, identifying relevant features to annotate, and on the characterization of what makes a particular corpus usable.

Chapman et al. [174] present an annotation schema to manually annotate clinical conditions. The schema was developed based on 40 emergency department reports and tested on 20 such reports. The two authors acted as anno-

tators and achieved a high agreement and an F-measure of 93%. They point out that there are no standard guidelines determining which words to include in the annotation of clinical texts; thus their proposal focuses on which semantic categories and words are important to include in such annotations. They suggest that similar methodology can be used to develop principled guidelines for other clinical text annotation. In a follow-up investigation, Chapman et al.[175] examined the improvement in agreement among annotators after they were trained with the annotation schema. For this investigation, three physicians and three lay people were used as annotators and concluded that physicians presented a higher agreement after training on the schema than when applying a baseline one-hour training; moreover, lay people performed almost as well as physicians when trained on the schema. These results suggest that good annotation guidelines are essential to good annotation quality, especially when the annotators are not domain experts.

Cohen et al. [176] examine six available corpora with respect to their design characteristics to determine which features may be responsible for their high or low usage rates by external systems. Their conclusion is that semantic annotation, standard formats for annotation and distribution, and high-quality annotation of structural and linguistic characteristics are relevant features and good predictors of usage. Cohen et al. [177] analyze in further detail corpus design characteristics and suggest that good documentation, balanced representation, the ability to recover the original text, and data on inter-annotator agreement are the main characteristics to promote a high-level use of a corpus. Wilbur et al. [178] discuss what properties make a text useful for data-mining applications. They identified 5 qualitative dimensions: focus, polarity,

certainty, evidence, and directionality and developed guidelines on how to annotate sentence fragments along these five dimensions. The guidelines were developed over a one-year period through multiple iterations of testing and revision. Results of 12 annotators on 101 sentences from biomedical periodicals are reported between 70-80%. This methodology and guidelines are being used to annotate a large corpus of 10,000 sentences to serve as training corpus for automated classifiers. An interesting point is that the difficulty of the annotation varies considerably depending on the dimension being annotated, with rating of the evidence being one of the most challenging tasks. Liu et al. [27] study and review the types of error a system that automatically extracts information from pathology reports makes. The information extracted was compared to a manually annotated gold standard. The authors classified the errors into: 1) system errors and 2) semantic disagreement between the report and the annotation. This second point shows that even when gold standard annotations are available they may still be difficult to interpret and automatic extraction may be more valid for some variables than for other.

Clinical Text Mining

Ananiadou and McNaught [8] and Hirschman and Blaschke [179] provide an extensive overview of the state-of-the-art of text mining and its challenges in biomedicine. In our review here, we focus on text mining in the clinical domain. We adhere to the widely-accepted definition of text mining by Hearst [9] and used in Ananiadou and McNaught [8] - the discovery and extraction of new knowledge from unstructured data - and contrast it with data mining, which finds patterns from structured data, and with information

extraction, which extracts the known facts from text and presents them in a structured form. The inspiration for text mining comes from the pioneering work of Swanson [180] in which he brilliantly demonstrates that chaining facts from disparate literature sources could lead to the generation of new scientific hypotheses.

Biomedical text mining has been primarily explored in relation to literature, the main reasons being the confidentiality provisions that govern patient clinical records and the limited number of investigators with access to such data. Clinical text mining has been investigated for finding association patterns. Chen et al. [181] employ text mining and statistical techniques to identify disease-drug associations in the biomedical literature and discharge summaries and conclude that there are distinct patterns in the drug usage as reported in the literature and as recorded in the patient record. Cao et al. [182] explore the automatic calibration of the statistic value and apply it for the discovery of disease-findings associations. In another study, Cao et al. [183] show that statistical methods are successful in finding strong disease-finding relations. Their use-case was a knowledge base construction for the patient problem list generation. Rindflesch et al. [184] use statistical methods to construct a database of drug-disorder co-occurrences from a large collection of clinical notes from the Mayo Clinic.

Conclusions and Future Challenges

In this paper, we reviewed the advances of information extraction from free-text EHR documents. IE is still a relatively new field of research in the biomedical domain, and the extraction of

information from clinical text is even newer. Compared to the IE tasks of the Message Understanding Conferences, results in clinical text IE were often mixed. Reasons proposed for this difference are that more experience is needed, annotated text corpora are rare and small, and clinical text is simply harder to analyze than biomedical literature, or even newswires. During the last several years, performance has gradually improved, exceeding 90% sensitivity and specificity in several cases. Systems are now mostly statistically-based, and therefore require annotated corpora for training. Creating annotated clinical text corpora is one of the main challenges for the future of this field. The effort required to develop annotated corpora is significant and patient data confidentiality issues hamper access to data. An issue that we observed in several publications is the quality of the evaluation of the systems. The study design might be prone to biases, and the reference standards used might have limited value, especially when created by only one reviewer. Robust evaluation practices in this domain are well described in Hripcsak et al. [100]. The potential uses of information extraction from clinical text are numerous and far-reaching. Current applications, however, are rarely applied outside of the laboratories they have been developed in, mostly because of scalability and generalizability issues. In the same way the MUCs have fostered the development of information extraction in the general domain, similar competitive challenges for information extraction from clinical text will undoubtedly stimulate advances in the field reviewed here. Organizing these competitive challenges is another challenge for the future. Some domains of research like discourse analysis and temporality analysis have not been investigated thoroughly yet and pose addi-

tional challenges that could also contribute to performance improvements. Improvements in system performance will subsequently enhance the acceptance and usage of IE in concrete clinical and biomedical research contexts.

References

1. Spyns P. Natural language processing in medicine: an overview. Methods Inf Med 1996 Dec;35(4-5):285-301.
2. Cohen AM, Hersh WR. A survey of current work in biomedical text mining. Brief Bioinform 2005 Mar;6(1):57-71.
3. Zweigenbaum P, Demner-Fushman D, Yu H, Cohen KB. Frontiers of biomedical text mining: current progress. Brief Bioinform 2007:358-75.
4. DeJong GF. An Overview of the FRUMP System. In: Ringle WGLaMH, editor. Strategies for Natural Language Processing. Hillsdale, NJ: Lawrence Erlbaum; 1982. p. 149-76.
5. Google. [cited 01/10/2008]; Available from: http://www.google.com
6. PubMed. [cited 01/10/2008]; Available from: http://www.ncbi.nlm.nih.gov/sites/entrez/
7. Carbonell JG, Hayes PJ. Natural Language Understanding. In: Shapiro SC, editor. Encyclopedia of Artificial Intelligence: Wiley;1992. p. 660-77.
8. Ananiadou S, McNaught J. Text Mining for Biology and Biomedicine: Artech House, Inc; 2006.
9. Hearst MA. Untangling text data mining. Proc 37th Annual meeting of the Association for Computational Linguistics. College Park, MD; 1999. p. 3-10.
10. Liu H, Lussier YA, Friedman C. Disambiguating ambiguous biomedical terms in biomedical narrative text: an unsupervised method. J Biomed Inform 2001:249-61.
11. Sager N, Friedman C, Chi E. The analysis and processing of clinical narrative. In: Salamon R, Blum B, Jørgensen M, editors. Medinfo 86; 1986; Amsterdam (Holland): Elsevier; 1986. p. 1101-5.
12. Friedman C, Johnson SB, Forman B, Starren J. Architectural requirements for a multipurpose natural language processor in the clinical environment. Proc Annu Symp Comput Appl Med Care. 1995:347-51.
13. Hripcsak G, Kuperman GJ, Friedman C. Extracting findings from narrative reports: software transferability and sources of physician disagreement. Methods Inf Med 1998:1-7.
14. Haug PJ, Ranum DL, Frederick PR. Computerized extraction of coded findings from free-text radiologic reports. Work in progress. Radiology 1990 Feb;174(2):543-8.
15. 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 1995:284-8.
16. McCray AT, Sponsler JL, Brylawski B, Browne AC. The role of lexical knowledge in biomedical text understanding. SCAMC 87; IEEE; 1987. p. 103-7.
17. Lindberg C. The Unified Medical Language System (UMLS) of the National Library of Medicine. Journal (American Medical Record Association). 1990 May;61(5):40-2.
18. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp 2001:17-21.
19. McNaught J, Black WJ. Information extraction: the task. In: Ananiadou S, McNaught J, editors. Text Mining for Biology and Biomedicine: Artech House Books; 2006. p. 143-76.
20. Hobbs JR. The generic information extraction system. Proc MUC-5; Baltimore, MD: Morgan Kaufmann; 1993. p. 87-92.
21. Hobbs JR. Information extraction from biomedical text. J Biomed Inform 2002 Aug;35(4):260-4.
22. Pakhomov S, Buntrock J, Duffy PH. High Throughput Modularized NLP System for Clinical Text 43rd Annual Meeting of the Association for Computational Linguistics; 2005; Ann Arbor, MI; 2005.
23. cancer Text Information Extraction System (caTIES) website. [cited 01/10/2008]; Available from: https://cabig.nci.nih.gov/tools/caties
24. Friedman C. A broad-coverage natural language processing system. Proc AMIA Symp 2000:270-4.
25. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc 2004:392-402.
26. Xu H, Anderson K, Grann VR, Friedman C. Facilitating cancer research using natural language processing of pathology reports. Medinfo 2004:565-72.
27. Liu K, Mitchell KJ, Chapman WW, Crowley RS. Automating tissue bank annotation from pathology reports - comparison to a gold standard expert annotation set. AMIA Annu Symp Proc 2005:460-4.
28. Hahn U, Romacker M, Schulz S. Creating knowledge repositories from biomedical reports: the MEDSYNDIKATE text mining system. Pac Symp Biocomput 2002:338-49.
29. International Challenge: Classifying Clinical Free Text Using Natural Language Processing. [cited 01/10/2008]; Available from: http://www.computationalmedicine.org/challenge/index.php
30. Pestian JP, Brew C, Matykiewicz P, Hovermale DJ, Johnson N, Cohen KB, et al. A Shared Task Involving Multi-label Classi?cation of Clinical Free Text. BioNLP 2007: Biological, translational, and clinical language processing. Prague, CZ; 2007.
31. i2b2 (Informatics for Integrating Biology and the Bedside) website. [cited 01/10/2008]; Available from: https://www.i2b2.org/
32. Uzuner O, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. J Am Med Inform Assoc 2007:550-63.
33. Uzuner O, Goldstein I, Luo Y, Kohane I. Identifying Patient Smoking Status from Medical Discharge Records. J Am Med Inform Assoc 2008 January-February;15(1):14-24. Epub 2007 Oct 18.
34. Ruch P, Baud R, Geissbuhler A. Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record. Artif Intell Med 2003:169-84.
35. Tolentino HD, Matters MD, Walop W, Law B, Tong

W, Liu F, et al. A UMLS-based spell checker for natural language processing in vaccine safety. BMC Med Inform Decis Mak 2007:3.

36. Miller G. WordNet: a dictionary browser. Proc of the First International Conference on Information and Data; 1985; Ontario, Canada; 1985.

37. Fellbaum C. WordNet: An electronic lexical database. Cambridge, MA: MIT Press; 1998.

38. Tomanek K, Wermter J, Hahn U. A reappraisal of sentence and token splitting for life sciences documents. Medinfo 2007:524-8.

39. Weeber M, Mork JG, Aronson AR. Developing a test collection for biomedical word sense disambiguation. Proc AMIA Symp 2001:746-50.

40. Liu H, Teller V, Friedman C. A multi-aspect comparison study of supervised word sense disambiguation. J Am Med Inform Assoc 2004:320-31.

41. Xu H, Markatou M, Dimova R, Liu H, Friedman C. Machine learning and word sense disambiguation in the biomedical domain: design and evaluation issues. BMC Bioinformatics 2006:334.

42. Pakhomov S, Pedersen T, Chute CG. Abbreviation and acronym disambiguation in clinical discourse. AMIA Annu Symp Proc 2005:589-93.

43. Coden A, Savova G, Buntrock J, Sominsky I, Ogren PV, Chute CG, et al. Text analysis integration into a medical information retrieval system: challenges related to word sense disambiguation. Medinfo; 2007; Brisbane, Australia; 2007.

44. Campbell DA, Johnson SB. Comparing syntactic complexity in medical and non-medical corpora. Proc AMIA Symp 2001:90-4.

45. Coden AR, Pakhomov SV, Ando RK, Duffy PH, Chute CG. Domain-specific language models and lexicons for tagging. J Biomed Inform 2005:422-30.

46. Liu K, Chapman W, Hwa R, Crowley RS. Heuristic sample selection to minimize reference standard training set for a part-of-speech tagger. J Am Med Inform Assoc 2007:641-50.

47. Hahn U, Wermter J. High-Performance Tagging on Medical Texts. 20th International Conference on Computational Linguistics, Geneva, Switzerland; 2004.

48. Campbell DA, Johnson SB. A transformational-based learner for dependency grammars in discharge summaries. Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain. Philadelphia, PN; 2002.

49. Clegg AB, Shepherd AJ. Benchmarking natural-language parsers for biological applications using dependency graphs. BMC Bioinformatics 2007:24.

50. Pyysalo S, Salakoski T, Aubin S, Nazarenko A. Lexical adaptation of link grammar to the biomedical sublanguage: a comparative evaluation of three approaches. BMC Bioinformatics 2006:S2.

51. Aronow DB, Fangfang F, Croft WB. Ad hoc classification of radiology reports. J Am Med Inform Assoc 1999:393-411.

52. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform 2001:301-10.

53. Mitchell KJ, Becich MJ, Berman JJ, Chapman WW, Gilbertson J, Gupta D, et al. Implementation and evaluation of a negation tagger in a pipeline-based system for information extract from pathology reports. Medinfo 2004:663-7.

54. Goryachev S, Sordo M, Zeng QT, Ngo L. Imple-

mentation and Evaluation of Four Different Methods of Negation Detection. DSG technical report?

55. Mutalik PG, Deshpande A, Nadkarni PM. Use of general-purpose negation detection to augment concept indexing of medical documents: a quantitative study using the UMLS. J Am Med Inform Assoc 2001:598-609.

56. Elkin PL, Brown SH, Bauer BA, Husser CS, Carruth W, Bergstrom LR, et al. A controlled trial of automated classification of negation from clinical notes. BMC Med Inform Decis Mak 2005:13.

57. Huang Y, Lowe HJ. A novel hybrid approach to automated negation detection in clinical radiology reports. J Am Med Inform Assoc 2007:304-11.

58. Hripcsak G, Zhou L, Parsons S, Das AK, Johnson SB. Modeling electronic discharge summaries as a simple temporal constraint satisfaction problem. J Am Med Inform Assoc. 2005 Jan-Feb;12(1):55-63.

59. Zhou L, Melton GB, Parsons S, Hripcsak G. A temporal constraint structure for extracting temporal information from clinical narrative. J Biomed Inform 2006:424-39.

60. Zhou L, Friedman C, Parsons S, Hripcsak G. System architecture for temporal information extraction, representation and reasoning in clinical narrative reports. AMIA Annu Symp Proc 2005:869-73.

61. Zhou L, Parsons S, Hripcsak G. The Evaluation of a Temporal Reasoning System in Processing Clinical Discharge Summaries. J Am Med Inform Assoc 2007:99-111.

62. Harkema H, Setzer A, Gaizauskas R, Hepple M. Mining and Modelling Temporal Clinical Data. Proceedings of the UK e-Science All Hands Meeting 2005 2005:507-14.

63. Bramsen P, Deshpande P, Lee YK, Barzilay R. Inducing Temporal Graphs. Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006). Sydney, Australia 2006:189-98.

64. Bramsen P, Deshpande P, Lee YK, Barzilay R. Finding temporal order in discharge summaries. AMIA Annu Symp Proc 2006:81-5.

65. Chu D, Dowling JN, Chapman WW. Evaluating the effectiveness of four contextual features in classifying annotated clinical conditions in emergency department reports. AMIA Annu Symp Proc 2006:141-5.

66. Chapman W, Chu D, Dowling JN. ConText: An Algorithm for Identifying Contextual Features from Clinical Text. BioNLP 2007: Biological, translational, and clinical language processing. Prague, CZ; 2007.

67. Meystre S, Haug PJ. Natural language processing to extract medical problems from electronic clinical documents: performance evaluation. J Biomed Inform 2006:589-99.

68. Long W. Extracting diagnoses from discharge summaries. AMIA Annu Symp Proc 2005:470-4.

69. McCray AT, Aronson AR, Browne AC, Rindflesch TC, Razi A, Srinivasan S. UMLS knowledge for biomedical language processing. Bull Med Libr Assoc 1993 Apr;81(2):184-94.

70. Turchin A, Kolatkar NS, Grant RW, Makhni EC, Pendergrass ML, Einbinder JS. Using regular expressions to abstract blood pressure and treatment intensification information from the text of physician notes. J Am Med Inform Assoc 2006:691-5.

71. Friedlin J, McDonald CJ. A natural language pro-

cessing system to extract and code concepts relating to congestive heart failure from chest radiology reports. AMIA Annu Symp Proc 2006:269-73.

72. Bashyam V, Divita G, Bennett DB, Browne AC, Taira RK. A normalized lexical lookup approach to identifying UMLS concepts in free text. Medinfo 2007:545-9.

73. Divita G, Tse T, Roth L. Failure analysis of MetaMap Transfer (MMTx). Medinfo. 2004;11(Pt 2):763-7.

74. Baud RH, Lovis C, Rassinoux AM, Scherrer JR. Morpho-semantic parsing of medical expressions. Proc AMIA Symp 1998:760-4.

75. Mamlin BW, Heinze DT, McDonald CJ. Automated extraction and normalization of findings from cancer-related free-text radiology reports. AMIA Annu Symp Proc 2003:420-4.

76. Spyns P, De Moor G. A Dutch medical language processor. Int J Biomed Comput 1996:181-205.

77. Dolin RH, Alschuler L, Beebe C, Biron PV, Boyer SL, Essin D, et al. The HL7 Clinical Document Architecture. J Am Med Inform Assoc 2001 Nov-Dec;8(6):552-69.

78. Huang Y, Lowe HJ, Klein D, Cucina RJ. 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 2005:275-85.

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 Mar-Apr;1(2):161-74.

80. Jain NL, Friedman C. Identification of findings suspicious for breast cancer based on natural language processing of mammogram reports. Proc AMIA Annu Fall Symp 1997:829-33.

81. Wilcox AB, Narus SP, Bowes WA, 3rd. Using natural language processing to analyze physician modifications to data entry templates. Proc AMIA Symp 2002:899-903.

82. Castilla AC, Furuie SS, Mendonca EA. Multilingual information retrieval in thoracic radiology: feasibility study. Medinfo 2007:387-91.

83. Schadow G, McDonald CJ. Extracting structured information from free text pathology reports. AMIA Annu Symp Proc 2003:584-8.

84. Chung J, Murphy S. Concept-value pair extraction from semi-structured clinical narrative: a case study using echocardiogram reports. AMIA Annu Symp Proc 2005:131-5.

85. Christensen L, Haug P, Fiszman M. MPLUS: A Probabilistic Medical Language Understanding System. BioNLP 2002.

86. Fiszman M, Haug PJ, Frederick PR. Automatic extraction of PIOPED interpretations from ventilation/perfusion lung scan reports. Proc AMIA Symp 1998:860-4.

87. 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 2000:593-604.

88. Trick WE, Chapman WW, Wisniewski MF, Peterson BJ, Solomon SL, Weinstein RA. Electronic interpretation of chest radiograph reports to detect central venous catheters. Infect Control Hosp Epidemiol 2003:950-4.

89. Shared Pathology Informatics Network (SPIN) website. [cited 01/10/2008]; Available from: http:/

/spin.nci.nih.gov/

90. General Architecture for Text Engineering (GATE) website. [cited 01/10/2008]; Available from: http://gate.ac.uk/

91. Fenstermacher D, Street C, McSherry T, Nayak V, Overby C, Feldman M. The Cancer Biomedical Informatics Grid (caBIG™). Conf Proc IEEE Eng Med Biol Soc 2005;1:743-6.

92. NCI Enterprise Vocabulary Services (EVS) website. [cited 01/10/2008]; Available from: http://evs.nci.nih.gov/

93. Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Med Inform Decis Mak 2006:30.

94. Denny JC, Spickard A, 3rd, Miller RA, Schildcrout J, Darbar D, Rosenbloom ST, et al. Identifying UMLS concepts from ECG Impressions using KnowledgeMap. AMIA Annu Symp Proc 2005:196-200.

95. Denny JC, Peterson JF. Identifying QT prolongation from ECG impressions using natural language processing and negation detection. Medinfo 2007:1283-8.

96. Taira RK, Soderland SG. A statistical natural language processor for medical reports. Proc AMIA Symp 1999:970-4.

97. Bashyam V, Taira RK. Indexing anatomical phrases in neuro-radiology reports to the UMLS 2005AA. AMIA Annu Symp Proc 2005:26-30.

98. Sibanda T, He T, Szolovits P, Uzuner O. Syntactically-informed semantic category recognition in discharge summaries. AMIA Annu Symp Proc 2006:714-8.

99. Friedman C, Hripcsak G, Shablinsky I. An evaluation of natural language processing methodologies. Proc AMIA Symp 1998:855-9.

100. Hripcsak G, Kuperman GJ, Friedman C, Heitjan DF. A reliability study for evaluating information extraction from radiology reports. J Am Med Inform Assoc 1999:143-50.

101. Friedlin J, McDonald CJ. Using a natural language processing system to extract and code family history data from admission reports. AMIA Annu Symp Proc 2006:925.

102. Friedman C. Towards a comprehensive medical language processing system: methods and issues. Proc AMIA Annu Fall Symp 1997:595-9.

103. Clark C, Good K, Jezierny L, Macpherson M, Wilson B, Chajewska U. Identifying Smokers with a Medical Extraction System. J Am Med Inform Assoc 2008 January-February;15(1):36-9. Epub 2007 Oct 18.

104. Cohen AM. Five-way Smoking Status Classification Using Text Hot-Spot Identification and Error-correcting Output Codes. J Am Med Inform Assoc 2008 January-February;15(1):32-5. Epub 2007 Oct 18.

105. Heinze DT, Morsch ML, Potter BC, Sheffer RE, Jr. Medical i2b2 NLP Smoking Challenge: The A-Life System Architecture and Methodology. J Am Med Inform Assoc 2008 January-February;15(1):40-3. Epub 2007 Oct 18.

106. Savova GK, Ogren PV, Duffy PH, Buntrock JD, Chute CG. Mayo Clinic NLP System for Patient Smoking Status Identification. J Am Med Inform Assoc 2008 January-February;15(1):25-8. Epub

2007 Oct 18.

107. Wicentowski R, Sydes MR. Using Implicit Information to Identify Smoking Status in Smoke-blind Medical Discharge Summaries. J Am Med Inform Assoc. 2008 January-February;15(1):29-31. Epub 2007 Oct 18.

108. Aronson AR, Bodenreider O, Demner-Fushman D, Fung KW, Lee VK, Mork JG, et al. From indexing the biomedical literature to coding clinical text: experience with MTI and machine learning approaches. BioNLP 2007: Biological, translational, and clinical language processing. Prague, CZ; 2007:105-12.

109. Crammer K, Dredze M, Ganchev K, Talukdar PP, Carroll S. Automatic Code Assignment to Medical Text. BioNLP 2007: Biological, translational, and clinical language processing. Prague, CZ; 2007:129-36.

110. Baud R. A natural language based search engine for ICD10 diagnosis encoding. Med Arh 2004:79-80.

111. Aramaki E, Imai T, Kajino M, Miyo K, Ohe K. Statistical selector of the best multiple ICD-coding method. Medinfo 2007;12(Pt 1):645-9.

112. Friedman C, Knirsch C, Shagina L, Hripcsak G. Automating a severity score guideline for community-acquired pneumonia employing medical language processing of discharge summaries. Proc AMIA Symp 1999:256-60.

113. Elkins JS, Friedman C, Boden-Albala B, Sacco RL, Hripcsak G. Coding neuroradiology reports for the Northern Manhattan Stroke Study: a comparison of natural language processing and manual review. Comput Biomed Res 2000:1-10.

114. Kukafka R, Bales ME, Burkhardt A, Friedman C. Human and automated coding of rehabilitation discharge summaries according to the International Classification of Functioning, Disability, and Health. J Am Med Inform Assoc 2006:508-15.

115. Lussier YA, Shagina L, Friedman C. Automating SNOMED coding using medical language understanding: a feasibility study. Proc AMIA Symp 2001:418-22.

116. Hasman A, de Bruijn LM, Arends JW. Evaluation of a method that supports pathology report coding. Methods Inf Med 2001;40(4):293-7.

117. Pakhomov SV, Buntrock JD, Chute CG. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. J Am Med Inform Assoc 2006:516-25.

118. Haug PJ, Christensen L, Gundersen M, Clemons B, Koehler S, Bauer K. A natural language parsing system for encoding admitting diagnoses. Proc AMIA Annu Fall Symp 1997:814-8.

119. Gundersen ML, Haug PJ, Pryor TA, van Bree R, Koehler S, Bauer K, et al. Development and evaluation of a computerized admission diagnoses encoding system. Comput Biomed Res 1996:351-72.

120. Kashyap V, Turchin A, Morin L, Chang F, Li Q, Hongsermeier T. Creation of structured documentation templates using Natural Language Processing techniques. AMIA Annu Symp Proc 2006:977.

121. Lovis C, Payne TH. Extending the VA CPRS electronic patient record order entry system using natural language processing techniques. Proc AMIA Symp 2000:517-21.

122. Cimino JJ, Bright TJ, Li J. Medication reconciliation using natural language processing and con-

trolled terminologies. Medinfo 2007:679-83.

123. Liu H, Friedman C. CliniViewer: a tool for viewing electronic medical records based on natural language processing and XML. Medinfo 2004:639-43.

124. Day S, Christensen LM, Dalto J, Haug P. Identification of trauma patients at a level 1 trauma center utilizing natural language processing. J Trauma Nurs 2007:79-83.

125. Mendonca EA, Haas J, Shagina L, Larson E, Friedman C. Extracting information on pneumonia in infants using natural language processing of radiology reports. J Biomed Inform 2005:314-21.

126. Fiszman M, Haug PJ. Using medical language processing to support real-time evaluation of pneumonia guidelines. Proc AMIA Symp 2000:235-9.

127. Jain NL, Knirsch CA, Friedman C, Hripcsak G. Identification of suspected tuberculosis patients based on natural language processing of chest radiograph reports. Proc AMIA Annu Fall Symp 1996:542-6.

128. Meystre S, Haug PJ. Automation of a problem list using natural language processing. BMC Med Inform Decis Mak 2005:30.

129. Meystre SM, Haug PJ. Comparing natural language processing tools to extract medical problems from narrative text. AMIA Annu Symp Proc 2005:525-9.

130. Hazlehurst B, Frost HR, Sittig DF, Stevens VJ. MediClass: A system for detecting and classifying encounter-based clinical events in any electronic medical record. J Am Med Inform Assoc 2005:517-29.

131. Hazlehurst B, Mullooly J, Naleway A, Crane B. Detecting possible vaccination reactions in clinical notes. AMIA Annu Symp Proc 2005:306-10.

132. Hazlehurst B, Sittig DF, Stevens VJ, Smith KS, Hollis JF, Vogt TM, et al. Natural language processing in the electronic medical record: assessing clinician adherence to tobacco treatment guidelines. Am J Prev Med 2005:434-9.

133. Johnson SB, Friedman C. Integrating data from natural language processing into a clinical information system. Proc AMIA Annu Fall Symp 1996:537-41.

134. Penz JF, Wilcox AB, Hurdle JF. Automated identification of adverse events related to central venous catheters. J Biomed Inform 2007:174-82.

135. Melton GB, Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. J Am Med Inform Assoc 2005:448-57.

136. Cao H, Stetson P, Hripcsak G. Assessing explicit error reporting in the narrative electronic medical record using keyword searching. J Biomed Inform 2003:99-105.

137. Chapman WW, Christensen LM, Wagner MM, Haug PJ, Ivanov O, Dowling JN, et al. Classifying free-text triage chief complaints into syndromic categories with natural language processing. Artif Intell Med 2005:31-40.

138. Haas JP, Mendonca EA, Ross B, Friedman C, Larson E. Use of computerized surveillance to detect nosocomial pneumonia in neonatal intensive care unit patients. Am J Infect Control 2005:439-43.

139. Chapman WW, Fiszman M, Dowling JN, Chapman BE, Rindflesch TC. Identifying respi-

ratory findings in emergency department reports for biosurveillance using MetaMap. Medinfo 2004:487-91.

140. Chapman WW, Dowling JN, Wagner MM. Fever detection from free-text clinical records for biosurveillance. J Biomed Inform 2004:120-7.

141. Brown SH, Speroff T, Fielstein EM, Bauer BA, Wahner-Roedler DL, Greevy R, et al. eQuality: electronic quality assessment from narrative clinical reports. Mayo Clinic proceedings 2006:1472-81.

142. 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 2007:281-8.

143. Pakhomov SS, Hemingway H, Weston SA, Jacobsen SJ, Rodeheffer R, Roger VL. Epidemiology of angina pectoris: role of natural language processing of the medical record. Am Heart J 2007:666-73.

144. Pakhomov SV, Buntrock J, Chute CG. Prospective recruitment of patients with congestive heart failure using an ad-hoc binary classifier. J Biomed Inform 2005:145-53.

145. Niu Y, Zhu X, Li J, Hirst G. Analysis of polarity information in medical text. AMIA Annu Symp Proc 2005:570-4.

146. Dorr DA, Phillips WF, Phansalkar S, Sims SA, Hurdle JF. Assessing the difficulty and time cost of de-identification in clinical narratives. Methods Inf Med 2006:246-52.

147. Sweeney L. Replacing personally-identifying information in medical records, the Scrub system. Proc AMIA Annu Fall Symp 1996:333-7.

148. Ruch P, Baud RH, Rassinoux AM, Bouillon P, Robert G. Medical document anonymization with a semantic lexicon. Proc AMIA Symp 2000:729-33.

149. Taira RK, Bui AA, Kangarloo H. Identification of patient name references within medical documents using semantic selectional restrictions. Proc AMIA Symp 2002:757-61.

150. Thomas SM, Mamlin B, Schadow G, McDonald C. A successful technique for removing names in pathology reports using an augmented search and replace method. Proc AMIA Symp 2002:777-81.

151. Berman JJ. Concept-match medical data scrubbing. How pathology text can be used in research. Archives of pathology & laboratory medicine 2003:680-6.

152. Fielstein EM, Brown SH, Speroff T. Algorithmic De-identification of VA Medical Exam Text for HIPAA Privacy Compliance: Preliminary Findings. Medinfo 2004:1590.

153. 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 2004:176-86.

154. Beckwith BA, Mahaadevan R, Balis UJ, Kuo F. Development and evaluation of an open source software tool for deidentification of pathology reports. BMC Med Inform Decis Mak 2006:12.

155. Sibanda T, Uzuner O. Role of Local Context in Automatic Deidenti?cation of Ungrammatical, Fragmented Text. ACL conference 2006.

156. Wellner B, Huyck M, Mardis S, Aberdeen J, Morgan A, Peshkin L, et al. Rapidly retargetable approaches to de-identification in medical records. J Am Med Inform Assoc 2007:564-73.

157. Szarvas G, Farkas R, Busa-Fekete R. State-of-the-art anonymization of medical records using an iterative machine learning framework. J Am Med Inform Assoc 2007:574-80.

158. Ananiadou S, Nenadic G. Automatic Terminology Management in Biomedicine. In: Ananiadou S, McNaught J, eds. Text Mining for Biology and Biomedicine: Artech House Books 2006:67-98.

159. Baneyx A, Charlet J, Jaulent MC. Methodology to build medical ontology from textual resources. AMIA Annu Symp Proc 2006:21-5.

160. Baneyx A, Charlet J, Jaulent MC. Building an ontology of pulmonary diseases with natural language processing tools using textual corpora. Int J Med Inform 2007:208-15.

161. Zhou L, Tao Y, Cimino JJ, Chen ES, Liu H, Lussier YA, et al. Terminology model discovery using natural language processing and visualization techniques. J Biomed Inform 2006:626-36.

162. Charlet J, Bachimont B, Jaulent MC. Building medical ontologies by terminology extraction from texts: an experiment for the intensive care units. Comput Biol Med 2006:857-70.

163. Kolesa P, Preckova P. Tools for Czech biomedical ontologies creation. Stud Health Technol Inform 2006:775-80.

164. Hersh WR, Campbell EH, Evans DA, Brownlow ND. Empirical, automated vocabulary discovery using large text corpora and advanced natural language processing tools. Proc AMIA Annu Fall Symp 1996:159-63.

165. Harris MR, Savova GK, Johnson TM, Chute CG. A term extraction tool for expanding content in the domain of functioning, disability, and health: proof of concept. J Biomed Inform 2003 Aug-Oct;36(4-5):250-9.

166. Savova GK, Harris M, Johnson T, Pakhomov SV, Chute CG. A data-driven approach for extracting "the most specific term" for ontology development. AMIA Annu Symp Proc 2003:579-83.

167. Savova G, Becker D, Harris M, Chute CG. Combining Rule-Based Methods and Latent Semantic Analysis for Ontology Structure Construction. Medinfo; 2004; San Francisco, CA; 2004. p. 1848.

168. do Amaral MB, Roberts A, Rector AL. NLP techniques associated with the OpenGALEN ontology for semi-automatic textual extraction of medical knowledge: abstracting and mapping equivalent linguistic and logical constructs. Proc AMIA Symp 2000:76-80.

169. Friedman C, Liu H, Shagina L. A vocabulary development and visualization tool based on natural language processing and the mining of textual patient reports. J Biomed Inform 2003:189-201.

170. UMLS Knowledge Source Server (UMLSKS). [cited 01/10/2008]; Available from: http://umlsks.nlm.nih.gov

171. The Lexical Grid (LexGrid). [cited 01/10/2008]; Available from: http://informatics.mayo.edu/LexGrid/index.php?page=

172. Tuttle MS, Olson NE, Keck KD, Cole WG, Erlbaum MS, Sherertz DD, et al. Metaphrase: an aid to the clinical conceptualization and formalization of patient problems in healthcare enterprises. Methods Inf Med 1998 Nov;37(4-5):373-83.

173. Cimino JJ, Clayton PD, Hripcsak G, Johnson SB. Knowledge-based approaches to the maintenance of a large controlled medical terminology. J Am Med Inform Assoc 1994 Jan-Feb;1(1):35-50.

174. Chapman WW, Dowling JN. Inductive creation of an annotation schema for manually indexing clinical conditions from emergency department reports. J Biomed Inform 2006:196-208.

175. Chapman WW, Dowling JN, Hripcsak G. Evaluation of training with an annotation schema for manual annotation of clinical conditions from emergency department reports. Int J Med Inform 2007.

176. Cohen KB, Fox L, Ogren PV, Hunter L. Empirical data on corpus design and usage in biomedical natural language processing. AMIA Annu Symp Proc 2005:156-60.

177. Cohen KB, Fox L, Ogren PV, Hunter L. Corpus design for biomedical natural language processing. AC-ISMB Workshop on Linking Biological Literature, Ontologies and Databases; 2005: Association for Computational Linguistics; 2005. p. 38-45.

178. Wilbur WJ, Rzhetsky A, Shatkay H. New directions in biomedical text annotation: definitions, guidelines and corpus construction. BMC Bioinformatics 2006:356.

179. Hirschman L, Blaschke C. Evaluation of Text Mining in Biology. In: Ananiadou S, McNaught J, editors. Text Mining for Biology and Biomedicine: Artech House Books 2006:67-98.

180. Swanson DR. Two medical literatures that are logically but not bibliographically connected. JASIS 1987;38(4):228-33.

181. Chen ES, Hripcsak G, Xu H, Markatou M, Friedman C. Automated Acquisition of Disease-Drug Knowledge from Biomedical and Clinical Documents: An Initial Study. J Am Med Inform Assoc 2007.

182. Cao H, Hripcsak G, Markatou M. A statistical methodology for analyzing co-occurrence data from a large sample. J Biomed Inform 2007 Jun;40(3):343-52.

183. Cao H, Markatou M, Melton GB, Chiang MF, Hripcsak G. Mining a clinical data warehouse to discover disease-finding associations using co-occurrence statistics. AMIA Annu Symp Proc 2005:106-10.

184. Rindflesch TC, Pakhomov SV, Fiszman M, Kilicoglu H, Sanchez VR. Medical facts to support inferencing in natural language processing. AMIA Annu Symp Proc 2005:634-8.

Correspondence to:
Stéphane M. Meystre
University of Utah
Department of Biomedical Informatics
26 South 2000 East, HSEB Suite 5700
Salt Lake City, UT 84112-5750
USA
Tel: +1 801 581 4080
Fax: +1 801 581 4297
E-mail: stephane.meystre@hsc.utah.edu