Part I Methods and techniques for mining biomedical literature and electronic health records

Amy Neustein, S. Sagar Imambi, Mário Rodrigues, António Teixeira and Liliana Ferreira

1 Application of text mining to biomedical knowledge extraction: analyzing clinical narratives and medical literature

Abstract: One of the tools that can aid researchers and clinicians in coping with the surfeit of biomedical information is text mining. In this chapter, we explore how text mining is used to perform biomedical knowledge extraction. By describing its main phases, we show how text mining can be used to obtain relevant information from vast online databases of health science literature and patients' electronic health records. In so doing, we describe the workings of the four phases of biomedical knowledge extraction using text mining (text gathering, text preprocessing, text analysis, and presentation) entailed in retrieval of the sought information with a high accuracy rate. The chapter also includes an in depth analysis of the differences between clinical text found in electronic health records and biomedical text found in online journals, books, and conference papers, as well as a presentation of various text mining tools that have been developed in both university and commercial settings.

1.1 Introduction

The corpus of biomedical information is growing very rapidly. New and useful results appear every day in research publications, from journal articles to book chapters to workshop and conference proceedings. Many of these publications are available online through journal citation databases such as Medline – a subset of the PubMed interface that enables access to Medline publications – which is among the largest and most well-known online databases for indexing professional literature. Such databases and their associated search engines contain important research work in the biological and medical domain, including recent findings pertaining to diseases, symptoms, and medications. Researchers widely agree that the ability to retrieve desired information is vital for making efficient use of the knowledge found in online databases. Yet, given the

current state of *information overload* efficient retrieval of useful information may be severely hampered. Hence, a retrieval system "should not only be able to retrieve the sought information, but also filter out irrelevant documents, while giving the relevant ones the highest ranking" (Ramampiaro 2010).

One of the tools that can aid researchers and clinicians in coping with the surfeit of information is text mining. Text mining refers to the process of deriving high-quality information from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. Those in the field have come to define text mining in rather broad terms. For some, text mining centers on finding implicit information, such as associations between concepts, by analyzing large amounts of text. For others it pivots on extraction of explicit, not implicit, information from texts, such as named entities mentions or relations explicitly such as "A leads to B." The task of identifying sentences with co-occurrences of a drug and a gene entity (for posterior manual curation into a database) is an example of the latter definition of text mining, which revolves around finding explicit information. Still, there are those who define text mining in the most stringent form: finding associations between a specific gene and a specific drug(s) based on clear-cut statistical analysis. No matter what view one subscribes to, text mining tools and methods are utilized, nonetheless, to significantly reduce human effort to build information systems and to automate the information retrieval and extraction process.

In particular, text mining aids in the search for information by using patterns for which the values of the elements are not exactly known in advance. In short, such tools are used to automate information retrieval and extraction systems, and by so doing, they help researchers to a large extent in dealing with the persistent problem of information overload. All in all, biomedical text mining "holds the promise of, and in some cases delivers, a reduction in cost and an acceleration of discovery, providing timely access to needed facts, as well as explicit and implicit associations among facts" (Simpson & Demner-Fushman 2012, p. 466). In this vein, biomedical text mining tools have been developed for the purpose of improving the efficiency and effectiveness of medical researchers, practitioners, and other health professionals so that they can deliver optimal health care. In the end, it is the patient who benefits from a more informed healthcare provider.

The field of text mining has witnessed a number of interesting applications. In Nahm and Mooney's (2002) AAAI technical report on text mining they describe how a special framework for text mining, called

DiscoTEX (Discovery from Text EXtraction), uses “a learned information extraction system to transform text into more structured data” so that it can be “mined for interesting relationships” (p. 60). In so doing, they define text mining as “the process of finding useful or interesting patterns, models, directions, trends or rules from *unstructured* text” (p. 61). In contrast to DiscoTEX, there are those applications that to try to infer higher-level associations or correlations between concepts. Arrowsmith[1] and BITOLA[2] are examples of such text mining applications that work on this higher level of association. Similarly, both MEDIE[3] and EvenMine[4] are examples of systems that perform more fine-grained linguistic analysis.

In Feldman and Sanger’s text mining handbook (2006) the authors show how text mining achieves its goal of extracting useful information from document collections “through the identification and exploration of interesting patterns.” Though the authors show that “text mining derives much of its inspiration and direction from seminal research on data mining,” they also emphasize that text mining is vastly different from data mining. This is so, because in text mining “the data sources are document collections” whereas in data mining the data sources are formal databases. As a result, in text mining, interesting patterns are found not among formalized database records” (as is the case with data mining), but rather “in the unstructured textual data in the documents in these collections” (p. 1).

Cohen and Hersh (2005) show that though text mining is concerned with unstructured text (as is likewise the case with natural language processing) it can, nevertheless, be “differentiated from ... natural language processing (NLP) in that NLP attempts to understand the meaning of text as a whole, while text mining and knowledge extraction concentrate on solving a specific problem in a specific domain identified *a priori* ...” The authors provide as an example the compilation of literature pertaining to migraine headache treatment, showing how the use of text mining “can aid database curators by selecting articles most likely to contain information of interest or potential new treatments for migraine [which] may be determined by looking for pharmacological substances that are associated with biological processes associated with migraine” (p. 58).

Current trends in biomedical text mining (Hakenberg et al. 2012; Gurulingappa et al. 2013; Zhao et al. 2014) include the extraction of information related to the recognition of chemical compound and drug mentions or drug dosage and symptoms. They also include extraction of drug-induced adverse effects, text mining of pathways and enzymatic reactions, and ranking of cancer-related mutations that cluster in

particular regions of the protein sequence.

In this chapter, we explore how text mining is used to perform biomedical knowledge extraction. By describing its main phases, we show how text mining can be used to obtain relevant information from vast online databases of health science literature and patients' electronic health records. In so doing, we describe the workings of the four phases of biomedical knowledge extraction using text mining (text gathering, text preprocessing, text analysis, and presentation) entailed in retrieval of the sought information with a high accuracy rate. The chapter also includes an in depth analysis of the differences between clinical text found in electronic health records and biomedical text found in online journals, books, and conference papers, as well as a presentation of various text mining tools that have been developed in both university and commercial settings.

1.2 Background

1.2.1 Clinical and biomedical text

In general, clinical text is written by clinicians in the clinical setting. This text describes patients in terms of their demographics, medical pathologies, personal, social, and medical histories and the medical findings made during interviews, laboratory workup, imaging and scans, or the medical or surgical procedures that are preformed to address the underlying medical problem (Meystre et al. 2008). Here is an example of what clinical text may look like: "a sixty five year old Caucasian female with acute pancreatitis with history of gall stones ... patient complains of severe weight loss and abdominal pain ... blood test shows increase in blood serum amylase and lipase ... abdominal ultrasound shows enlarged bile duct ... ERCP (endoscopic retrograde cholangiopancreatography) scheduled for patient next week for removal of stones from bile duct ... patient to be placed on low fat diet ..." (Though in actual clinical notes, abbreviations and symbols, such as those that indicate the patient's gender, are often used, we chose to omit such shorthand text for the purpose of giving a clear example here.)

As this example shows, clinical text describes a sequence of events and narratives, with the goal in mind of producing as precise and comprehensive an explanation as possible when describing the health status of a patient. This type of expressive description found in the clinical narrative understandably inheres a fair amount of ambiguity and personal

differences in both vocabulary and style (Lovis et al. 2000; Suominen 2009). The main purpose of clinical text is to serve as a summary or "handover note" of patient care (documentation relating to the transfer of responsibility of the patient to another care provider either within the same healthcare setting or at another health facility), but it can also be used for legal requirements, care continuity, reimbursement, case management and research. Clinical text covers every phase of care, and depending on the purpose, the documents may differ in style, lengthiness, conformity to grammatical rules and so on. As such, documents describing lab results and medical examinations are very different from those that describe patient care outcome in both the long run and short run.

There are other variations of clinical text as well. That is, clinical text may be entered either in *real time* or in retrospect, as a summary. In addition, clinical text may be entered at the patient's bedside or elsewhere (Thoroddsen et al. 2009). Clinical text contrasts with biomedical text, which is the kind of text that appears in books, articles, literature abstracts, posters, and so forth (Meystre et al. 2008). This is the kind of text that appears in MEDLINE/PubMed resources. Although both types of text do have some similarities, in that the heavy use of domain-specific terminology and the frequent inclusion of acronyms and polysemic words are found in both mediums, there are several features that make clinical text different from biomedical text. It is these differences that make clinical text especially challenging to NLP. Here are some of the reasons:

- Some clinical texts do not conform to the rules of grammar, are short, and are composed of telegraphic phrases;
- Clinical narratives are full of abbreviations, acronyms, and other shorthand phrases. Also, these shorthand lexical units are often overloaded, i.e., the same set of letters has multiple interpretations (Liu, Lussier & Friedman 2001);
- Misspellings are frequent in clinical text, as it is often produced without any spelling support;
- Clinical narratives often contain pasted sets of laboratory values or vital signs with embedded non-text strings, complicating otherwise straightforward NLP tasks like sentence splitting; and
- Templates and pseudo tables are often composed in plain text that are made to look tabular by the use of white space or lists.

Information search from this type of narrative text is difficult and time consuming. Standardization and structuring have been proposed as possible solutions. However, such solutions are not free of problems. For

example, converting narratives to numerical and structured data is laborious and easily leads to differences and errors in coding. Moreover, if these tasks are performed manually, which is currently the most common approach, text ambiguity and personal differences may cause inconsistencies (Suominen 2009). Also, converting narratives into structured data may lead to significant information losses, as it limits the expressive power of free-text (Lovis et al. 2000; Walsh 2004).

1.2.2 Information retrieval

The term “Information Retrieval” was coined in 1952; a decade later this term came to be popularly used in the research community (Van Rijsbergen 1979) and has continued to date. When the first automated information retrieval systems were actually introduced during the 1960s, the field of information retrieval (IR) was born. Information retrieval can be defined as the art and science of searching for information in large collections of documents; and, likewise, searching for text, sound, or images within those documents themselves. In addition, the search for metadata about documents is also part of information retrieval. According to Manning, Raghavan and Schutze (2008) “Information Retrieval (IR) is finding documents of an unstructured nature that satisfies an information need from within large collections (usually stored on computers).” As such, the field of information retrieval (IR) is the study of techniques for organizing and retrieving unstructured text stored on the computer. However, working with unstructured text, such as web pages, text documents, office documents, presentations and emails, can be quite difficult. That is, since unstructured text does not have a data model, it cannot be easily processed by a machine. *Structured* data, on the other hand, is either, in general, annotated or contained in databases (e.g., library catalogues and phone numbers), whereas unstructured data is not. (See Appendix “A” for list of open-sourced structured databases.)

Singhal (2001) opined that since the quantity of electronic information has increased dramatically with the widespread adoption of World Wide Web during the 1990s, information retrieval has become a sphere of great interest. Similarly, he saw the research and growth in this area as a natural consequence of the increasing interest in information retrieval.[5]

1.2.2.1 Information retrieval process

Information retrieval is used to locate specific items in a set of natural-language documents, such as finding specific gene-related information from the biomedical literature. IR systems provide a way for a user to

enter a *query*, using keywords, wherein the system will return the documents considered relevant to the query from the document collection. To do so, Herrera-Viedma (2001) explains, "both documents and user queries must be formally represented in a consistent way so that IRS [Information Retrieval System] can satisfactorily develop the retrieval activity." IR is achieved by scanning the collection for matched terms when a search is performed. The author shows that, basically, three components are involved in the information retrieval process:

1. *A Database:* which stores the documents and the representation of their information contents (index terms). It is built using tools for extracting index terms and for representing the documents.
2. *A Query Subsystem:* which allows users to formulate their queries by means of a query language.
3. *An Evaluation Subsystem:* which evaluates the documents for a user query. It presents an inference procedure that establishes a relationship between the user request and the documents in the database to determine the relevance of each document to the user query (p. 460).

The author points out that to help overcome the "lack of flexibility and precision for representing document contents, for describing user queries and for characterizing the relevance of the documents retrieved for a given user query" weights are incorporated at these three levels of information representation. Namely, at the document representation level in which a database is built, "by computing weights of index terms, the system specifies to what extent a document matches the concept expressed by the index terms"; at the query representation level "by attaching weights in a query" which allows the user to "provide a more precise description of his or her information needs or desired documents"; and at the evaluation representation level "by assigning weights to characterize the relationship between user queries and document representation" so that the evaluation subsystem can provide a means, known as the retrieval status value (RSV) of a document "to discriminate the documents retrieved by relevance judgments" (pp. 460–461).

In fact, a number of researchers in the field of informational retrieval have been encouraged to devise ways of making the entire information retrieval process more efficient. Some have, for example, embarked on various ways of streamlining the index size of the IR system. Gonzalez (2008) showed how the index system, also known as the inverted file (IF) that "serves as the data structure in charge of storing the information handled in the retrieval process" can be compressed using "document

reordering and static index pruning." The author shows how this new approach differs from the traditional "static compression schemes" though they are deemed "complementary to them," and that regardless of the approach used they all "have one thing in common: they make use of some of the properties inherently related to document collection."

1.2.3 Information extraction

Information extraction (IE) systems analyze *unstructured* text in order to extract information about pre-specified types of events, entities or relationships, such as the relationship between disease and genes or disease and food items. In other words, information extraction is all about deriving structured information from unstructured text. This differs from information retrieval (IR), described above, in that the purpose of IE is to *add* value and insight to the data whereas IR simply locates information in the same form(s) that it is stored without supplying any additional analytical insight about correlations, co-morbidity, or any other co-occurrence.

In addition, IE may be seen as a subtask of text mining, since the latter is a vast area that includes document classification, document clustering, building ontologies and other tasks, whereas IE is primarily concerned with crawling, parsing, and indexing documents so as to extract useful information from the data. In recent years, however, IE has distinguished itself from text mining as multimedia document processing, involving automatic annotation and content extraction from images, audio and video clips, has become more widely used. In fact, radiologists have come to depend on information extraction from medical images, using automatic image annotation systems in some of the more novel and creative ways.

1.2.4 Challenges to biomedical information extraction systems

Biomedical information extraction can build a database with the information on a given relationship or event drawn from a variety of sources such as online medical news, biomedical literature, or electronic health records. Since the documents are *unstructured* and expressed in a natural language format, it is very difficult for a computer to understand and analyze them. Yet, scientists and clinicians need to keep up-to-date with all of the new discoveries and theories presented in the biomedical literature, and they must, likewise, make efficient use of this ever-expanding reservoir of biomedical information. Undoubtedly, there is a

significant degree of information overload.

Not surprisingly, information overload places a heavy burden on biomedical information extraction systems to perform efficiently. However, biomedical IE systems face yet another problem, one that is undoubtedly *sui generis* to the biomedical domain. Ramampiaro (2010) describes how medical terms often cross over to vernacular usage, thereby causing false positives that artificially boost ranking scores. The duality of meaning ascribed to words, which can be found in both the vernacular or, alternatively, in biomedical literature and in clinical documents, constitutes a persistent problem associated with biomedical IE. Krauthammer and Nenadic (2004) point out that this duality of usage presents one of the biggest challenges to biomedical extraction in that "biomedical information typically contains large amounts of domain-specific terminology with *high ambiguity*" (emphasis supplied). This makes indexing particularly difficult.

For example, *heart* means the hollow muscular organ located behind the sternum and between the lungs in the medical context, but in the vernacular English language, it may be used to convey a different meaning, as in "the child won everyone's *heart*." Such linguistic ambiguities may create serious problems with how to rank the documents at hand. Finding the occurrence of the word "heart" many times in an online news article, for example, may give a speciously high ranking to the document if indeed the word "heart" had been used in a vernacular rather than in a biomedical context.

Furthermore, the need to learn and derive new knowledge also remains a challenge for biomedical information extraction systems. For all these reasons, there remains a growing need for the development of effective tools to meet these challenges and obstacles head-on so as to enable researchers and practitioners (and lay members who may need to research certain health issues) to access and extract useful information from the biomedical literature. It is understandable that this will require better machine learning tools that can perform heuristic discoveries so as to learn new relationships between entities and events that are not previously stored in the system.

In addition, the rapid increase in the sheer volume of biomedical literature necessitates the design of information extraction tools similar to the "open discovery" algorithm introduced by Srinivasan and Libbus (2004), which they used to "uncover information that could form the basis of new hypotheses." Or, the MedMeSH Summarizer System described by Kankar et al. (2002) to help streamline the process of cross-referencing "experimental and analytical results with previously known biological

facts, theories, and results." This is much needed given the breadth of biomedical databases, which can ordinarily make "the task of cross-referencing very lengthy, tedious, and daunting."

In sum, it is these special requirements of the biomedical domain that call for a new set of text mining tools, since the tools used for other domains have not proven entirely successful when applied to the biomedical sciences.

1.2.5 Applications of biomedical information extraction tools

Information extraction tools are used across various domains such as security, online media, marketing applications (Coussement & Poel 2008), and web mining (Zanasi 2009). Biomedical information extraction tools are used to perform a variety of functions. Text mining applications in biomedical area are diverse and they include:

1. The identification of chemical compounds: identifying their structures and the relations between them; and identifying drugs in which the particular compound is used, along with their respective side effects and toxicity (Vazquez et al. 2011);
2. Disease research such as cancer: several applications were developed to provide easy access to the most recent developments in cancer research (Zhu et al. 2013);
3. Genetics: gathering the most recent information about complex processes involving genes, proteins and phenotypes (Jensen, Jensen & Brunak 2012; Rebholz-Schuhmann, Oellrich & Hoehndorf 2012);
4. Extracting gene-based patterns using natural language processing techniques to extract the rhetoric information (the intention to be conveyed to the reader by the author(s) of the paper) contained in technical abstracts (Atkinson, Ferreira & Aravena 2004);
5. Indexing Medline documents (Kankar et al. 2002);
6. Finding the relationship between curcumin longa (a dietary substance) and retinal diseases (Srinivasan, Bisharah & Sehgal 2004);
7. Developing an expert system to perform medical diagnosis from clinical patient records and patient histories (Moumtzoglou & Kastania 2011); and
8. Finding risk factors of a disease (Imambi & Sudha 2010).

1.3 Biomedical knowledge extraction using text mining

The main phases, as shown in Fig. 1.1, of biomedical knowledge extraction using text mining are: (1) Unstructured text gathering and preprocessing; (2) Extraction of features and semantic information (including information extraction and creation of semantic metadata) to produce annotated texts; (3) Analysis of the annotated texts (using data mining, semantic search and knowledge discovery); and (4) Presentation. Each phase will be discussed in turn.

Typical text mining applications include the following: identification of facts in specialized (domain-based) literature, discovery of implicit and unknown facts, document summarization, and entity-relation modeling (i.e., learning relations between named entities). Applications usually scan sets of document to identify relevant information. The relevant information can be identified by either modelling the document set, using one or more classification schemes, or populating a database (adding information to a database or adding fields to a database in order to be able to fill it with information) or search index with the information that is extracted.

Some important subtasks are:

- Information retrieval or identification of a corpus, a preparatory step for collecting or identifying a set of textual materials (that either appear on the Web or are held in a file system, database, or content management system) for analysis.
- Named entity recognition is the use of gazetteers or statistical techniques to identify named text features: diseases, drugs, anatomical structures, dysfunctions, lab procedures, certain abbreviations, and so on. Disambiguation by using contextual clues that may be required in order to decide whether, for instance, “block” refers to a specific medical condition such as *intraventricular* block or *heart* block, or some other entity for that matter.
- Natural language processing (which are considered complex tasks that can take more time to complete), such as part of speech tagging, syntactic parsing, and other types of linguistic analysis. These tasks are performed less frequently, in part, as they require a longer processing time. Machine learning approaches usually include these tasks to generate features to be analyzed in the learning process and to support decision in runtime. Features can be at the token level, as lemmas and part of speech tags, or at the sentence level using syntactic parsing.

Fig. 1.1: Main phases of biomedical knowledge extraction using text mining.

1.3.1 Unstructured text gathering and preprocessing

1.3.1.1 Text gathering

The text-gathering phase provides an "interface" to collect the raw documents from online sources, such as online journals, books, and conference papers and from electronic health records compiled at major teaching hospitals and at local community medical facilities. Biomedical information is, thus, made available through such online literary databases and health records, as well from the web in general. One such interface for published materials is PubMed, whose largest component is MEDLINE, which serves as a freely accessible online database of biomedical journal citations and abstracts created by the U.S. National Library of Medicine.

As of 2014, Medline includes citations from over 5600 scholarly journals published in more than 80 countries around the world. PubMed comprises more that 22 million references that include the entire MEDLINE database and other types of citations, such as in-process citations, which provide records for articles before those records go through quality control and are indexed; citations to articles that are out-of-scope from certain MEDLINE journals; citations that precede the date that a journal was selected for MEDLINE indexing; and other works such as chapters and books that are likewise outside the purview of MEDLINE.[6] This repository of scientific literature provides a vast amount of text data that has helped researchers to implement their classification algorithms (Imambi & Sudha 2011).

The electronic health record (EHR) constitutes another major source of digital web-based data, primarily existing as part of the hospital's own collection of private computer networks (*an intranet*) rather than as part of the World Wide Web. Yet, this source of data can serve a goldmine of valuable clinical and demographic information on patient care. Such records contain a large repository of Patient Notes that describe the patient's medical history and treatments, plans for follow-up treatment after the patient is discharged from the hospital, the test results and lab reports of the patient during in-hospital care, and the many other aspects of patient care that had not been captured in the structured part of the EHR. The information in the notes can be found in the form of descriptive

and semi-structured format. These data can be mined for genomic research purposes as well. In fact, Denny (2012) showed that "when linked to biological data such as DNA and tissue biorepositories, EHRs can become a powerful tool for genomic analysis."

1.3.1.2 Text preprocessing

The document set obtained is prepared for processing. First, the document text is tokenized. Tokenization is the division of a text into meaningful units called "tokens." A token is a group of characters that is categorized according to a set of rules. For instance, NUMBER, COMMA and DOT are examples of token categories. It is an important task since all the following tasks will be based on tokens resulting from this process. Thus, several tokenization solutions were developed for several domains and languages. For instance, OpenNLP[7] has models for biomedical documents in English and Portuguese, and SPECIALIST NLP (Browne, McCray & Srinivasan 2000) supports English biomedical text. This process is also referred to as "feature generation."

Next, some words are removed. These words are called *stop words*. They consist of words that are frequently used, such as "it," "are," and "is." Though such words are quite common, since they are not useful in the classification of documents they are summarily removed (National Center for Biotechnology Information 2010).

Since in most cases morphological variants of words have similar semantic interpretations, which can be considered as equivalent, words are stemmed as part of preprocessing. Word stemming reduces inflected and derived word forms to their root or stem, mapping related words to the same stem, for example, the words "retrieval," "retrieve," and "retrieving" become *retrie* when stemmed.

1.3.2 Extraction of features and semantic information

This next phase usually starts with named entity recognition (NER), which aims to detect specific terms that represent relevant entities such as genes, proteins, diseases, and drugs. There still exist important challenges in named entity recognition that derive from the fact that there are different ways of referring to the same phenomena. For instance, "epilepsy" and "falling sickness" refer to the same disease: a central nervous system disorder characterized by the loss of consciousness (Zhu et al. 2013).

The natural language text of biomedicine, found in articles, books, reports, and other unstructured sources, present several challenges that

can make the application of information extraction and retrieval techniques even harder. The main challenge is related to terminology, and is a result of the complexity of the terms used in biomedical entities and processes (Zhou et al. 2004; Ananiadou & McNaught 2006):

- Non-standardized naming convention: an entity name could be found in various spelling forms (e.g., “N-acetylcysteine,” “N-acetyl-cysteine,” and “NAcetylCysteine”);
- Ambiguous names: a same name could be related with more than one entity, depending on the text context;
- Abbreviations: biomedical abbreviations are frequently used (e.g., “TCF” may refer to “T cell factor” or to “Tissue Culture Fluid”);
- Descriptive naming convention: many entity names are descriptive, which makes its recognition a complex task (e.g., “normal thymic epithelial cells”);
- Conjunction and disjunction: two or more entity names sharing one head noun (e.g., “91 and 84 kDa proteins” refers to “91 kDa protein” and “84 kDa protein”);
- Nested names: one name may occur within a longer name, as well as occur independently (e.g., “T cell” is nested within “nuclear factor of activated T cells family protein”)
- Names of newly discovered entities: there is an overwhelming growth rate and constant discovery of novel biomedical entities, which takes time to register in curated nomenclatures.

In general, there have been several approaches to NER in the clinical and biomedical literature. These can be roughly divided into the following four groups: (1) Dictionary-basedapproaches that try to find names of the well-known nomenclatures in texts; (2) Rule-based approaches that manually or automatically construct rules and patterns to directly match them to candidate named entities in the texts; (3) Machine learning approaches that employ machine learning techniques, such as Hidden Markov Models and Support Vector Machines, to develop models for NER; and (4) Hybrid approaches that merge two or more of the above approaches, mostly in a sequential way, to deal with different aspects of NER.

1.3.3 Analysis of annotated texts

In this next phase, various text mining techniques can be applied to the preprocessed data. Frequent tasks associated with this phase are the following:

Relation extraction: After having identified named entities, several

information extraction tasks in the biomedical domain involve determination of relationships among those entities. The goal of the relation extraction task is to identify occurrences of particular types of relationships between pairs of entities. Although common entity classes, such as genes or drugs, are in general quite specific, relations may be broad, including any type of biomedical association. Alternatively, such relations may be very specific, for example, by characterizing only gene regulatory associations (Simpson & Demner-Fushman 2012). Relation extraction approaches have shown an evolution from simple systems that rely solely on co-occurrence statistics to complex systems utilizing syntactic analysis and dependency parsing.

Event detection: Recently, there has been a shift in biomedical information extraction from recognizing binary relations to the more ambitious task of identifying complex, nested event structures. Events are typically characterized by verbs or nominalized verbs. For example, in the sentence “glnAP2 may be activated by NifA,” the verb activated specifies the event, and glnAP2 and NifA are the event’s arguments. Unlike the case of simple binary relations, both concept labels and semantic roles are assigned to an event and its arguments. In this example, the verb activated indicates a positive regulation type event, which expects a protein (NifA) to act as the event’s cause and a gene (glnAP2) to act as the event’s theme (Ananiadou et al. 2010).

Semantic search and inference: Search in large collections of documents, as those in biomedical and health domains, presents a series of challenges. A highly relevant one is vocabulary mismatch because it can severely decrease the performance of keyword-based search. This can happen when a user’s query contains little or no shared terms with relevant documents for that query. For example, when querying “lung cancer treatment,” documents using specialized terms such as “lung excision” or “chemotherapy” may receive a low rank or even be left out of the result set altogether. Vocabulary mismatch is dealt with by using techniques such as query term expansion and inference (Liu & Chu 2007; Koopman et al. 2011).

Text summarization: Medical information is often fragmented, existing in a wide range of locations and formats. This fragmentation makes the creation of an optimal clinical summary more challenging (Feblowitz et al. 2011). The availability of a great amount of clinical information that can be accessed rapidly increases the risk of inefficacy due to information overload (Hall & Walton 2004). This problem is likely to increase over time with the sharing of patient data more broadly. This makes clinical text summarization an important task. It can be divided

into three interrelated categories: source-oriented, time-oriented and concept-oriented views (Feblowitz et al. 2011).

Text clustering: The objective is to organize text in a small number of meaningful clusters of the same type or class. Classes are usually obtained from the set of relevant and frequent words of the text, and thus the number of classes that will be assigned is not known beforehand. Text clustering finds applicability for a number of tasks, such as document organization and browsing, corpus summarization, and document classification (Simpson & Demner-Fushman 2012).

Automated text categorization: Is the process of assigning unseen documents to user-defined categories. An important goal in biomedical text mining is automatic classification of electronic documents. Computer programs scan text in a document and generate a model that assigns the document to one or more pre-specified topics/categories using classification techniques. Those categories are usually organized in taxonomies (Fang, Parthasarathy & Schwartz 2001). Text classification, adopted as an example, is the subject of next section.

1.3.3.1 Algorithms for text classification

Several approaches have been proposed. Text classification is based on the supervised learning model. In this learning the total documents are divided into two parts. One part is called “training data” and the other part is called “test data.” A model or classifier is generated with training data. Once a classifier is created, it is applied to test the dataset in order to calculate the accuracy of the classifier. The frequently used text classification algorithms are Naïve Bayesian, k-NN, Decision Trees, and SVM.

Naive Bayesian (NB) algorithm

Naïve Bayesian (NB) algorithm has been generally used for text classification. This algorithm is based on Bayes’ theorem and is used to predict the probability of categories for a given document. The classifier predicts posterior probability of documents for each category and assigns the category which has highest posterior probability. Naive Bayesian classifier assumes that the effect of the probability of the term on a given category is independent of the probability of the other terms in the same category (Zhang, Chen & Xiong 2007; Yuan 2010).

There are two versions of the NB algorithm. One is the multi-variate Bernoulli event model that only takes into account the presence or absence of a particular term so that it doesn’t capture the number of

occurrences of each word. The other model is the multinomial model that captures the word frequency information in documents. Li and Jain (1998) showed that Naïve Bayesian classifier does not provide efficient classification with smaller training data sets. If the training set is limited in size, then there may be a chance that the term frequency of some of words will become zero and, at the same time, the probability of the word in a given category also becomes zero.

k-Nearest Neighbor algorithm (k-NN)

k-NN classifier is an instance-based learning algorithm that is based on a distance function for pairs of observations, such as the Euclidean distance or cosine. In this classification process, an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors. k is a positive integer, typically small. If k = 1, then the object is simply assigned to the class of its nearest neighbor. In binary classification problems, it is helpful to choose k to be an odd number as this avoids tied votes.

"One of the advantages of k-NN is that it is well suited for multi-modal classes as its classification decision is based on a small neighborhood of similar objects. So, even if the target class is multi-modal it can still lead to good accuracy."[8] The drawback of k-NN is that it uses all features in the documents to compare them. It affects the similarity measure and consequently the efficiency of classification. The traditional k-NN text classification algorithmic limitations are: calculation complexity mainly due to the usage of all the training samples for classification; dependency on the training set; and equal weighting of all samples. To overcome these challenges researchers developed variations of k-NN algorithms.

Decision trees

Decision trees are one of the most widely used inductive learning methods. Decision tree algorithms are suitable for document classification because of their robustness to noisy data. Two widely known algorithms for building decision trees are classification and regression trees. ID3 and its successor C4.5 (Quinlan 1993) and booster version of C 4.5 (Quinlan 1998) are famous for classification. It is a top-down approach which recursively constructs a decision tree classifier. At each level of the tree, ID3 selects the attribute that has the highest *information gain.* "ID3 is a supervised machine learning algorithm that automatically derives a decision tree from a set of training instances once each instance is tagged

with its correct classification. A fully trained decision tree can then be used to classify previously unseen instances from a test set" (Lehnert et al. 1995). The tree tries to split the training data based on the values of the available features to produce a good generalization. The node which has highest information gain is used to make a split. Each leaf node represents a class label. The given document is classified by following a path from the root node to a leaf node, where at each node a test is performed on some feature of that document. The leaf node reached is considered as the class label for that document. Decision tree algorithms are suitable for both binary and multiclass classification.

Support vector machines

Support vector machine (SVM) is a popular technique for classification. In recent years, the SVM has become an effective tool for pattern recognition, machine learning, and data mining because of its high generalization performance. The goal of SVM is to produce a model that predicts target value of data instances in the testing set, which are only given the attributes. Support vector machines (SVM) is a new technique for data mining, which has received increasing popularity in the machine learning and statistics community. SVM has been introduced by Vapnik (1995) for solving pattern recognition and nonlinear function estimation problems. SVM has become the tool of choice for the fundamental classification problem of machine learning and data mining. "Unlike traditional methods which minimize the empirical training error, SVM aims at minimizing an upper bound of the generalization error through maximizing the margin between the separating hyperplane and the data. This can be regarded as an approximate implementation of the structure risk minimization principle" (Wang, Chen & Chen 2004, p. 512).

Support vector machines are among the most robust and successful classification algorithms. They are based upon the idea of maximizing the margin i.e., maximizing the minimum distance from the separating hyperplane to the nearest example. The basic SVM supports only binary classification, but several extensions of these algorithms can deal with multiclass classification as well (Bredensteiner & Bennett 1999). SVM is frequently used in the medical domain. For example, it is used to generate a decision support system for heart disease classification (Bhatia, Prakash & Pillai 2008).

1.3.3.2 Classification evaluation measures

The evaluation is essential for understanding the quality of the learning

model, for tuning the parameters in the iterative process of classification, and for selecting the best model. There are several measures for evaluating models such as complexity, computational cost, computational time, mean absolute error, sensitivity, specificity, and accuracy.

Confusion matrix

A classification model classifies each instance into one of the classes. The confusion matrix shows how the predictions are made by the model. The rows correspond to the class labels in the data set. The columns show the predictions made by the model. The value of each element in the matrix is the number of predictions made with the class corresponding to the column. Thus, the diagonal elements show the number of correct classifications made for each class, and the off-diagonal elements show the errors made.

There are four possible classifications for each instance: i.e., true positive, true negative, false positive, and false negative. This is represented in matrix form and is called confusion matrix. If the accuracy of the classification model is 100% then all predictions are correct, which means that false positives and false negatives have a value of zero. The below Tab. 1.1 shows how the results are tabulated in a confusion matrix.

Mean absolute error

The mean absolute error (MAE) is a quantity used to measure how close forecasts or predictions are to eventual outcomes. The mean absolute error is given by

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|f_i - y_i\right| = \frac{1}{n}\sum_{i=1}^{n}\left|e_i\right|$$

The mean absolute error is an average of the absolute error $e_i = |f_i - y_i|$, where f_i is the prediction and y_i is the true value.

Kappa statistics

Kappa is a chance-corrected measure of agreement between the classifications and the true classes. It is calculated by taking the agreement expected by chance away from the observed agreement and dividing by the maximum possible agreement:

$$K = \frac{P_o - P_c}{1 - P_c}$$

where P_o is the proportion of observed agreement and P_c is the proportion of agreements expected by chance. A value greater than "0" means the classifier is doing better than chance.

Tab. 1.1: Confusion matrix.

		Observed	
		True	False
Predicted	True	True Positive rate (tp)	False Positive rate (fp)
	False	False Negative rate (fn)	True Negative rate (tn)

Accuracy

Accuracy is the overall correctness of the model and is calculated as the sum of correct classifications divided by the total number of classifications:

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$

Precision

Precision is a measure of the accuracy provided that a specific class has been predicted. Precision is the probability that a retrieved document is relevant. From the confusion matrix it is calculated by:

$$Precision = \frac{tp}{tp + fp}$$

where *tp* and *fp* are the numbers of true positive and false positive predictions for the considered class. Precision is 1 when *fp* is 0, which indicates there were no spurious results.

Recall

Recall is the probability that a relevant document is retrieved in a search. Recall is also referred to as the true positive rate or sensitivity and is given by:

$$Recall = \frac{tp}{tp + fn}$$

Recall becomes 1 when *fn* is 0, and it indicates that 100% of the *tp* were

discovered.

F-measure
The F-measure is the harmonic mean of precision and recall. It is calculated by using the formula:

$$F = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

The behavior of the performance measures is the function of the decision threshold for classification. When decision threshold increases, the recall will increase and precision will decrease.

1.3.4 Presentation

In this last phase, the result of classification is represented in graphical format, so that even non-technical people can also easily interpret the result. There are several presentation tools available. These tools are also called *data visualization tools*. Some of them are Plotly,[9] IBMMany Eyes,[10] Grapheur,[11] Visumap[12], etc. These tools are not only used to represent the relationships and co-relations, but they are also used to represent patterns of data.

1.4 Text mining tools

Text mining tools help in discovering structure and patterns in unstructured data –usually text. These tools are available from many commercial and open source companies. Some relevant general-purpose tools are:

SAS Text miner: This tool extracts knowledge from unstructured data with text mining software. It provides interactive GUIs which makes it easy to identify relevance, modify algorithms, document assignments, and group materials into meaningful aggregations. This makes it easy for the user to guide machine-learning results with human insights. It extends text mining efforts beyond basic start-and-stop lists by using custom entities and term-trend discovery to refine automatically generated rules.[13]

NetOwl Text Analytics: NetOwl offers a suite of best-of-breed text and entity analytics products. “NetOwl analyzes Big Data in the form of text data – news, email, web, social media, and any other text document that

organizations would like to exploit as well as structured entity data about people, organizations, places, and things."[14] It provides tools to analyze an extremely large volume of data in a variety of forms and languages and offers advanced text analytics products to meet today's Big Data challenges.

IBM Intelligent Miner: IBM Intelligent Miner for Text is a knowledge discovery software development toolkit. It contains tools for application programmers who want to build applications to extract key information from very large quantities of documents, e-mails, or Web pages stored online, often on the Internet or on intranets, without having to read them all. IBM Text Analysis Tools include a Language Identification tool, comprehensive Clustering tools, a Topic Categorization tool, a Summarization tool, and Feature Extraction tools. These tools identify document language, group conceptually related documents, classify documents by content, generate document summaries, and extract key elements of text.[15]

Weka: WEKA is an open-source machine learning tool. It was developed at the University of Waikato, New Zealand to implement data mining algorithms. WEKA is a state-of-the-art facility for developing machine learning (ML) techniques and their application to real-world data mining problems. WEKA is a collection of machine learning algorithms for data mining tasks. The algorithms are applied directly to a dataset. WEKA implements algorithms for data preprocessing, classification, regression, clustering, and association rules; it also includes visualization tools. The new machine learning schemas can also be developed with this package. WEKA is open-source software issued under General Public License.[16]

Adding to these general purpose tools, some specialized tools were developed for specific topics related to biomedical and health domains. Simpson and Demner-Fushman (2012) present a comprehensive review of recent works; an extensive list can be found in the Bio-NLP resources database.[17] Some relevant systems are:

Becas: becas[18] is a web application, API, and widget for biomedical concept identification that helps researchers, healthcare professionals, and developers in the identification of over 1,200,000 biomedical concepts in text and PubMed abstracts (Nunes et al. 2013). It provides annotations for isolated, nested, and intersected entities, and identifies concepts from multiple semantic groups. It has the ability to provide preferred names for concept identification and is able to enrich them with references to public knowledge resources.

KLEIO: enhances search facilities across the MEDLINE collection by identifying key entities within the text, such as gene names or proteins,

and improves the querying method with unique identifiers by automatically including synonyms, spelling variants and, even, disambiguating acronyms (Nobata et al. 2008). It combines these features with the common features found in other interfaces to provide a solution to the growing problem of finding valuable information within the ever increasing volume of modern publications.[19]

PIE the search: *PIE* (Protein Interaction information Extraction) *the search* is a web service to extract protein-protein interaction relevant articles from MEDLINE (Kim et al. 2012). It accepts PubMed input formats to make available up-to-date protein-protein interaction information which cannot be found in manually curated databases. *PIE the search* is targeted at providing protein-protein interaction relevant articles for biologists, baseline system performance for bio-text mining researchers, and a compact PubMed-search environment for PubMed users.[20]

MEDIE: is a framework for accurate, real time, retrieval of relational concepts from MEDLINE (Miyao et al. 2006). Prior to retrieval, a semantically annotated text base is prepared and stored in a structured database. The preparation of the text base includes applying natural language processing tools, including deep parsers and term recognizers. User requests are converted on the fly into patterns of these semantic annotations, and texts are retrieved by matching these patterns with the pre-computed semantic annotations. Real-time retrieval is possible because semantic annotations are computed in advance.[21]

MedInX: is a Medical Information eXtraction system tailored to process textual clinical discharge records, performing automatic and accurate mapping of free reports onto a structured representation (Ferreira, Teixeira & Cunha 2012). MedInX is designed to be used by health professionals, and by hospital administrators and managers, allowing a search of the contents of its automatically populated ontologies. (Further details on this system can be found in Chapter 3 of this book.)

NextBio: aggregates large quantities of genomic data for research and clinical applications. It contains the world's largest repository of curated and correlated public and private genomic data, including data from multiple public repositories of genomic studies and patient molecular profiles, up-to-date reference genomes, and clinical trial results (Kupershmidt et al. 2010). Several molecular data types from these resources are systematically processed, curated, and integrated into the data center based platform. This allows applying genomic data in novel and useful ways, both in the research laboratory and in the clinic.[22]

The Neuroscience Information Framework: is a dynamic inventory of Web-based neuroscience resources: data, materials, and tools (Akil,

Martone & Van Essen 2011). It helps in advancing neuroscience research by enabling discovery and access to public research data and tools worldwide through an open source networked environment. It offers the following: a search portal for researchers, students, or anyone looking for neuroscience information, tools, data, or materials; access to content normally not indexed by search engines; and tools for resource providers to make resources more discoverable, such as ontologies, data federation tools, and vocabulary services.[23]

1.5 Summary

This chapter shows how biomedical information is successfully retrieved by using text mining techniques. The sources of biomedical information, found in both clinical narratives and biomedical literature, and the available tools for text mining are described in this chapter, which highlights various text mining techniques and evaluation measures. Future work, however, requires an interdisciplinary approach to text mining of biomedical information. Such coordinated efforts of biologists and clinicians, medical researchers and epidemiologists, computer scientists and computational linguists, library scientists and statisticians, and others are imperative to exploit the full scientific potential of biomedical text mining. The field has promise but much more effort must be made in choosing tasks and evaluating results based on real-world requirements and needs. In the end it is the patient population and the public writ large who will reap the full benefits of the application of text mining tools that successfully perform biomedical knowledge extraction.

Appendix “A”

Open-sourced Structured Databases

- Diseases Database:[24] It provides Cross-referenced database of clinical medicine and it links to topic categorical pages from other websites.
- DynaMed:[25] A medical information database with over 2000 diseases.
- General Practice Notebook:[26] Database of clinical medicine with a search facility.
- ICD-9 Data:[27] Offers drillable dataset of ICD-9-CM medical

diagnosis codes.

- ICD-9 Search:[28] Search ICD-9 for medical diagnosis, codes, and procedures. Find related diseases, treatments and related news.
- ICD-9-CM Online:[29] Searchable database of disease classification.
- IndMED:[30] Indian Biomedical Journals Database: Bibliographic aggregation of peer-reviewed biomedical journals.
- OpenMED:[31] An international open-access archive of scientific and technical documents for Medical and Allied Sciences.
- AIDSLIN database: It provides the literature on AIDS and HIV back to 1980.
- AMED Database:[32] This database covers a range of complementary and alternative medicine including homeopathy, chiropractic, and acupuncture and so on.
- Bandolier:[33] Award-winning summary journal with searchable index produced by Andrew Moore and colleagues in Oxford, UK.
- Cochrane database.[34]
- English National Board Health Care Database:[35] A database of journal references primary of interest to nurses, midwives and health visitors.
- POPLINE database:[36] The world's largest online bibliographic database on population, family planning, and related health issues. It is also available in CD-ROM which is free of charge to developing countries.
- STRIDE Clinical Data Warehouse[37] is the source of historical clinical data from both hospitals for research purposes.

References

Akil, H., Martone, M. E. & Van Essen, D. **C.** (2011) 'Challenges and opportunities in mining neuroscience data', *Science*, 331:708–712.

Ananiadou, **S.** & McNaught, J. (2006) 'Text mining for biology and biomedicine', *Comput Ling*, 135–140.

Ananiadou, S., Pyysalo, S., Tsujii, J. & Kell, D. B. (2010) 'Event extraction for systems biology by text mining the literature', *Tends Boitech*, 28:381–390.

Atkinson, J., Ferreira, A. & Aravena, E. (2004) 'Discovering implicit intention-level knowledge from natural-language texts', *Knowl-*

Based Styt, 22:502–508.

Bhatia, S., Prakash, P. & Pillai, G. N. (2008) 'SVM based decision support system for heart disease classification with integer-coded genetic algorithm to select critical features', In *Proceedings of the World Congress on Engineering and Computer Science*, pp. 34–38.

Bredensteiner, E. J. & Bennett, K. P. (1999) 'Multi category classification by support vector machines', In *Computational Optimization*, Heidelberg, Germany: Springer. pp. 53–79.

Browne, A. C., McCray, A. T. & Srinivasan, S. (2000) 'The specialist lexicon', *Natl Libr Med Tech Rep*, 18–21.

Cohen, A. M. & Hersh, W. R. (2005) 'A survey of current work in biomedical text mining', *Briefings in Bioinformatics,* 6(1):57–71.

Coussement, K. & Poel, V. D. (2008) 'Integrating the voice of customers through call center e-mails into a decision support system for churn prediction', *Inform Manage*, 45(3):164–174.

Denny, J.C. (2012) 'Mining electronic health records in the genomics era', *PLoS Comput Biol,* 8(12).

Fang, Y. C., Parthasarathy, S. & Schwartz, F. (2001) 'Using clustering to boost text classification', In *ICDM Workshop on Text Mining* (TextDM'01).

Feblowitz, J. C., Wright, A., Singh, H., Samal, L. & Sittig, D. F. (2011) 'Summarization of clinical information: A conceptual model', *J Biomed Inform*, 44:688–699.

Feldman, R. & Sanger, J. (2006) '*The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*', Cambridge: Cambridge University Press.

Ferreira, L., Teixeira, A. & Cunha J. P. (2012) '*Medical Information Extraction: Information Extraction from Portuguese Hospital Discharge Letters*', Saarbrücken, Germany: Lambert Academic Publishing.

Gonzalez, R. B. (2008) 'Index Compression for Information Retrieval Systems', Ph.D. Thesis, University of A Coruña.

Gurulingappa, H., Toldo, L., Rajput, A. M., Kors, J. A., Taweel, A. & Tayrouz, Y. (2013) 'Automatic detection of adverse events to predict drug label changes using text and data mining techniques', *Pharmacoepidemiology Dr S*, 22:1189–1194.

Hakenberg, J., Voronov, D., Nguyên, V. H., Liang, S., Anwar, S., Lumpkin, B., Leaman, R., Tari L. & Baral, C. (2012) 'A SNPshot of PubMed to associate genetic variants with drugs, diseases, and adverse reactions', *J Biomed Inform*, 45:842–850.

Hall, A. & Walton, G. (2004) 'Information overload within the health care system: a literature review', *Health Inform Libr J*, 21:102–108.

Herrera-Viedma, E. (2001) 'Modeling the retrieval process for an information retrieval system using an ordinal fuzzy linguistic approach', *J Am Soc Inform Sci Tech*, 52(6):460–475.

Imambi, S. S. & Sudha, T. (2010) 'Building classification system to predict risk factors of diabetic retinopathy using text mining', *Int J Comput Sci Eng*, 2(7):2309–2312.

Imambi, S. S. & Sudha, T. (2011) 'Classification of Medline documents using global relevant weighting schema', *Int J Comput Appl*, 16(3):45–48.

Jensen, P. B., Jensen, L. J. & Brunak, S. (2012) 'Mining electronic health records: towards better research applications and clinical care', *Nat Rev Gen*, 13:395–405.

Kankar, P., Adak, S, Sarkar, A. & Sharma, G. (2002) 'MedMeSH Summarizer: Text mining for gene clusters', *Proceedings of the Second SIAM International Conference on Data Mining.*

Kim, S., Kwon, D., Shin, S.-Y. & Wilbur, W. J. (2012) 'PIE the search: searching PubMed literature for protein interaction information', *Bioinformatics*, 28:597–598.

Koopman, B., Bruza, P. D., Sitbon, L. & Lawley, M. (2011) 'Towards semantic search and inference in electronic medical records: an approach using concept-based information retrieval', *Proceedings of the 1st Australian Workshop on Artificial Intelligence in Health (AIH 2011*), pp. 1–11.

Krauthammer, M. & Nenadic, G. (2004) 'Term identification in the biomedical literature', *J Biomed Inform*, 37(6):512–526.

Kupershmidt, I., Qiaojuan, J. S., Grewal, A., Sundaresh, S., Halperin, I., Flynn, J., Shekar, M., Wang, H., Park, J., Cui, W., Wall, G. D., Wisotzkey, R., Alag, S., Akhtari, S. & Ronaghi, M. (2010) 'Ontology-based meta-analysis of global collections of high-throughput public data', *PLoS ONE*, 5.

Latha, K., Kalimuthu, S. & Rajaram, R. (2007) 'Information extraction from biomedical literature using text mining framework', *IJISE*, GA, USA, 1(1):1–5.

Lehnert, W., Soderland, S., Aronow, D., Feng, F. & Shmueli, A. (1995) 'Inductive text classification for medical applications', *J Exp Theor Artif In*, 7(1):49–80.

Li, Y. H. & Jain, A. K. (1998) 'Classification of text documents', *Comput J*, 41(8).

Liu, Z. & Chu, W. W. (2007) 'Knowledge-based query expansion to support scenario-specific retrieval of medical free text', *Inform Ret*, 10:173–202.

Liu, H., Lussier, Y. A. & Friedman, C. (2001) 'Disambiguating ambiguous biomedical terms in biomedical narrative text: an unsupervised method', *J Biomed Inform*, 34:249–261.

Lovis, C., Baud, R. H. & Planche, P. (2000) 'Power of expression in the electronic patient record: Structured data or narrative text?' *Int J Med Inform*, 58–59:101–110.

Manning, C., Raghavan, P. & Schutze, H. (2008) '*Introduction to Information Retrieval,*' Cambridge: Cambridge University Press.

Meystre, S. M., Savova, G. K., Kipper-Schuler, K. **C.** & Hurdle, J. F. (2008) 'Extracting information from textual documents in the electronic health record: a review of recent research', *Yearb Med Inform*, pp. 128–144.

Mitchell, T. M. (1997) '*Machine Learning,*' New York: McGraw-Hill.

Miyao, Y., Ohta, T., Masuda, K., Tsuruoka, Y., Yoshida, K., Ninomiya, T. &

Tsujii, J. (2006) 'Semantic retrieval for the accurate identification of relational concepts in massive text bases', In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL* - ACL '06. pp. 1017–1024.

Moumtzoglou, A. & Kastania, A. (2011) 'E-Health systems quality and reliability: models and standards', *Medical Information Science Reference*. New York: Hershey.

Nahm, U. Y. & Mooney, R. J. (2002) 'Text Mining with Information Extraction', *AAAI Tech Rep SS-02-06*, pp. 60–67.

Nunes, T., Campos, D., Matos, S. & Oliveira, J.L. (2013) 'BeCAS: b Quinlan, Biomedical concept recognition services and visualization', Bioinformatics, vol. 29, no. 15, p. 1915–1916, June 2013.

National Center for Biotechnology Information 2010 PubMed stop words.

Nobata, C., Cotter, P., Okazaki, N., Rea, B., Sasaki, Y., Tsuruoka, Y., Tsujii, J. & Ananiadou, S. (2008) 'Kleio: a knowledge-enriched information retrieval system for biology', In SIGIR '08: *Proceedings of the 31st annual international ACM SIGIR Conference on Research and Development in Information Retrieval,* pp. 787–788.

Nunes, T., Campos, D., Matos, S. & Oliveira, J. L. (2013) 'BeCAS: biomedical concept recognition services and visualization', *Bioinformatics*, 29:1915–1916.

Quinlan, J. (1993) '*C4.5: Programs for machine learning*', Morgan Kaufmann: San Matteo, CA.

Quinlan, J. R. (1998) 'Mini boosting decision trees', *J Artif Intell Res*, 1–15.

Ramampiaro (2010) 'Retrieving biomedical information with BioTracer: Challenges and possibilities', *NIK-2009*.

Rebholz-Schuhmann, D., Oellrich, A. & Hoehndorf, R. (2012) 'Text-mining solutions for biomedical research: enabling integrative biology', *Nat Rev Gen*, 13:829–839.

Simpson, M. S. & Demner-Fushman, D. (2012) 'Biomedical text mining: a survey of recent progress', In C. C. Aggarwal and C. X. Zhai (eds.), *Mining Text Data*, Heidelberg: Springer Verlag, pp. 465–517.

Singhal, A. (2001) 'Modern information retrieval: a brief overview', *IEEE Data Eng Bull*, 24(4):35–43.

Srinivasan, P., Bisharah, L. & Sehgal, A. (2004) 'Mining MEDLINE: Postulating a beneficial role for curcumin longa in retinal diseases', Boston, MA: *Workshop: Biolink, Linking Biological Literature, Ontologies and Databases*, pp. 33–40.

Srinivasan, P. & Libbus, B. (2004) 'Mining MEDLINE for implicit links between dietary substances and diseases', *Bioinformatics*, 20:290–296.

Suominen, H. (2009) 'Machine learning and clinical text: Supporting health information flow', *TUCS Dissertations,* (125).

Thoroddsen, A., Saranto, K., Ehrenberg, A. & Sermeus, W. (2009) 'Models, standards and structures of nursing documentation in European countries', *Stud Health Tech Inform*, 146:327–331.

Van Rijsbergen, C. J. (1979) *Information Retrieval*, 2nd edition, Newton, MA: Butterworth Heinemann.

Vapnik, V. (1995) '*The Nature of Statistical Learning Theory*', 2nd edition, Heidelberg, Germany: Springer-Verlag. pp. 138–141.

Vazquez, M., Krallinger, M., Leitner, F. & Valencia, A. (2011) 'Text mining for drugs and chemical compounds: Methods, tools and applications', *Molecular Informatics*, 30:506–519. Available at: http://doi.wiley.com/10.1002/minf.201100005.

Walsh, S. H. (2004) 'The clinician's perspective on electronic health records and how they can affect patient care', *Br Med J*, 328:1184–1187.

Wang, J., Chen, Q. & Chen, Y. (2004) 'RBF kernel based Support Vector Machine with universal approximation and its application', In *Lecture Notes in Computer Science 3174*, F. Yin, J. Wang, & C. Guo (eds.), Springer Verlag: Heidelberg.

Yuan, L. (2010) 'An improved Naive Bayes text classification algorithm in Chinese information processing', *Proceedings of the Third International Symposium on Computer Science and Computational Technology (ISCSCT'10).*

Zanasi, A. (2009) 'Virtual weapons for real wars: Text mining for national security', *Proceedings of the International Workshop on Computational Intelligence in Security for Information Systems CISIS'08, Advances in Soft Computing*, 53:53–60.

Zhang,Y, Chen, J. & Xiong (2007) 'Improved Naive Bayes text classification algorithm', *J Guangxi Normal University* (Natural Science Edition), 2.

Zhao, L.-L., Zhang, T., Zhuang, L.-W., Yan, B.-Z., Wang, R.-F. & Liu, B.-R. (2014) 'Uncovering the pathogenesis and identifying novel targets of pancreatic cancer using bioinformatics approach', *Mol Biol Rep*, 1–8.

Zhou, G., Zhang, J., Su, J., Shen, D. & Tan, C. L. (2004) 'Recognizing names in biomedical texts: a machine learning approach', *Bioinformatics*, 20:1178–1190.

Zhu, F., Patumcharoenpol, P., Zhang, C., Yang, Y., Chan, J., Meechai, A., Vongsangnak, W. & Shen, B. (2013) 'Biomedical text mining and its applications in cancer research', *J Biomed Inform*, 46:200–211.

Hua Xu and Joshua C. Denny

2 Unlocking information in electronic health records using natural language processing: a case study in medication information extraction

Abstract: Clinical natural language processing (NLP), which can unlock detailed patient information from clinical narratives stored in electronic health records, has been frequently used to support clinical research and operations. This chapter introduces the state-of-the-art work in clinical NLP. Using medication information extraction as a use case, we describe different methods to build clinical NLP systems, including rule-based, machine learning-based, and hybrid approaches. Applications of medication information extraction systems, such as *pharmacovigilance* (post-market surveillance of drugs) research, are also discussed in this chapter.

2.1 Introduction to clinical natural language processing

Electronic health record (EHR) systems have been increasingly adopted in the United States and worldwide (Jha et al. 2009; Shea and Hripcsak 2010). This growth is fueled, in part, by recent federal legislation that provides significant financial incentives to institutions demonstrating aggressive application and “meaningful use” of comprehensive EHRs (http://www.hhs.gov/news/press/2010pres/07/20100713a.html). The ever-growing availability of EHR data has become an enabling resource for clinical and translational research (Kohane 2011). However, the majority of EHR data is narrative text, given that clinical documentation is the primary form of communication in clinical practice. Unstructured clinical texts contain rich patient information, though such texts are not immediately accessible to computerized applications that rely on structured inputs, such as decision support systems and healthcare analytic tools. As a result, there has been a great interest in developing clinical natural language processing (NLP) methods to unlock information embedded in clinical narratives (Meystre et al. 2008; Nadkarni et al.

2011).

Various clinical NLP systems have been developed in past decades to extract information from clinical narratives to facilitate patient care and clinical research. The Linguistic String Project (LSP) (Sager et al. 1987, 1994) led by Naomi Sager at New York University was one of the earliest attempts to formulate comprehensive semantic and syntactic rules to parse clinical text. Later, Friedman and her colleagues (1994) developed a clinical NLP system called MedLEE (Medical Language Extraction and Encoding System), which was originally designed for decision-support applications in the domain of radiology reports of the chest. MedLEE has been shown to be as accurate as physicians at extracting clinical concepts from chest radiology reports (Hripcsak et al. 2002). It is routinely used to process and encode clinical text at New York Presbyterian Hospital. MPLUS and its ancestor SymText (Haug et al. 1995) are NLP systems developed at the University of Utah, which have been used for various applications such as encoding chief complaints into ICD-9 codes and extracting pneumonia-related findings from chest radiograph reports (Fiszman et al. 2000). KnowledgeMap Concept Indexer (Denny et al. 2003), an NLP system developed at Vanderbilt University Medical Center (VUMC) around 2000, has been used at Vanderbilt to extract clinical concepts from clinical documents (Denny et al. 2005). Other research groups have also developed various NLP systems for processing clinical text in different sub-domains of medicine (Hahn et al. 2002; Zeng et al. 2006; Harkema et al. 2009; Yetisgen-Yildiz et al. 2013).

Two widely used clinical NLP systems that are freely available to the public are MetaMap and cTAKES (clinical Text Analysis and Knowledge Extraction System). MetaMap (Aronson 2001; Aronson and Lang 2010) is a general biomedical NLP system developed by Aronson et al. at National Library Medicine. It was originally developed to map biomedical literature (e.g., MEDLINE abstracts) to concepts in the Unified Medical Language System (UMLS) Metathesaurus. Many researchers have used MetaMap to extract information from clinical text (Schadow and McDonald 2003; Chung and Murphy 2005; Meystre and Haug 2006; Friedlin and Overhage 2011). For example, Meystre and Haug (2006) applied MetaMap to extract medical problems from clinical text and reported a recall of 0.74 and a precision of 0.76. cTAKES (Savova et al. 2010b) is another freely available comprehensive clinical NLP system, which is developed on the Unstructured Information Management Architecture (UIMA, http://uima.apache.org/) framework and the OpenNLP toolkit (http://opennlp.apache.org/). cTAKES is a pipeline-based system that consists of different modules such as sentence boundary detector, part-of-

speech tagger, shallow parser, and named entity recognizer. Many studies have reported the use of cTAKES for different clinical information extraction tasks such as determining patient smoking status (Savova et al. 2008) and identifying disease cohorts (Savova et al. 2010a).

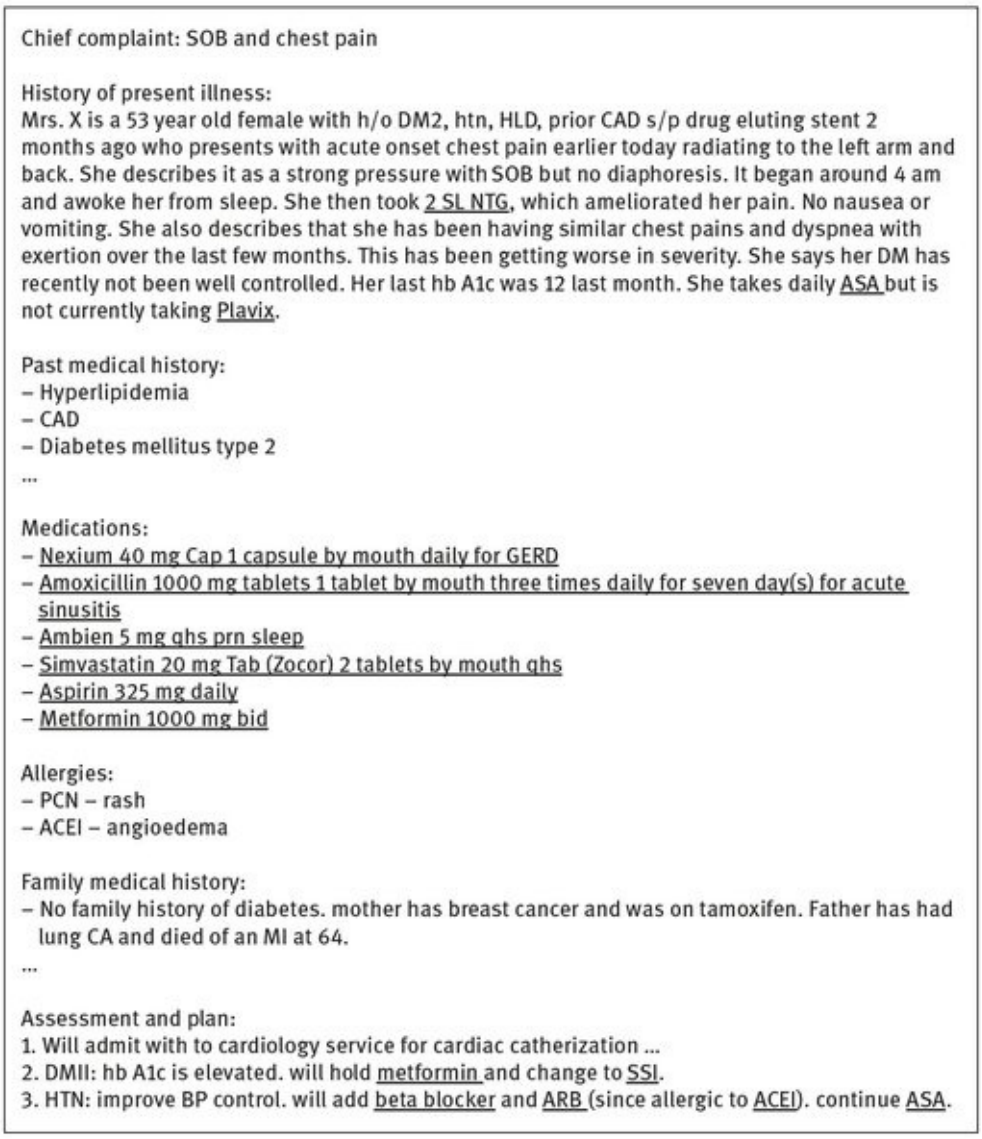
General-purpose clinical NLP systems such as MedLEE, MetaMap, and cTAKES are often comprehensive, requiring different methodologies for various components. To better describe methods in clinical NLP research, we decided to focus on a more narrowly defined topic. In this chapter, we will use medication information extraction as a use case to explain how state-of-the-art clinical NLP systems work.

2.2 Medication information in EHRs

The use of computer applications for recording and processing drug information are becoming increasingly available in most EHRs. For the inpatient setting, computerized provider order entry (CPOE) systems and electronic medication administration record (eMAR) systems have been widely adopted. Many EHR systems have also incorporated e-prescribing systems in the outpatient setting, which create structured records during generation of new prescriptions and refills. Adoption is increasing, as Meaningful Use Stage 1 requires that 40% of permissible prescriptions are generated and sent to pharmacies electronically by e-prescribing tools. Nevertheless, e-prescribing tools are still not yet widely adopted by physicians. Furthermore, it is often the case that historical medication information is not even generated through the use of such tools. Currently, outpatient medication information is frequently recorded via narrative text entries within clinical documentation or patient problem lists. Not surprisingly, many times this information is transmitted in the course of communications with the patient through telephone calls or patient portals for which there is corresponding notation in the patient's file. For all these reasons given above, an accurate construction of a patient's medication exposure history often requires extraction of information embedded in clinical narratives.

Figure 2.1 shows an example of an outpatient clinic visit note, with medication information highlighted using the underline. As shown in the example, some medication mentions are recorded in a semi-structured list (e.g., in the MEDICATIONS section); while other medications are recorded in narrative sentences (e.g., in the ASSESSMENT AND PLAN section). In the MEDICATION section, a medication entry often contains the medication name (generic or brand) and its signature information, such as dose, form,

route, frequency and, sometimes, the reason(s) for giving the patient the medication. Though context-specific information such as reasons or duration of a medication are also important, it is often much more challenging to extract from the patient's record. To complicate things further, medications are often mentioned in the patient's file for other purposes than for treating the medical condition at hand. For example, as shown in Fig. 2.1, medications are mentioned in the patient's file for a variety of reasons: (1) to indicate possible allergies or adverse reactions; (2) to formulate a family medical history: the name of the medication taken by the patient's mother is useful for defining the exact kind of breast cancer the mother had; and (3) medications that the patient is *not* taking may be mentioned in the "history of present illness" section of patient's chart to indicate possible lapse in medical care that must be addressed. For example, the blood thinner "Plavix" (the brand name for clopidogrel) for treating cardiac problems appears in this section of the patient's record.



Chief complaint: SOB and chest pain

History of present illness:
Mrs. X is a 53 year old female with h/o DM2, htn, HLD, prior CAD s/p drug eluting stent 2 months ago who presents with acute onset chest pain earlier today radiating to the left arm and back. She describes it as a strong pressure with SOB but no diaphoresis. It began around 4 am and awoke her from sleep. She then took 2 SL NTG, which ameliorated her pain. No nausea or vomiting. She also describes that she has been having similar chest pains and dyspnea with exertion over the last few months. This has been getting worse in severity. She says her DM has recently not been well controlled. Her last hb A1c was 12 last month. She takes daily ASA but is not currently taking Plavix.

Past medical history:
– Hyperlipidemia
– CAD
– Diabetes mellitus type 2
...

Medications:
– Nexium 40 mg Cap 1 capsule by mouth daily for GERD
– Amoxicillin 1000 mg tablets 1 tablet by mouth three times daily for seven day(s) for acute sinusitis
– Ambien 5 mg qhs prn sleep
– Simvastatin 20 mg Tab (Zocor) 2 tablets by mouth qhs
– Aspirin 325 mg daily
– Metformin 1000 mg bid

Allergies:
– PCN – rash
– ACEI – angioedema

Family medical history:
– No family history of diabetes. mother has breast cancer and was on tamoxifen. Father has had lung CA and died of an MI at 64.
...

Assessment and plan:
1. Will admit with to cardiology service for cardiac catherization ...
2. DMII: hb A1c is elevated. will hold metformin and change to SSI.
3. HTN: improve BP control. will add beta blocker and ARB (since allergic to ACEI). continue ASA.

Fig. 2.1: An example of outpatient clinic visit note, where medication information is highlighted with underline.

2.3 Medication information extraction systems and methods

2.3.1 Relevant work

Early studies on medication information extraction in EHRs have been focused on identifying drug names and selected signature information such as dosage. For example, Chhieng et al. (2007) used a string-matching method to identify drug names in clinical records and reported a precision rate of 83%. Levin and colleagues (2007) extracted drug names from anesthesia records and reported high performance with a sensitivity of 92.2% and a specificity of 95.7%. Evans et al. (1996) developed the

CLARIT system and showed that it could extract drug name and dosage phrases in patient discharge summaries with an accuracy of 80%.

More recent studies extended the scope to additional drug signature information such as *route* and *frequency*. Gold et al. (2008) developed a regular expression based approach to extracting drug names and signature information including dose, route, and frequency. They evaluated the system using a data set of 26 discharge summaries and showed that drug names were identified with a precision of 94.1% and a sensitivity of 82.5%, but other signature information such as dose and frequency had much lower precisions. In a study by Jagannathan et al. (2009), several commercial systems, such as LifeCode™, FreePharma™, and Coderyte, were assessed for their ability to extract medication information (including drug names, strength, route, and frequency) from clinical notes. Their evaluation showed a high F-measure of 93.2% on capturing drug names, but lower F-measures of 85.3%, 80.3%, and 48.3% on retrieving strength, route, and frequency, respectively (Jagannathan et al. 2009). At VUMC, a medication information extraction system called MedEx (Xu et al. 2010) was developed. It achieved F-scores over 90% on extracting drug names, strength, route and frequency information in discharge summaries and clinic visit notes from VUMC's EHR.

In 2009, i2b2 (Center of Informatics for Integrating Biology and Beside) organized a clinical NLP challenge to extract medication-related information in discharge summaries from Partners Healthcare (Uzuner et al. 2010). The goal of the challenge was to identify and determine boundaries of six types of drug information, which consisted of 1) drug name, 2) dosage, 3) route, 4) frequency, 5) duration; and 6) reason for drug administration. In addition, it required determination of whether medication information was found in a list or a narrative sentence. Figure 2.2 (Uzuner et al. 2010) shows the examples of inputs and outputs of the 2009 i2b2 challenge. Twenty teams, including international entries, participated in the medication challenge using various approaches including rule-based, machine learning and hybrid methods (see Section 2.3.2) (Deleger et al. 2010; Doan et al. 2010; Hamon and Grabar 2010; Li et al. 2010; Meystre et al. 2010; Mork et al. 2010; Patrick and Li 2010; Spasic et al. 2010; Tikk and Solt 2010; Yang 2010). While names of medications were well identified by all of the top 10 systems, the performance for durations and reasons were still low, with the best F-measure of 0.525 and 0.459, respectively (Uzuner et al. 2010).

The i2b2 medication challenge created an important asset to enhance research in this area by generating an annotated dataset for medication information extraction in EHRs. Using the i2b2 data set, researchers

investigated different aspects of machine learning-based approaches to medication-entity recognition, including different machine learning algorithms (Doan 2010) and the ensemble method, which combines predictions from multiple classifiers (Doan et al. 2012). Furthermore, Li et al. (2013) extended such medication information extraction methods to other clinically relevant text such as clinical trial documents.

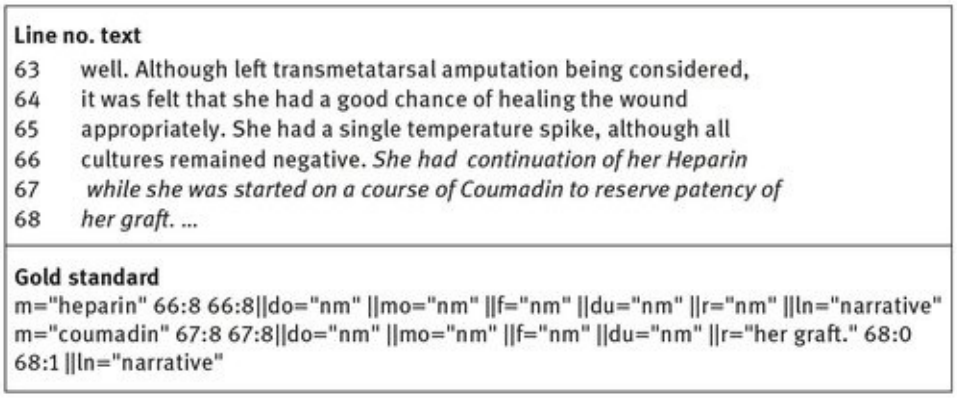


Fig. 2.2: Examples of the input and output in the 2009 i2b2 medication information extraction challenge. The challenge requires to identify six types of drug information including drug name (m), dosage (do), route (mo), frequency (f), duration (du), and reason (r), as well as their exact offset (by line number and token position) in the clinical documents. The figure was taken from (Uzuner et al. 2010).

2.3.2 Summary of approaches

Although many systems have been developed to extract medication information from clinical text, their methodological approaches can be mainly divided into three categories: rule-based (Gold et al. 2008; Deleger et al. 2010; Mork et al. 2010; Spasic et al. 2010; Xu et al. 2010), machine learning-based (Li et al. 2010; Patrick and Li 2010; Li et al. 2013), and hybrid methods (Meystre et al. 2010; Tikk and Solt 2010).

2.3.2.1 Rule-based methods

A rule-based medication information extraction system often works as following: (1) identify medication-related entities by using rules and dictionaries (e.g., lists of drug names) based on domain-specific resources; (2) filter medication entries based on context-specific information; and (3) link signature modifiers to corresponding drug names using specific rules.

Medication name identification is the crucial step in medication information extraction. Rule-based systems often leverage existing medical knowledge to build a comprehensive list of drug names. In Mork et al. (2010), the drug dictionary was built on various medical resources such as UMLS (http://www.nlm.nih.gov/research/umls), RxNorm

(http://www.nlm.nih.gov/research/umls/rxnorm), and DailyMed (http://dailymed.nlm.nih.gov). While lexicons for some signature fields such as route could be built in a similar way by creating an exhaustive list; other signature fields such as dose and frequency have to be recognized by defining regular expression patterns. Figure 2.3 shows some examples of regular expression rules used in Yang's (2010) system for recognizing frequency expressions. In the system developed by Spasic and colleagues (2010), all rules were implemented as expression in Mixup (My Information eXtraction and Understanding Package), a simple pattern-matching language.

After a possible drug entity is recognized, context-based rules have to be applied to verify its inclusion. A drug entry could be excluded in the final output due to several reasons, such as drug-allergy information, negation, and non-patient experience (e.g., about a family member). Specific rules could be developed based on context information, e.g., to remove drugs in the Allergy section. In addition, drug names could be ambiguous as well. For example, in the sentence "The patient was found to be iron deficient and she continued on iron supplements," the first occurrence of "iron" should not be labeled as a medication, as it is associated with "deficient," a term that is used here to indicate a medical problem in this particular patient. However, the second occurrence of "iron" should be marked as a *medication*, because it is collocated with the word "supplements." What we learn from this is that it is very import to construct rules to resolve such ambiguities so that we can improve the performance of the medication information extraction system.

Token-based rule	Example
[after|before|at|following|with|w/] + <Meal>	after breakfast, before meals, at supper, following lunch
[in|on|at|during]+<Daytime>	in the a.m., at bedtime, on p.m., during the evening
[each|every|on]+<Weekday>	each Monday, every Sunday, on tues,
[every]+<Num>+<TimeUnit>	every 3 hour, every 3–5 min
<Num>+[x|x/]+<TimeUnit>	2×/wk, 2–3×/day, 2×wk
[q|q.]+<TimeUnit>	qhr, q day, q.wk, q. week
[q|q.]+<Num>+<TimeUnit>	q2h, q 4 h, q. 2 weeks, q.6 h
[q|q.]+<Meal>	qlunch, q breakfast, q.meal, q. dinner
[q|q.]+<Daytime>	qam, q p.m., q. afternoon, q.evening
[q|q.]+<Weekday>	qwed, q monday, q. friday, q.saturday
[once|twice] + <OneTimeUnit>	once a day, twice per day
<Num>+[times|x] +<OneTimeUnit>	2 times a day, 3×daily

Fig. 2.3: Examples of regular expressions for recognizing frequency expressions, as specified in Yang (2010).

The last step is to link drug names with their corresponding signature modifiers. A simple but effective approach is to use regular expression to recognize drug names together with their associated signature fields, as implemented in the MERKI system developed by (Gold et al. 2008). However, sometimes clinical sentences could be very complex, containing multiple medications with repetitive signature text, e.g., "*Midrin 2 po initial then 1 po q6hrs prn.*" To better handle such complex cases, Xu et al. (2010) developed a more robust approach that uses a chart parser and a semantic grammar to parse medications in a sentence based on a formal representation model.

2.3.2.2 Machine learning-based methods

From the perspective of supervised machine learning, the medication information extraction task in the 2009 i2b2 challenge can be divided into two steps: (1) identifying medication-related entities (e.g., drug names and other signature fields); and (2) determining the linkage between the detected medication names and the other signature modifiers. Both tasks can be converted into classification problems and resolved, using supervised machine-learning approaches.

Identification of medication-related entities is a typical named entity recognition (NER) problem, which is to determine boundary and semantic classes (e.g., medication, dosage, or frequency) of words/phrases in free

text. To apply machine learning algorithms to an NER task, annotated text are typically converted into a "BIO" format. Specifically, it assigns each token into one of the three classes: **B –** beginning of an entity, **I –** inside an entity, and **O –** outside of an entity. Thus, an NER problem now is converted to a classification problem – to determine a correct label {B, I, O} for each token. Figure 2.4 shows a clinical sentence and its corresponding BIO labels. As multiple types of drug-related entities (e.g., drug name, dose, and frequency) need to be identified, we can extend the BIO labels by adding a suffix to indicate its entity type. For example, "B-m" indicates the beginning of a medication entity, "B-d" indicates the beginning of a dose entity, and "B-f" is used to indicate the beginning of a frequency entity. Different machine learning algorithms have been used for NER tasks. For example, Conditional Random Fields (CRFs) (Lafferty et al. 2001), a representative model for sequence labeling, is one of the most widely used algorithms. In the studies by Patrick and Li (2010) and Li et al. (2010), CRF was used to recognize medication-related entities. Doan and Xu (2010) developed a medication entity recognition approach using Support Vector Machines (SVMs) and reported reasonable performance as well.

As mentioned above, machine learning-based approaches can also be implemented to determine the linkage between the detected medication names and signature modifiers. An intuitive approach to linkage detection would be to build a binary classifier to determine if a candidate pair of medication name and signature modifier is linked or not. Candidate linkage pairs can be generated by taking all possible medication name and signature modifier pairs in one sentence. For example, Patrick and Li (2010) developed an SVM-based classifier to determine the linkage between drug names and signature modifiers in the 2009 i2b2 challenge. Li et al. (2013) proposed a multi-layered sequence labeling for medication-signature linkage detection. However, their evaluation showed that the multi-layered sequence labeling approach did not perform as well as the SVM-based binary classifier.

Token:	In	addition	,	start	Percocet	1-2	tablets	twice	a	day
Label:	O	O	O	O	B-m	B-d	I-d	B-f	I-f	I-f

Fig. 2.4: An example of the BIO representation of an annotated clinical sentence. Upper case letters {B, I, O} stand for beginning of an entity (B), inside an entity (I), and outside of an entity (O) respectively. Lower case letters {m, d, f} stand for entity types, medication name (m), dose (d), and frequency (f).

2.3.2.3 Hybrid methods

A hybrid system takes advantages of both rule-based methods and machine learning-based methods. It often consists of modules that are based on machine learning algorithms and modules that use regular expressions, rules, and dictionaries. Different approaches have been developed to combine machine learning and rule-based methods for medication information extraction. In Patrick and Li's system (Patrick and Li 2010), context-specific rules were applied to outputs of machine learning-based modules in a post-processing fashion. Others used outputs of rule-based modules to improve machine learning classifiers. For example, Doan and Xu (2010) used outputs of a rule-based system as features to feed into a machine learning classifier and demonstrated improved performance. Tikk and Solt (2010) used a rule-based system to create additional training datasets for a machine learning system and also showed better performance.

2.4 Uses of medication information extraction tools in clinical research

Practice-based structured medication data (e.g., claims) have long been used for a large variety of drug outcome studies, including pharmacoepidemiology, pharmacoeconomic, and service-related healthcare investigations (Strom 2005). EHR data, which can include more comprehensive lists of patients' drug exposure (including both over-the-counter and prescription medications) and clinical outcomes, have emerged as a new enabling resource to facilitate broad types of drug-related clinical studies, including pharmacovigilance (Wang et al. 2009) and pharmacogenomics (Wilke et al. 2011). All such studies rely on medication information extraction tools to automatically and accurately extract patient medication exposure information from EHRs.

Pharmacovigilance: Post-market surveillance (also called pharmacovigilance) is an important step to establish complete safety profiles of drugs by detecting additional adverse drug reactions (ADRs) that are not captured during clinical trial phases. Current pharmacovigilance databases such as US Food and Drug Administration's Adverse Event Reporting System (FAERS) have limitations. As a result, EHRs are emerging as a promising new data source for pharmacovigilance (Wood and Martinez 2004; Wysowski and Swartz 2005). Wang et al. (2009) conducted a feasibility study that used the MedLEE system (Friedman et al. 1994) to extract medication and adverse drug events from hospital discharge summaries and then calculated co-occurrence

statistics between these events. Their evaluation showed it was feasible to detect known drug ADRs, as well as novel ADRs from EHRs. The same group then applied similar informatics methods to detect two serious ADRs: rhabdomyolysis and agranulocytosis from EHRs, and showed promising results (Haerian et al. 2012). More recently, La Pendu et al. (2013), a group of researchers at Stanford University, also demonstrated the use of EHRs and NLP methods to conduct pharmacovigilance studies, including detecting ADRs associated with drug-drug interactions.

Pharmacogenomics: Recently, huge efforts have been initiated to link new and existing EHR databases with archived biological material, to accelerate research in personalized medicine, such as pharmacogenomics that aims to identify common and rare genetic variants that contribute to variability in drug response specifically within the context of relevant clinical covariates (McCarty and Wilke 2010). One such effort has been the NIH-funded eMERGE network (*electronic MEdical Records and GEnomics*) (Manolio 2009), a consortium of institutions with DNA biobanks coupled with large comprehensive EHRs. For example, the research team at Vanderbilt has used a DNA biobank linked to de-identified EHR data to successfully replicate pharmacogenomics associations between cardiovascular risk and *CYP2C19**2 and *ABCB1* in patients receiving clopidogrel (Delaney et al. 2012). They also looked at associations between variants in *CYP2C9, VKORC1* and *CYP4F2* and a steady state Warfarin (a blood thinner) dose in individuals of European and African ancestry (Ramirez et al. 2012). In both studies, MedEx (Xu et al. 2010) was used to help identify drug exposure information of patients in EHRs. In addition, the Vanderbilt team also extended MedEx to automatically extract weekly dose of Warfarin (Xu et al. 2011) and daily dose of statins (Wei et al. 2014) from EHRs to facilitate pharmacogenomic studies of both drugs.

2.5 Challenges and future work

Clinical NLP has become an enabling technology to unlock unstructured data in EHRs to support the secondary use of EHR data for clinical and translational research. This chapter briefly introduces relevant work, methods, and applications of clinical NLP technologies, using the task of medication information extraction as an example. Although the content is specific for medication information in EHRs, the NLP methods described here are generalizable to other types of clinical information found in EHRs.

Despite the promising results achieved by current medication

information extraction systems, it is still challenging to accurately extract contextual information for medications, such as duration and reason of medication administration (Uzuner et al. 2010). These types of context are often loosely attached with medication mentions, e.g., the reason may be located in a different sentence than the drug name and using a variety of different words to indicate the linkage, if specified at all. In such situations, sentence syntactic structures could be helpful though parsing clinical text is still an under-explored area of clinical NLP. In addition, existing knowledge bases about drug and indication pairs could potentially be helpful. Existing resources, such as SIDER (Kuhn et al. 2010) and MEDI (Wei et al. 2013), may allow a hybrid approach to improved drug-indication linkage. All in all, integrating domain specific knowledge with machine learning-based information extraction systems remains a challenging task (Friedman et al. 2013).

Another exciting direction of future work, described by Liu and her colleagues (2011) is to build longitudinal drug profiles of patients (e.g., to link longitudinal drug mentions of the same patient to form a timeline of drug exposure events such as the "start" or "discontinuation" of a drug), which is important for any drug-related clinical study.

References

Aronson, A. R. (2001) 'Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program', *Proc AMIA Symp*, 17–21.

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

Chhieng, D., Day, T., Gordon, G. & Hicks, J. (2007) 'Use of natural language programming to extract medication from unstructured electronic medical records', *AMIA Annu Symp Proc*, 908.

Chung, J. & Murphy, S. (2005) 'Concept-value pair extraction from semi-structured clinical narrative: a case study using echocardiogram reports', *AMIA Annu Symp Proc*, 131–135.

Delaney, J. T., Ramirez, A. H., Bowton, E., Pulley, J. M., Basford, M. A., Schildcrout, J. S., Shi, Y., Zink, R., Oetjens, M., Xu, H., Cleator, J. H., Jahangir, E., Ritchie, M. D., Masys, D. R., Roden, D. M., Crawford, D. C. & Denny, J. C. (2012) 'Predicting clopidogrel response using DNA

samples linked to an electronic health record', *Clin Pharmacol Ther,* 91:257–263.

Deleger, L., Grouin, C. & Zweigenbaum, P. (2010) 'Extracting medical information from narrative patient records: the case of medication-related information', *J Am Med Inform Assoc,* 17:555–558.

Denny, J. C., Smithers, J. D., Miller, R. A. & Spickard, A., 3rd (2003) '"Understanding" medical school curriculum content using KnowledgeMap', *J Am Med Inform Assoc*, 10:351–362.

Denny, J. C., Spickard, A., 3rd, Miller, R. A., Schildcrout, J., Darbar, D., Rosenbloom, S. T. & Peterson, J. F. (2005) 'Identifying UMLS concepts from ECG Impressions using KnowledgeMap', *AMIA Annu Symp Proc*, 196–200.

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

Doan, S., Collier, N., Xu, H., Pham, H. D. & Tu, M. P. (2012) 'Recognition of medication information from discharge summaries using ensembles of classifiers', *BMC Med Inform Decis Mak,* 12:36.

Doan, S. X. H. (2010) 'Recognizing Medication related Entities in Hospital Discharge Summaries using Support Vector Machine'. *Coling 2010, The 23rd International Conference on Computational Linguistics* Beijing.

Evans, D. A., Brownlow, N. D., Hersh, W. R. & Campbell, E. M. (1996) 'Automating concept identification in the electronic medical record: an experiment in extracting dosage information', *Proc AMIA Annu Fall Symp*, 388–392.

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

Friedlin, J. & Overhage, M. (2011) 'An evaluation of the UMLS in representing corpus derived clinical concepts,' *AMIA Annu Symp Proc,* 2011:435–444.

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

Friedman, C., Rindflesch, T. C. & Corn, M. (2013) 'Natural language processing: state of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine', *J Biomed Inform,* 46:765–773.

Gold, S., Elhadad, N., Zhu, X., Cimino, J. J. & Hripcsak, G. (2008) 'Extracting structured medication event information from discharge summaries', *AMIA Annu Symp Proc*, 237–241.

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

Hahn, U., Romacker, M. & Schulz, S. (2002) 'MEDSYNDIKATE–a natural language system for the extraction of medical information from findings reports', *Int J Med Inform,* 67:63–74.

Hamon, T. & Grabar, N. (2010) 'Linguistic approach for identification of medication names and related information in clinical narratives', *J Am Med Inform Assoc,* 17:549–554.

Harkema, H., Dowling, J. N., Thornblade, T. & Chapman, W. W. (2009) 'ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports', *J Biomed Inform,* 42:839–851.

Haug, P. J., Koehler, S., Lau, L. M., Wang, P., Rocha, R. & Huff, S. M. (1995) 'Experience with a mixed semantic/syntactic parser', *Proc Annu Symp Comput Appl Med Care*, 284–288.

Hripcsak, G., Austin, J. H., Alderson, P. O. & Friedman, C. (2002) 'Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports', *Radiology,* 224:157–163.

Jagannathan, V., Mullett, C. J., Arbogast, J. G., Halbritter, K. A., Yellapragada, D., Regulapati, S. & Bandaru, P. (2009) 'Assessment of commercial NLP engines for medication information extraction from dictated clinical notes', *Int J Med Inform,* 78:284–291.

Jha, A. K., Desroches, C. M., Campbell, E. G., Donelan, K., Rao, S. R., Ferris, T. G., Shields, A., Rosenbaum, S. & Blumenthal, D. (2009) 'Use of electronic health records in U.S. hospitals', *N Engl J Med*, 360:1628–1638.

Kohane, I. S. (2011) 'Using electronic health records to drive discovery in disease genomics', *Nat Rev Genet,* 12:417–428.

Kuhn, M., Campillos, M., Letunic, I., Jensen, L. J. & Bork, P. (2010) 'A side effect resource to capture phenotypic effects of drugs', *Mol Syst Biol,* 6:343.

Lafferty, J., Mccallum, A. & Pereira, F. (2001) 'Conditional random fields: Probabilistic models for segmenting and labeling sequence data', *Proc. 18th International Conf. on Machine Learning*, 282–289.

Lependu, P., Iyer, S. V., Bauer-Mehren, A., Harpaz, R., Mortensen, J. M., Podchiyska, T., Ferris, T. A. & Shah, N. H. (2013) 'Pharmacovigilance using clinical notes', *Clin Pharmacol Ther,* 93:547–555.

Levin, M. A., Krol, M., Doshi, A. M. & Reich, D. L. (2007) 'Extraction and mapping of drug names from free text to a standardized nomenclature', *AMIA Annu Symp Proc*, 438–442.

Li, Q., Zhai, H., Deleger, L., Lingren, T., Kaiser, M., Stoutenborough, L. & Solti, I. (2013) 'A sequence labeling approach to link medications and their attributes in clinical notes and clinical trial announcements for information extraction', *J Am Med Inform Assoc,* 20:915–921.

Li, Z., Liu, F., Antieau, L., Cao, Y. & Yu, H. (2010) 'Lancet: a high precision medication event extraction system for clinical text', *J Am Med Inform Assoc,* 17:563–567.

Liu, M., Jiang, M., Kawai, V. K., Stein, C. M., Roden, D. M., Denny, J. C. & Xu, H. (2011) 'Modeling drug exposure data in electronic medical records: an application to warfarin', *AMIA Annu Symp Proc,* 2011:815–823.

Manolio, T. A. (2009) 'Collaborative genome-wide association studies of diverse diseases: programs of the NHGRIs office of population genomics', *Pharmacogenomics,* 10:235–241.

Mccarty, C. A. & Wilke, R. A. (2010) 'Biobanking and pharmacogenomics', *Pharmacogenomics,* 11:637–641.

Meystre, S**.** & Haug, P. J. (2006) 'Natural language processing to extract medical problems from electronic clinical documents: performance evaluation', *J Biomed Inform,* 39:589–599.

Meystre, S**.** M., Savova, G. K., Kipper-Schuler, K. C**.** & Hurdle, J. F. (2008) 'Extracting information from textual documents in the electronic health record: a review of recent research', *Yearb Med Inform*, 128–144.

Meystre, S. M., Thibault, J., Shen, S., Hurdle, J. F. & South, B. R. (2010) 'Textractor: a hybrid system for medications and reason for their prescription extraction from clinical text documents', *J Am Med Inform Assoc,* 17:559–562.

Mork, J. G., Bodenreider, O., Demner-Fushman, D., Dogan, R. I., Lang, F. M., Lu, Z., Neveol, A., Peters, L., Shooshan, S**.** E. & Aronson, A. R. (2010) 'Extracting Rx information from clinical narrative', *J Am Med Inform Assoc,* 17:536–539.

Nadkarni, P. M., Ohno-Machado, L. & Chapman, W. W. (2011) 'Natural language processing: an introduction', *J Am Med Inform Assoc,* 18:544–551.

Patrick, J. & Li, M. (2010) 'High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge', *J Am Med Inform Assoc,* 17:524–527.

Ramirez, A. H., Shi, Y., Schildcrout, J., Delaney, J. T., Xu, H., Oetjens, M., Zuvich, R., Basford, M., Bowton, E., Jiang, M., Zink, R., Cowan, J. D., Pulley, J. M., Ritchie, M. D., Peterson, J. F., Masys, D. R., Roden, D. M., Crawford, D. C**.** & Denny, J. C. (2012) 'Predicting warfarin dosage in European and African Americans using DNA samples linked to an electronic health record'. *Pharmacogenomics,* in press.

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

Sager, N., Lyman, M., Bucknall, C., Nhan, N. & Tick, L. J. (1994) Natural language processing and the representation of clinical data. *J Am*

Med Inform Assoc, 1:142–160.

Savova, G. K., Fan, J., Ye, Z., Murphy, S. P., Zheng, J., Chute, C. G. & Kullo, I. J. (2010a) 'Discovering peripheral arterial disease cases from radiology notes using natural language processing', *AMIA Annu Symp Proc,* 2010:722–726.

Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C. & Chute, C. G. (2010b) 'Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications', *J Am Med Inform Assoc,* 17:507–513.

Savova, G. K., Ogren, P. V., Duffy, P. H., Buntrock, J. D. & Chute, C. G. (2008) 'Mayo clinic NLP system for patient smoking status identification', *J Am Med Inform Assoc,* 15:25–28.

Schadow, G. & Mcdonald, C. J. (2003) 'Extracting structured information from free text pathology reports', *AMIA Annu Symp Proc*, 584–588.

Shea, S. & Hripcsak, G. (2010) 'Accelerating the use of electronic health records in physician practices', *N Engl J Med,* 362:192–195.

Spasic, I., Sarafraz, F., Keane, J. A. & Nenadic, G. (2010) 'Medication information extraction with linguistic pattern matching and semantic rules,' *J Am Med Inform Assoc,* 17:532–535.

Strom, B. L. (2005) *Pharmacoepidemiology,* J. Wiley: Chichester; Hoboken, NJ.

Tikk, D. & Solt, I. (2010) 'Improving textual medication extraction using combined conditional random fields and rule-based systems', *J Am Med Inform Assoc,* 17:540–544.

Uzuner, O., Solti, I. & Cadag, E. (2010) 'Extracting medication information from clinical text', *J Am Med Inform Assoc,* 17:514–518.

Wang, X., Hripcsak, G., Markatou, M. & Friedman, C. (2009) 'Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study', *J Am Med Inform Assoc,* 16:328–337.

Wei, W. Q., Cronin, R. M., Xu, H., Lasko, T. A., Bastarache, L. & Denny, J. C. (2013) 'Development and evaluation of an ensemble resource

linking medications to their indications', *J Am Med Inform Assoc,* 20:954–961.

Wei, W. Q., Feng, Q., Jiang, L., Waitara, M. S., Iwuchukwu, O. F., Roden, D. M., Jiang, M., Xu, H., Krauss, R. M., Rotter, J. I., Nickerson, D. A., Davis, R. L., Berg, R. L., Peissig, P. L., Mccarty, C. A., Wilke, R. A. & Denny, J. C. (2014) 'Characterization of statin dose response in electronic medical records', *Clin Pharmacol Ther,* 95(3):331–338.

Wilke, R. A., Xu, H., Denny, J. C., Roden, D. M., Krauss, R. M., Mccarty, C. A., Davis, R. L., Skaar, T., Lamba, J. & Savova, G. (2011) 'The emerging role of electronic medical records in pharmacogenomics', *Clin Pharmacol Ther,* 89:379–386.

Wood, L. & Martinez, C. (2004) 'The general practice research database: role in pharmacovigilance', *Drug Saf,* 27:871–881.

Wysowski, D. K. & Swartz, L. (2005) 'Adverse drug event surveillance and drug withdrawals in the United States, 1969–2002: the importance of reporting suspected reactions', *Arch Intern Med,* 165:1363–1369.

Xu, H., Jiang, M., Oetjens, M., Bowton, E. A., Ramirez, A. H., Jeff, J. M., Basford, M. A., Pulley, J. M., Cowan, J. D., Wang, X., Ritchie, M. D., Masys, D. R., Roden, D. M., Crawford, D. C. & Denny, J. C. (2011) 'Facilitating pharmacogenetic studies using electronic health records and natural-language processing: a case study of warfarin', *J Am Med Inform Assoc,* 18:387–391.

Xu, H., Stenner, S. P., Doan, S., Johnson, K. B., Waitman, L. R. & Denny, J. C. (2010) 'MedEx: a medication information extraction system for clinical narratives', *J Am Med Inform Assoc,* 17:19–24.

Yang, H. (2010) 'Automatic extraction of medication information from medical discharge summaries', *J Am Med Inform Assoc,* 17:545–548.

Yetisgen-Yildiz, M., Gunn, M. L., Xia, F. & Payne, T. H. (2013) 'A text processing pipeline to extract recommendations from radiology reports', *J Biomed Inform,* 46:354–362.

Zeng, Q. T., Goryachev, S., Weiss, S., Sordo, M., Murphy, S. N. & Lazarus, R. (2006) 'Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system', *BMC Med Inform Decis Mak,* 6:30.

António Teixeira, Liliana Ferreira and Mário Rodrigues

3 Online health information semantic search and exploration: reporting on two prototypes for performing information extraction on both a hospital intranet and the world wide web

Abstract: In this chapter, we apply ontology-based information extraction to unstructured natural language sources to help enable semantic search of health information. We propose a general architecture capable of handling both private and public data. Two of our novel systems that are based on this architecture are presented here. The first system, MedInX, is a Medical Information eXtraction system which processes textual clinical discharge records, performing automatic and accurate mapping of free text reports onto a structured representation. MedInX is designed to be used by health professionals, and by hospital administrators and managers, allowing its users to search the contents of such automatically populated ontologies. The second system, SPHInX, attempts to perform semantic search on health information publicly available on the web in Portuguese. The potential of the proposed approach is clearly shown with usage examples and evaluation results.

3.1 Introduction

More and more healthcare institutions store vast amounts of information about users, procedures, and examinations, as well as the findings, test results, and diagnoses, respectively. Other institutions, such as the Government, increasingly disclose health information on varied topics of concern to the public writ large. Health research is one of the most active areas, resulting in a steady flow of publications reporting on new findings and results.

In recent years, the Internet has become one of the most important tools to obtain medical and health information. Standard web search is by far the most common interface for such information (Abraham & Reddy 2007). Several general search engines such as Google, Yahoo! and Bing currently play an important role in obtaining medical information for both

medical professionals and lay persons (Wang et al. 2012). However, these general search engines do not allow the end-user to obtain a clear and organized presentation of the available health information. Instead, it is more or less of a hit or miss, random return of information on any given search. In fact, medicine-related information search is different from other information searches, since users often use medical terminology, disease knowledge, and treatment options in their search (Wang et al. 2012).

Much of the information that would be of interest to private citizens, researchers, and health professionals is found in unstructured text documents. Efficient access to this information implies the development of search systems capable of handling the technical lexicon of the domain area, entities such as drugs and exams, and the domain structure. Such search systems are said to perform semantic search as they base the search on the *concepts* asked and not so much on the words used in the query (Guha, McCool & Miller 2003). Semantic search maintains several advantages over search based on surface methods, such as those that directly index text words themselves rather than underlying concepts. Three main advantages of concept-based search are: (1) they usually produce smaller sets of results, as they are able to identify and remove semantically duplicated results and/or semantically irrelevant results; (2) they can integrate related information scattered across documents; frequently answers are obtained by compounding information from two or more sources; and (3) they can retrieve relevant results even when the question and answer do not have common words, since these systems can be aware of similar concepts, synonyms, meronyms, antonyms, etc.

Semantic search involves representing the concepts of a domain and the relations between them, organizing, in this way, the information according to its semantics and forming a knowledge representation of the world or some part of it. This representation is called *ontology*, which is formally defined as an explicit specification of a shared conceptualization (Gruber 1993). Ontology describes a hierarchy of concepts related by subsumption relationships, and can include axioms to express other relationships between concepts and to constrain their intended interpretation. The usage of ontology to explicitly define the application domain brings large benefits from the viewpoint of information accessibility, maintainability and interoperability, as it formalizes and allows the application’s view of the world to be made public (Guarino 1998). Also, with the emergence of semantic reasoners, software that is able to infer logical consequences from a set of asserted facts or axioms, it is possible to verify the coherence of the stored information and to infer new information from the contents of ontology (Sirin & Parsia 2004).

However, there is still the challenge of bridging the gap between the needed semantically structured information and the original text content. The acquisition of specific and relevant pieces of information from texts, and respective storage in a coherent framework, is called *information extraction* (Cowie & Lehnert 1996). The general problem of information extraction (IE) involves the analysis of natural language texts, such as English or Portuguese texts, to determine the semantic relations among the existing entities and the events they participate in, namely their relations. Natural language texts can be unstructured, plain texts, and/or semi-structured machine-readable documents, with some kind of markup. The information to be retrieved can be entities, classes of objects and events, and relationships between them. Informally, IE is the task of detecting elements such as “who” did “what” to “whom,” “when” and “where,” in unstructured free text information sources, and using those elements to populate structured information sources (Gaizauskas & Wilks 1998; Màrquez et al. 2008).

IE is different from information retrieval (IR), which is the task usually performed by current search engines such as Google and Bing. Whereas IE aims to extract relevant information from documents, IR aims to retrieve those relevant documents themselves from collections. For example, universities and other public libraries use IR systems to provide access to books, journals and other documents. However, in such cases, after querying search engines the users *still* have to read through those documents brought up in their search to find the information they were looking for. When the goal is to explore data, obtain a summary of facts reported in large amounts of documents or have facts presented in tables, IE becomes a much more relevant technology than IR (McNaught & Black 2006).

A typical IE system has two main subtasks: entity recognition and relation extraction. Entity recognition seeks to locate and classify atomic elements in natural language texts into predefined categories, while relation extraction tries to identify the relations between the entities in order to fill predefined templates. Two important challenges exist in IE. One arises from the variety of ways of expressing the same fact. The other challenge, shared by almost all NLP tasks, is due to the great expressiveness of natural languages, which can have ambiguous structure and meaning.

The chapter is structured as follows: the next two sections provide background information and an overview of related work about information search in general and in the health domain, information extraction in health, and ontology-based information extraction. Section 4

contains the general vision/ proposal of an ontology-based information extraction system to feed a search engine. The respective instantiation in two systems with different purposes and some illustrative results are presented in Section 5. Chapter ends with conclusions, provided in Section 6.

3.2 Background

Several approaches to IE have been followed over the years. One common approach is based on pattern matching and exploits basic patterns over a variety of structures: text strings, part-of-speech tags, semantic pairs, and dictionary entries (Pakhomov 2005). However, this type of approach does not generalize well, which limits its extension to new domains. The need for IE systems that can be easily adapted from one domain to another leads to the development of different approaches based on adaptive IE, starting with the Alembic Workbench (Aberdeen et al. 1995). The idea behind these approaches is to use various kinds of machine learning algorithms to allow IE systems to be easily targeted to new problems. The effort required to redesign a new system is replaced with that of generating batches of training data and applying learning algorithms.

A more recent approach is the ontology-based IE (OBIE), which aims at using ontology to guide the information extraction process (Hahn, Romacker & Schulz 2002). Since Berners-Lee et al. (1994) and Berners-Lee & Fischetti (1999) began to endorse ontologies as the backbone of the semantic web in the 1990s, a whole research field has evolved around the fundamental engineering aspects of ontologies, such as their generation, evaluation and management.

A relevant number of approaches need seed examples to train the IE systems. As such, several tools to annotate the semantic web were developed. Some earlier systems involved having humans annotate texts manually, using user-friendly interfaces (Handschuh, Staab & Studer 2003; Schroeter, Hunter & Kosovic 2003). Others featured algorithms to automate part of the annotation process. Those algorithms were based on manually constructed rules or extraction patterns, to be completed based on the previous annotations (Alfonseca & Manandhar 2002; Ciravegna et al. 2002).

As manual annotation can be a time consuming task, some approaches involved using ontology class and subclass names to generate seed examples for the learning process. Those names are used to learn contexts from the web and then those contexts are used to extract information

(Kiryakov et al. 2004; Buitelaar et al. 2008). Other approaches added Hearst patterns to increase the amount of seed examples (McDowell & Cafarella 2008). For instance, considering the Bird class, useful patterns would be “birds such as X,” “birds including X,” “X and other birds,” and “X or other birds,” among others.

3.3 Related work

Relevant related work on search, particularly search applied to health information is the focus of this section. Here, relevant work related to ontology-based health information search is presented. First, recent trends in semantic search are presented; thereafter we discuss some of the trends related to health search and its specificities followed by a discussion of information extraction applied to health. The section ends with recent relevant work on OBIE.

3.3.1 Semantic search

In recent years, the interest in semantic search has increased. Even mainstream search engines such as Google or Bing are evolving to include semantics. Some systems do not assume that all, or most data, have a formal semantic annotation. One approach is expanding the user query by including synonyms and meronyms of the queried terms (Moldovan & Mihalcea 2000; Buscaldi, Rosso & Arnal 2005). Term expansion is made using the “OR” operation available in most search engines. A somewhat similar approach is followed by Kruse et al. (2005), which uses WordNet ontology and the “AND” operation of search engines to provide semantic clarification on concepts that have more than one meaning in WordNet.

Other approaches combine full text search and ontology search. ESTER (Bast et al. 2007) features an entity recognizer that assigns words or phrases to the entities of the ontology. Then, when searching for information, two basic operations are used: prefix search and join. This allows discrimination of different meanings of a concept but without logic inference. A different approach is adopted by Rocha, Schwabe & Aragao (2004). This approach involves using a regular full text search plus locating additional relevant information by using other document data such as document creator. This additional data is stored in a RDF graph, which is traversed in order to find similar concepts.

Systems that process only data with a formal semantic annotation use SPARQL queries to retrieve results (Guha & McCool 2003; Lei, Uren &

Motta 2006; Esa, Taib & Thi 2010). The problems usually addressed in these cases are performance, in terms of reasoning speed, and how to rank the result set. A discussion on semantically enhanced search engines for web content discovery can be found in Kamath et al. (2013). Jindal, Bawa & Batra (2013) present a detailed review of ranking approaches for semantic search on web.

3.3.2 Health information search and exploration

In the health domain, web-available search engines are mainly targeted at retrieving information from related knowledge resources such as PubMed, the Medical Subject Headings thesaurus (MeSH) of the U.S. National Library of Medicine and the Unified Medical Language System (UMLS) of the U.S. National Library of Medicine.

CISMeF and HONselect are examples of such systems. The objective of CISMeF (Darmoni et al. 2000) is to assist health professionals and consumers in their search for electronic health information available on the Internet. CISMeF, initially only available in French, has recently improved in two ways, being currently: (1) a generic tool able to describe and index web resources and PubMed citations or Electronic Health Records; (2) multi-lingual by allowing queries in multiple terminologies and several languages. HONselect (Boyer et al. 2001) presents medical information arranged under MeSH, offering advanced multilingual features to facilitate comprehension of web pages in languages other than those of the user.

Another health-specific information search engine is WRAPIN (Gaudinat et al. 2006). WRAPIN combines search in medical Web pages with other “hidden” online documents that are not referenced by other search engines. WRAPIN analyses a page for the most important medical terms, performing frequency analysis on MeSH terms found on the page. It identifies keywords which are then used for weighted queries to its indexes and to translate into languages other than that of the initial query. WRAPIN also allows the most important medical concepts in the document to be highlighted.

Can & Baykal (2007) designed MedicoPort, a medical search engine designed for users with no medical expertise. It is enhanced with domain knowledge obtained from UMLS in order to increase search effectiveness. MedicoPort is semantically enhanced by transforming a keyword search into a conceptual search, both for web pages and user queries.

As an example of recent work, Mendonça et al. (2012) designed and developed a proof-of-concept system for a specific group of target users

and a specific domain, namely, *Neurological Diseases*. The application allows users to search for neurologic diseases, and collects a set of relevant documents with the support of ontology navigation as an auxiliary tool to redefine a query and change previous results.

Another example of recent work is that of Dragusin et al. (2013) who introduced FindZebra, a specialized rare disease search engine powered by open-source search technology. FindZebra uses freely available online medical information, but also includes specialized functionalities such as exploiting medical ontological information and UMLS medical concepts to demonstrate different ways of displaying results to medical experts. The authors concluded that specialized search engines can improve diagnostic quality without compromising the ease of use of the current and widely popular web search engines.

3.3.3 Information extraction for health

In the clinical domain, IE was initially approached with complete systems, i.e., systems including all functions required to fully analyze free-text. Examples of these large-scale projects are:

- The Linguistic String Project – Medical Language Processor of New York University
- The Specialist system (McGray et al. 1987) developed at the United States National Library of Medicine as part of UMLS project. This system includes the Specialist Lexicon, the Semantic Network, and the UMLS Metathesaurus (USNLM 2008)
- The Medical Language Extraction and Encoding system (MedLEE) system (Friedman et al. 1995) developed at the New York Presbyterian Hospital at Columbia University. MedLEE is mainly semantically driven; it is used to extract information from clinical narrative reports, to participate in an automated decision-support system, and to allow natural language queries.

Significant resources were required to develop and implement these complete medical language processing systems. Consequently, several authors experimented over time with simpler systems that were focused on specific IE tasks and on limited numbers of different types of information to extract. Some of the areas currently benefiting from IE methods are biomedical and clinical research, clinical text mining, automatic terminology management, decision support and bio-surveillance. These narrowly focused systems demonstrated such good performance that they now constitute the majority of systems used for IE.

Relevant examples of this type of system are those from the International Classification of Diseases (ICD) (Aronson et al. 2007; Crammer et al. 2007).

3.3.4 Ontology-based information extraction – OBIE

Different approaches to OBIE have been proposed and developed over the years. The approaches differ in some dimensions as follows: (1) the identification and extraction of information can be performed using probabilistic methods or explicitly defined sets of rules; (2) the types of document from which information is extracted can be unstructured, plain text, or semi-structured and structured sources; (3) the ontology can be constructed from the document's content or exist before the process started, and has the option of being updated automatically while processing documents; and (4) the kind of information extracted varies from extracting only ontological instances to extracting entire ontological classes and its associated properties.

Several IE groups focused on the development of extraction methods that use the content and predefined semantics of an ontology to perform the extraction task without human intervention and dependency on other knowledge resources (Embley et al. 1998; Maedche et al. 2002; Buitelaar & Siegel 2006; Yildiz 2007).

Considering IE for generic domains, a frequent approach is to use Wikipedia to build their knowledge base. Relevant examples of such systems were developed by Bizer et al. (2009), Suchanek, Ifrim & Weikum (2007), and Wu, Hoffmann & Weld (2008). Wikipedia structure is used to infer the semantics, and the knowledge base is populated by extracting information from pages of text and infoboxes. Other approaches, that do not take advantage of Wikipedia structure, acquire information from generic web pages. The knowledge base structure is often inferred from pages of content and the knowledge base is populated using the same sources. Two good examples are Etzioni et al. (2004) and Yates et al. (2007).

Most approaches, whether using Wikipedia or not, use shallow linguistic analysis to detect the information to extract. Shallow analysis involves detecting text patterns and, at most, using part-of-speech information: e.g., which words are nouns, verbs, adverbs, or adjectives. The use of shallow linguistic information, however, makes it difficult to acquire information from complex sentences. A comprehensive survey of current approaches to OBIE can be found in Wimalasuriya & Dou (2010).

3.4 A general architecture for health search: handling both private and public content

Health-related information can have quite different access restrictions. Personal health-related information is confidential and has restricted access even inside health organizations. On the other hand, other health-related information, such as general information on drugs and disease characterization, is available to the general public. Not every case falls neatly into either personal health information that is confidential or more broadly into the category of general health info accessible to the public writ large. It is thus important that both well-defined cases which are either personal health-related or general-health related, and those that are not as extreme but share features of both categories, be addressed in a similar way. In this section we present a unified view of these cases (Fig. 3.1).

The main differences between the two extreme cases of personal health info versus general health pivots on (1) who are the target users of the information, i.e., if the search is made available to a general audience or only to authorized users (power users in the figure); and (2) if such a search is only possible inside the intranet of the organization holding the original source of information. In our view, these differences do not need different architectures to be handled effectively. In both cases, processing of the public and private documents can be performed by a similar pipeline, combining IE and semantic integration, and making use of ontologies. The Search and Inference Engine can also be the same. The only requirement is that the interfaces can handle access restrictions, preventing both unauthorized user access and access from the web when the information is for an organization's restricted internal use.

The proposed architecture is composed of a set of modules in which the IE component consists of a basic set of processing elements similar to the basic set illustrated in Fig. 3.2 and described by Hobbs (2002). The semantic integration module is responsible for merging the information extracted in the previous modules and reasoning. The following sections describe in detail the architecture of the two different systems used in this work.

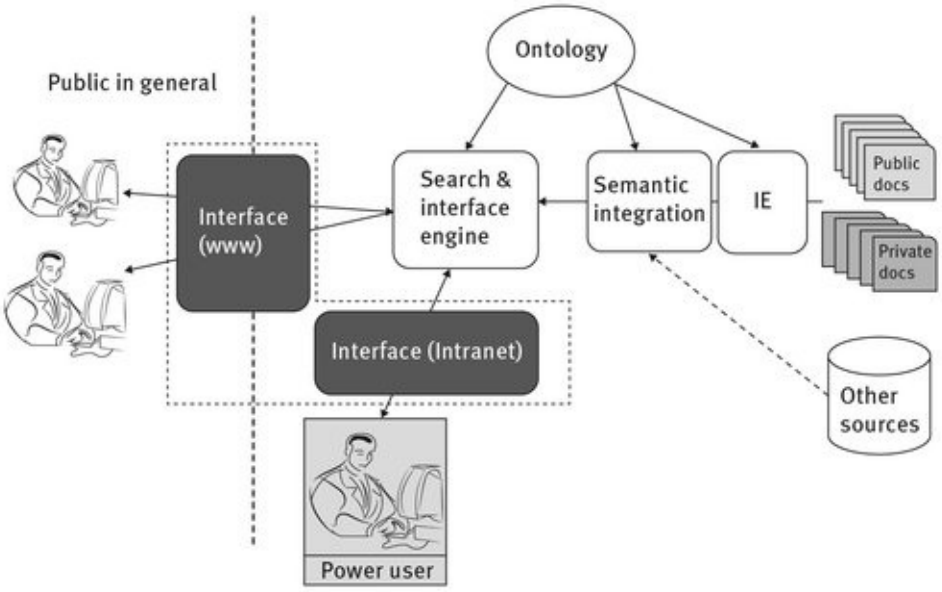


Fig. 3.1: Unified view of semantic search for health, handling both access by the general public and restricted access by authorized users inside an organization.

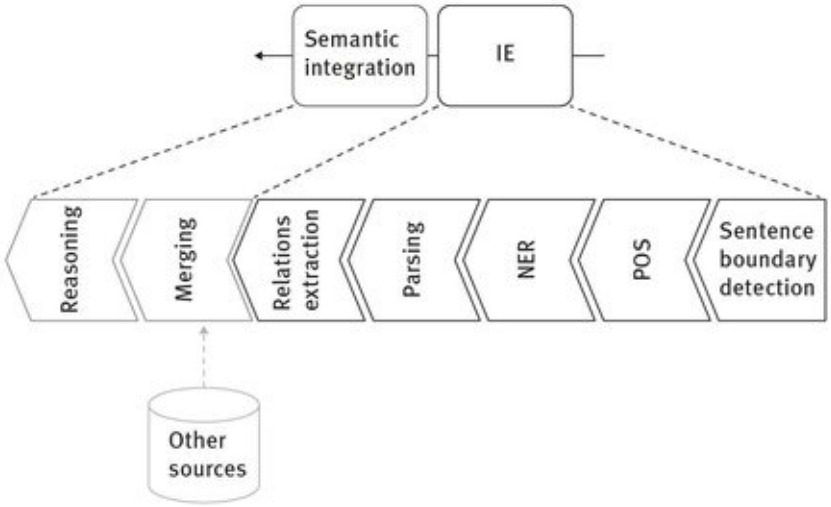


Fig. 3.2: The basic set of modules to be included in the processing pipeline to extract information from natural language sources and feeding the search engine.

3.5 Two semantic search systems for health

In this section we present in detail two different systems providing semantic search supported by information extraction for both private and public content. The first system targets the search inside a health institution such as a hospital, or more generally what is considered an *Intranet* search, which is a search within an organization's own internal website or group of websites. The second system in contrast targets the search outside the confines of the organization so as to enable the general public to semantically search and explore health-related information

made available on the World Wide Web in Portuguese. Both of these semantic search systems were designed for Portuguese, but can be readily extended to other languages.

3.5.1 MedInX

MedInX (Ferreira, Teixeira & Cunha 2012) is a medical information eXtraction system tailored to process textual clinical discharge records in order to perform automatic and accurate mapping of free text reports onto a structured representation. MedInX is designed to be used by health professionals, and by hospital administrators and managers, as it also allows its users to search the contents of such automatically populated ontologies. MedInX uses IE technology to structure the information present in discharge reports originated by the electronic health record (EHR) system used in the region of Aveiro, Portugal in the Telematic Healthcare Network RTS® (Cunha et al. 2006). The way it works is by automatically instantiating a knowledge representation model from the free-text patient discharge letters (PDL) issued by the hospital.

During a patient's hospitalization, a large amount of data is produced in textual form as in the case of patient discharge letters. The purpose of these documents is to transfer summarized information from the hospital setting to other places, normally to the general practitioner, in order to assure continuity of patient care. MedInX addresses this type of narrative since they cover the whole inpatient period and summarize the major occurrences during that period.

The first step in development of MedInX was the creation of a corpus of authentic health records to be used in the development and evaluation of the system. This corpus was gathered through a list of hospital episodes for which a code had been assigned relative to the diagnosis of a cerebrovascular disease. If more than one cerebrovascular disease were found in the patient, additional codes pertaining to those conditions were entered in the patient's record. The corpus consists, thus, of 915 discharge letters written in Portuguese, corresponding to patients admitted with at least one of the several cerebrovascular disease-related codes. Table 3.1 gives statistics pertaining to the amount of documents, sentences, and tokens included in the development and test-evaluation set of MedInX.

Tab. 3.1: Number of documents, tokens and sentences in the MedInX corpus and its subsets.

Number of	Development set	Test set	Total
Documents	829	86	915
Tokens	215,730	21,788	237,518
Tokens/Document	260	253	260
Sentences	12,974	1,346	14,320
Sentences/Document	16	16	16

Figure 3.3 presents a style-preserving illustration of a PDL, showing how it is possible to analyze the general content and structure of the documents. To begin with, the discharge documents have several interesting contextual features. In general, it is evident that the narratives are written from one professional to another in order to support information transfer, remind them about important medical facts, and supplement with crucial numerical data such as blood pressure and lab test results. The texts are normally intelligible and the meaning becomes evident from the context even in the presence of numerous linguistic and grammatical mistakes, word abbreviations, acronyms, signs, and other communicative features.

MedInX is a system designed for the clinical domain, which contains components for the extraction of hypertension-specific characteristics from unstructured PDLs. Its components are based on NLP principles; as such, they contain several mechanisms to read, process, and utilize external resources, such as terminologies and ontologies. These external resources represent an important part of the system by providing structured representations of the domain, clinical facts, and events that are present in the texts. The MedInX ontologies allow the assignment of domain-specific meanings to terms and use these meanings in their operations.

3.5.1.1 MedInX ontologies

In MedInX four new ontologies were created. The first two consist of two formalizations of the international classification systems that are supported by the World Health Organization (WHO): the International Classification of Diseases (ICD); and the International classification of functioning, disability and health (ICF). A drugs ontology and a conceptualization of the structure and content of the discharge reports comprise the last two of the MedInX ontologies.

Motivo Internamento AVC isquémico de repetição; HTA	**Admission Reason** Recurrent ischemic stroke, HTA
História Clinica Doente de 83 anos,com antec. de HTA e depressão nervosa.Faz uso regular de fármacos que desconhe nomes.Hoje veio transferida do Hospital do Visconde de Salreu por desvio da comissura labial para esquerda. Medicada regular/ com Ogasto; Tenoretic; Micardis e Motillium. A doente refere que cerca das 09H00 teve episodio de disartra acompanhado de desvio da comissura labial a dta, sem alterações da FM ou da sensibilidade. Recorreu ao Hosp de Estarreja tendo sido encaminhada a esta Urgencia por hipotese de AVC. Desde a entrada no HVS que refere melhoria progressiva das alterações da fala, sem deficites neurologicos de novo. Sem queixas sugestivas de sindrome infeccioso ou de febre.	**Clinical History** 83 years old patient, with history of HTA and clinical depression. Uses drugs regularly of which does not know names. Was transferred today from the Hospital Visconde Salreu by deviation of the left lip. Regularly medicated with Ogasto; Tenoretic; Motillium and Micardis. The patient states that at approximately 09:00 had an episode of dysarthria accompanied by deviation of the right lip without changes in MS or sensation. Appealed to the Hospital of Estarreja and was forwarded to this Urgency due to stroke suspicion. Refers progressive improvement of speech changes since entering the HVS, no neurological deficit again. No complaints suggestive of infectious syndrome or fever.
Exame Fisico COC, eupneica sem SDR. Apiretica. Chorosa. Corada e hidratada. TA- 193/68 mmHg; spO2 (AA)- 99%; PR- 60/min. AC- irregular, por ES. AP- Mv+ sem RA valorizaveis. Sem edemas dos MI's. ENS: EG- 15; sem lateralização motora; FM preservada; sem alteração da linguagem; pupilas I/R; olhos na linha media; esboço de paresia facial central a Dta.	**Physical Examination** COC, eupneic without RDS. Afebrile. Tearful. Healthy coloring and hydrated. BP-193 / 68 mmHg; spO2 (AA) - 99%, PR-60/min. CA-irregular, due to ES. PA-Bs + without considerable rales. No edema in the IMs. ENS: EG-15, no motor lateralization; MS preserved, without changes in language, PERRLA; eyes in the midline; outline of central facial palsy at the right.
Terapêutica Efectuada Terapeutica efectuada:-aas 100, enoxaparina, esomeprazol, insulina sos, nitratos transdermicos, soros.	**Therapeutics** Therapeutic: -asa 100, enoxaparin, esomeprazole, sos insulin, transdermal nitrates, salines.
Destino Hos. de Salreu	**Destination** Salreu Hos.
Evolução Tranferida para o Hosp. de Salreu	**Evolution** Transferred to Salreu Hosp.

Fig. 3.3: A style-preserving illustration of a patient discharge letter. The original document written in portuguese is presented on the left of the figure and, on the right, its English translation.

This last ontology, in particular, was designed as an extensible knowledge model and used for storing and structuring the PDLs' entities and their relations, including temporal and modifying information. For instance, the MedInX ontology describes the fact that a Sign or Symptom is a Condition, a prescribed Medication is a Therapeutic, a Procedure can be targeted to an Anatomical Site, and so forth. To identify the concepts in the ontology a middle-out strategy was used, i.e., first the core basic terms were identified in text and then specified and generalized as required.

3.5.1.2 MedInX system

MedInX components run within the unstructured information

management architecture (UIMA) framework (Ferrucci & Lally 2004). Figure 3.4 illustrates MedInX architecture, identifying the following components:

1. **Document Reader:** a component which converts PDL files into plain text and extracts implicit meaning from the structure of the document by converting the embedded tags of the input document into annotations;
2. **General natural language processing:** components for sentence discovery, tokenization and part-of-speech tagging;
3. **REMMIX: the Named Entity Recognition component of MedInX** which concentrates on the later stages of IE, i.e., takes the linguistic objects as input and finds domain-dependent classifications and patterns among them. REMMIX is made up of three other components:
    a. **Context Dependent annotator:** an annotator that creates annotations from one or more tokens, using regular expressions and surrounding tokens as clues;
    b. **Concept finding:** a component which extracts concepts based on specified terminologies and ontologies, and determines negation, lateralization and modifiers;
    c. **Relation extraction:** a component which extracts relations between concepts using contextual information;
4. **Consumers:** responsible for ending the process and creating the desired output. The three main consumers of MedInX are:
    a. **XML consumer:** which produces an XML file with the annotations of the previous annotators.
    b. **Ontology population:** which populates the MedInX ontology. This component produces an OWL file with the information extracted.
    c. **Template filling**: which outputs the knowledge extracted from the narratives to a template containing information about the patient and their health related state, with the correspondent ICF codes.

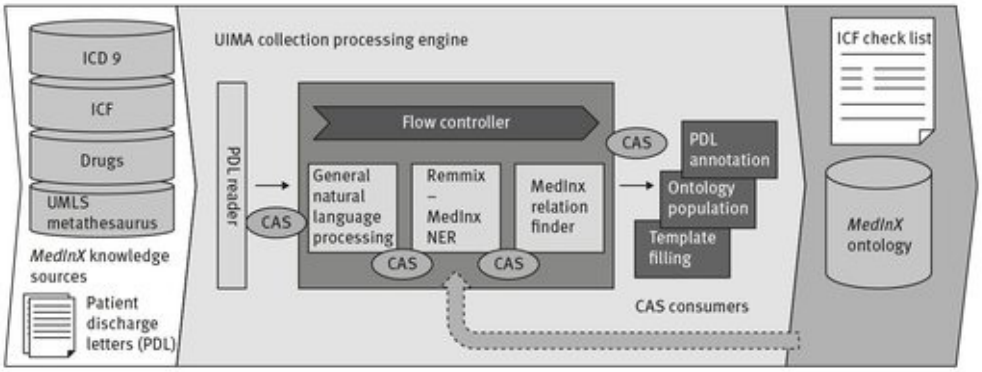


Fig. 3.4: MedInX architecture.

3.5.1.3 Representative results

Figure 3.5 presents MedInX evaluation interface. The extracted entities are identified by the filled boxes while the arrows represent the relations between these.

MedInX was first evaluated in 2011, in the task of extracting information from PDLs. Seven judges, belonging to different specialized areas ranging from medicine to computer science to linguistics, participated in the assessment by using the web-based evaluation interface (Fig. 3.5). To wit, the jury was made up of two computer scientists, a linguist, a radiologist, two psychologists and a physician. The 86 PDLs of the evaluation set initially selected from the MedInX corpus were automatically annotated by MedInX and made available for evaluation for four months. During this four-month period a total of 30 different reports were reviewed.

Fig. 3.5: MedInX evaluation interface.

The results obtained in the task of semantic classification are presented

in Tab. 3.2 in terms of precision, recall, and F-measure. These values indicate a good performance of MedInx all the way around. For example, given that MedInX is tailored to the medical domain and intended to process clinical text, its proficiency with *correctly* extracting the entities and relations described in text makes it very well suited to the task. That is, only a precise system is capable of producing a correct, consistent, and concise ontology. Nonetheless, we were also concerned with the completeness of such an ontology, i.e., with the recall of the system. All in all, the results obtained indicate that MedInX performs with *both* high precision and recall, each showing an evaluation at approximately 95%. This is supported by an F-Measure, the Harmonic mean of recall and precision, whose evaluation is likewise at approximately 95%.

The clinical data included in the PDLs is a rich source of information, not only about the patient's medical condition, but also about the procedures and treatments performed in the hospital. Searching the content of the PDLs and ensuring the completeness of these documents is a process that still needs to be performed manually by expert physicians in health institutions. In order to support this process, we developed the MedInX clinical audit system. The main objective of the audit system is to help not only physicians, but hospital administrators and managers as well, to access the contents of the PDLs.

The clinical audit system uses the automatically populated MedInX ontology, which contains the structured information automatically extracted from the PDLs, and performs an automatic analysis of the content and completeness of the documents. The MedInX audit system uses the RDF query language SPARQL to query the ontology and retrieve relevant information from this resource. Several levels of information can be retrieved from this resource. An example of a developed rule, describing a complex scenario is given in Fig. 3.6. With this rule, we can find the PDLs that refer to *less than* a number of clinical Conditions and *over* a certain number of Chemical Procedures, namely, Medications and Active Substances. Both numbers used by the rule are user defined. The example of the figure uses the value 17 as the defined threshold after analysis of the most common values for these entities in the PDLs. The result of this rule allows identification of the outlier reports and suggests the need for content verification.

Tab. 3.2: Results of MedInX in the task of semantic classification.

	Precision	Recall	F-measure
Semantic classification	94.87	94.84	94.85

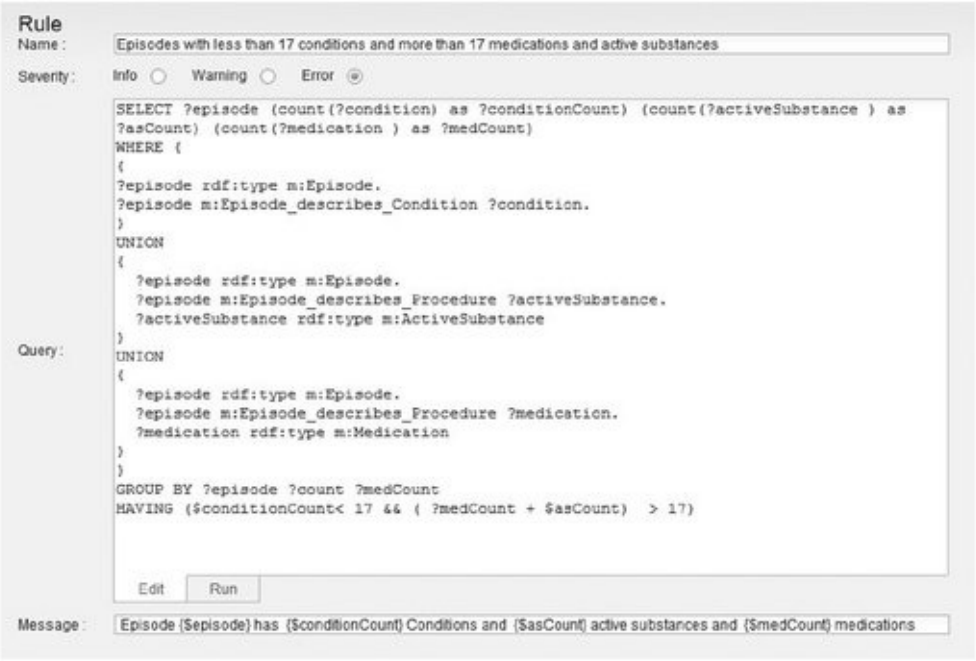


Fig. 3.6: MedInX audit system rule editor.

3.5.2 SPHInX – Semantic search of public health information in portuguese

The proof of concept system described next aims to perform semantic search on health information publicly available on the web in Portuguese.

3.5.2.1 System architecture

The SPHInX implementation is organized in four modules, respectively:

- Natural Language Processing: includes document content processing technologies to retrieve structured information from natural language texts. It is named NLP because it is based on technologies from the NLP area. In this part of the prototype, text is extracted from documents and enriched with the inclusion of POS tags, identification of named entities, and the construction of syntactic structures.
- Domain Representation: has tools for defining data semantics and associates it with samples of the NLP module output. System semantics is defined via ontology and, according to the ontology defined, it is necessary to provide examples of ontological classes and relations in sample documents. The examples are used to train semantic extraction models.
- Semantic Extraction and Integration: trains and applies semantic extraction models to all texts in order to obtain meaningful semantic information. It complements the extracted information

with external structured sources, e.g., geocodes and stores everything in a knowledge base conforming to the defined ontology.
– Search: information in the knowledge base can be searched and explored using natural language queries or via SPARQL.

3.5.2.2 Natural language processing

SPHInX was developed to process generic unstructured documents written in Portuguese, not only PDLs. Thus the NLP part was developed to handle the Portuguese language. The processing is organized in four sequential steps: (1) end of sentence detection; (2) Part-of-Speech (POS) tagging; (3) Named Entity Recognition (NER); and (4) syntactic parsing.

Text is extracted from documents, and then sentences are separated using the sentence boundary detector Punkt (Kiss & Strunk 2006). The sentence boundary detector step is highly relevant because all natural language processing is done in a per sentence fashion. This means that sentences define the processing context in the next NLP steps, which include algorithms able to use all the content of a sentence without using any content of the previous or following sentences.

After the split, sentences are enriched with POS tags assigned by TreeTagger (Schmid 1994): noun, verb, adjective, etc. TreeTagger was trained with a European Portuguese lexicon in order to be integrated in the system. Its outputs contain the word form followed by the assigned POS tag and the word lemma.

Named entities are discovered and classified by REMBRANDT (Cardoso 2012). Words belonging to a named entity are grouped using underscores. For instance, the names of the person John Stewart Smith become the single token John_ Stewart_Smith. Then, sentences are analyzed to determine their grammatical structure. This is done by MaltParser (Hall et al. 2007) and the result is a planar graph encoding the dependency relations among the words of each sentence.

3.5.2.3 Semantic extraction models

SPHInX creates one semantic extraction model for each ontology class and ontology relation. A model is a set of syntactic structure examples and counter examples that were found to encode the meaning represented by the model. It also contains a statistical classifier that measures the similarity between a given structure and the model's internal examples. The model is said to have positively evaluated a sentence fragment if the similarity is higher than a given threshold.

The algorithm for creating semantic extraction models was inspired in

two studies. The first is about extracting instances of binary relations using deep syntactic analysis. Suchanek, Ifrim & Weikum (2006) extracted one-to-one and many-to-one relations such as place and date of birth. They used custom-built decision functions to detect facts for each relation, and a set of statistical classifiers to decide if new patterns are similar to the learned facts. In our proof-of-concept prototype, this work was extended to include the extraction of one-to-many and many-to-many relations. The proof-of-concept prototype also implements a general purpose decision function based on the annotated examples instead of a custom-built function for each relation.

The second work is about improving entity and relation extraction when the process is learned from a small number of labeled examples, using linguistic information and ontological properties (Carlson et al. 2009). Improvements are made using class and relation hierarchy, information about disjunctions, and confidence scores of facts. This information is used to bootstrap more examples thereby generating more data to train statistical classifiers. For instance, when the system is confident about a fact, such as when it was annotated by a person, this fact is used as an instance of the annotated class and/or relation. This fact can also be used as a counter-example of all classes/relations disjoint with the annotated class/relation, and as an instance of super-class/super-relation. Moreover, facts discovered by the system with a high confidence score can be promoted to examples and included in a new round of training. In the proof-of-concept prototype, this creation of more examples is not active by default as it can lead to data over-fitting and should therefore be used carefully.

For the first version of SPHInX, the ontology about neurological diseases used in Mendonça et al. (2012) was adopted. The semantic extraction models were trained with a set of six manually annotated documents, of around fifty pages each, by a person familiar with the ontology but not related to the prototype development. The annotations were related to neurological diseases and respective symptoms, risk factors, treatments and related drugs.

3.5.2.4 Semantic extraction and integration

All sentence graphs are evaluated by the classifiers of all semantic models, and are collected in the case of forming a triple. A sentence fragment forms a *triple* if it is positively evaluated by two class models, one for subject and the other for object, along with one relation model binding the subject and object (Rodrigues, Dias & Teixeira 2011). Missing information according to the ontology is searched in external structured

information sources. For instance, unknown locations of entities with a fixed place (such as streets, organizations' headquarters, and some events) are queried using Google Maps API.

All collected triples are tentatively added to the knowledge base and their coherence is verified by a semantic reasoner. In SPHInX, reasoning is performed by an open-source reasoner for OWL-DL named Pellet (Sirin & Parsia 2004). All triples not coherent with the rest of the knowledge base are discarded, and a warning is issued. The remaining triples become part of the knowledge base.

3.5.2.5 Search and exploration

The search and exploration part of the system will be explained based on an illustrative example of use. The data for this example was obtained by having the system process fourteen previously unseen documents using the semantic extraction models trained earlier, plus the data already on the ontology at the time. Then, the same person that annotated the training documents was asked to suggest a few possible questions in Portuguese. Those questions were submitted to the system and one of them was selected for the example.

The interface, based on NLP-Reduce (Kaufmann, Bernstein & Fischer 2007), accepts natural language questions and generates SPARQL queries that are passed to a SPARQL engine. The system allows the user to enter a sentence, such as "memory loss is a symptom of what diseases?"

First, the question is transformed by removing all stop words and punctuation marks. The remaining words are stemmed and passed to a query generator that will use them to produce a SPARQL query in four steps:

1. Search for triples that contain one or more words of the query in the object property label. Triples are ranked according to the amount of words included in the label.
2. Search for properties that can be joined with the triples found in step 1. Thus, properties are searched using domain and range information of triples from step 1 along with the remaining query words. In the case of query words producing triples based on alternative object properties, the triples favored are those with the highest score from step 1. The triple set of this step is combined with the set of step 1, according to the ontology rules.
3. Search for data type property values that match the query words not matched in steps 1 and 2. Triples found are once again ranked considering the amount of words included in the property values. All triples found respecting the domain and range

restrictions of the set created in step 2 are added to it.

4. When there are no more query words left, the SPARQL query is generated to join the retrieved triples that achieved the highest scores in steps 1 to 3. Semantically equivalent duplicates are removed and the query is ready to be passed to a SPARQL endpoint.

So a question like “a perda de memória é um sintoma de que doenças?” which is roughly the Portuguese equivalent to “memory loss is a symptom of what diseases?” originates the SPARQL code presented in Fig. 3.7.

The output of the query is a table containing the variables of the SPARQL query. Depending on the type of information asked, the data on the table can be plain text as in case of data type property, presented in Tab. 3.3, or links to other ontological entities as in the case of object properties. In the case of the latter, it is then possible to navigate through the ontology by following those links.

Another way to output results is by presenting the graph of ontological concepts involved in the query. Figure 3.8 depicts the graph of the ontological elements used to compute the answer to the example query. As can be seen, there are more concepts involved than the ones included in the output table. The concepts involved in queries and not presented in the output table are typically the ones used to compute logical inferences.

Presenting the results in a graphic format allows users to navigate the information stored in the knowledge base while keeping a good overview of the ontology and how different concepts relate to each other.

SPARQL code	Explanation
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> prefix owl: <http://www.w3.org/2002/07/owl#> prefix MedInX: <http://www.MedInX.com/MedInX.owl#>	
select distinct * WHERE { ?NamedIndividual MedInX:hasSinalSintoma ?Sinal_Sintoma .	Get instances of relation hasSinalSintoma
?Sinal_Sintoma rdfs:label ?Sinal_Sintoma_label .	Get labels of the relation objects
FILTER(REGEX(?Sinal_Sintoma_label,'memoria','i')) .	... and accept those including word “memoria”
?NamedIndividual rdfs:label ?NamedIndividual_label .	Get labels of the relation subjects
?Sinal_Sintoma rdf:type MedInX:Sinal_Sintoma .	Relation objects have the type Sinal_Sintoma
?NamedIndividual rdf:type owl:NamedIndividual	... and subjects have the type NamedIndividual
}	

Fig. 3.7: SPARQL code generated by the natural language interface.

Tab. 3.3: Result set for the query “a perda de memória é um sintoma de que doenças?” (“memory loss is a symptom of what diseases?”).

NamedIndividual	Sinal_Sintoma	Sinal_Sintoma_label	NamedIndividual_label
“Doenca de Parkinson”	“Perda de memoria”	“Perda de memoria”@pt	“Doenca de Parkinson”@pt
“Doenca de Huntington”	“Perda de memoria”	“Perda de memoria”@pt	“Doenca de Huntington”@pt
“Esclerose Multipla”	“Perda de memoria”	“Perda de memoria”@pt	“Esclerose Multipla”@pt
“Demencia de tipo Alzheimer”	“Perda de memoria”	“Perda de memoria”@pt	“Demencia de tipo Alzheimer”@pt

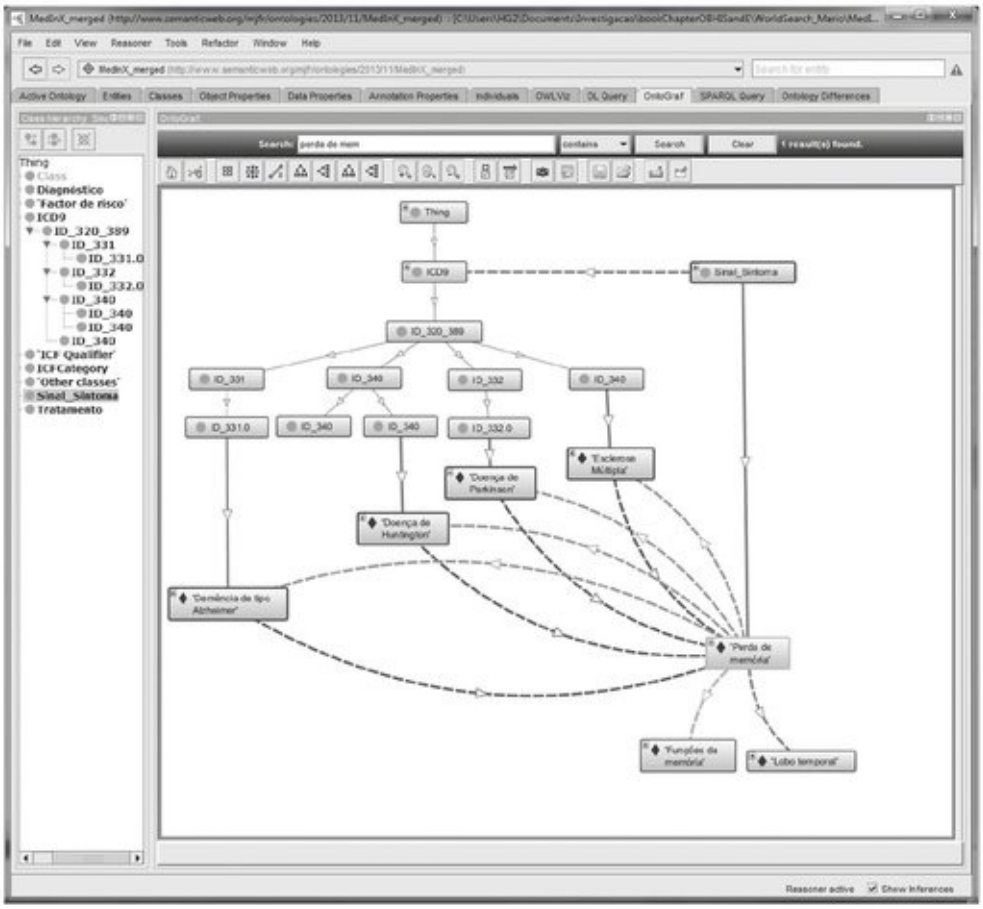


Fig. 3.8: Graph of the ontology concepts involved in the query example.

3.6 Conclusion

Taking into consideration the increasing need for semantic search of health information available originally in natural language, in this chapter a general architecture predicated on ontology-based information extraction to feed a search engine is proposed and instantiated in two systems. The first system, MedInx, allows semantic search of the information regarding a hospital’s discharge letters, and can be

generalized to the vast information in natural language stored in internal web-based hospital information systems. The second system, SPHInX, currently at an early stage of development, is capable of extracting information from public documents in Portuguese. For both systems, we present information on its architecture and components, and show via demonstration how these systems work.

We envision future developments of both systems that would address a much broader area, as both systems that we presented here only address a limited medical domain. Another equally important goal pivots on the improvement of the interaction of the user with these systems, making search and exploration a natural experience for professionals and laity alike.

Acknowledgments

This work was partially supported by World Search, a QREN project (QREN 11495) co-funded by COMPETE and FEDER, and the Portuguese Foundation for Science and Technology PhD grant SFRH/BD/27301/2006 to Liliana Ferreira. The authors also acknowledge the support from IEETA Research Unit, FCOMP-01-0124-FEDER-022682 (FCT-Pest C/EEI/UI0127/2011).

References

Aberdeen, J., Burger, J., Day, D., Hirschman, L., Robinson, P. & Vilain, M. (1995) MITRE: description of the Alembic system used for MUC-6. *Proceedings of the 6th conference on Message understanding* (pp. 141–155). New York, NY: Association for Computational Linguistics.

Abraham, J. & Reddy, M. (2007) Quality of healthcare websites: A comparison of a general-purpose vs. domain-specific search engine. *AMIA Annual Symposium Proceedings*, 858.

Aronson, A., Bodenreider, O., Demmer-Fushman, D., Fung, K., Lee, V. & Mork, J. (2007) From indexing the biomedical literature to coding clinical text. *BioNLP '07: Proceedings of the Workshop on BioNLP 2007*, (pp. 105–112).

Bast, H., Chitea, A., Suchanek, F. & Weber, I. (2007) ESTER: Efficient Search on Text, Entities, and Relations. *Proceedings of the 30th ACM*

SIGIR, (pp. 679–686).

Berners-Lee, T., Cailliau, R., Luotonen, A., Nielsen, H. F. & Secret, A. (1994) 'The World-Wide Web', *Commun ACM*, 37:76–82.

Berners-Lee, T. & Fischetti, M. (1999) *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor.* San Francisco: Harper Collins.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R. & Hellmann, S. (2009) DBpedia – A crystallization point for the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web – The Web of Data, 7*, 154–165.

Boyer, C., Baujard, V., Griesser, V. & Scherrer, J. R. (2001) 'HONselect: a multilingual and intelligent search tool integrating heterogeneous web resources', *Int J Med Inform,* 64:253–258.

Buitelaar, P. & Siegel, M. (2006) Ontology-based Information Extraction with SOBA. *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, (pp. 2321–2324).

Buscaldi, D., Rosso, P. & Arnal, E. S. (2005) A wordnet-based query expansion method for geographical information retrieval. *Working Notes for the CLEF Workshop.*

Can, A. B. & Baykal, N. (2007) 'MedicoPort: a medical search engine for all', *Comput Meth Prog Biomed,* 86:73–86.

Cardoso, N. (2012) 'Rembrandt – a named-entity recognition framework'. *LREC*, (pp. 1240–1243).

Carlson, A., Betteridge, J., Hruschka, E. R. & Mitchell, T. M. (2009) Coupling Semi-Supervised Learning of Categories and Relations. *SemiSupLearn '09: Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing* (pp. 1–9). Association for Computational Linguistics.

Cowie, J. & Lehnert, W. (1996) 'Information extraction', *Communications of the ACM,* 39:80–91.

Crammer, K., Dredze, M., Ganchev, K., Talukdar, P. P. & Carroll, S. (2007) Automatic code assignment to medical text. *BioNLP '07:*

Proceedings of the Workshop on BioNLP 2007 (pp. 129–136). Association for Computational Linguistics.

Cunha, J. P., Cruz, I., Oliveira, I., Pereira, A. S., Costa, C. T., Oliveira, A. M. & Pereira, A. (2006) The RTS project: Promoting secure and effective clinical. *eHealth 2006 High Level Conference*, (pp. 1–10).

Darmoni, S. J., Leroy, J. P., Baudic, F., Douyère, M. A. & Thirion, B. (2000) 'CISMeF: a structured health resource guide', *Met Inform Med*, 30–35.

Dragusin, R., Petcu, P., Lioma, C., Larsen, B., Jorgensen, H. L., Cox, I. J., Hansen, L. K., Ingwersen, P. & Winther, O. (2013) 'FindZebra: a search engine for rare diseases', *Int J Med Inform,* 82:528–538.

Embley, D. W., Campbell, D. M., Smith, R. D. & Liddle, S. W. (1998) Ontology-based extraction and structuring of information from data-rich unstructured documents. *Proceedings of the seventh international conference on Information and knowledge management* (pp. 52–59). ACM.

Esa, A. M., Taib, S. M. & Thi, H. N. (2010) Prototype of semantic search engine using ontology. *Open Systems (ICOS), 2010 IEEE Conference on* (pp. 109–114). IEEE.

Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S. & Yates, A. (2004) Web-Scale Information Extraction in KnowItAll (Preliminary Results). *WWW '04 – Proceedings of the 13th International World Wide Web Conference* (pp. 100–110). New York, NY, USA: Association for Computational Linguistics.

Ferreira, L., Teixeira, A. & Cunha, J. P. (2012) *Medical Information Extraction – Information Extraction from Portuguese Hospital Discharge Letters*. Lambert Academic Publishing.

Ferrucci, D. & Lally, A. (2004) 'UIMA an architectural approach to unstructured information', *Nat Lang Eng,* 10:327–348.

Friedman, C., Johnson, S. B., Forman, B. & Starren, J. (1995) 'Architectural requirements for a multipurpose natural language processor in the clinical environment'. *Proc Annu Symp Comput Appl Med Care*, (pp. 347–351).

Gaizauskas, R. & Wilks, Y. (1998) 'Information extraction: beyond document retrieval', *J Doc,* 54:70–105.

Gaudinat, A., Ruch, P., Joubert, M., Uziel, P., Strauss, A., Thonnet, M., Baud, R., Spahni, S., Weber, P., Bonal, J., Boyer, C., Fieschi, M. & Geissbuhler, A. (2006) 'Health search engine with e-document analysis for reliable search results', *Int J Med Inform,* 75:73–85.

Gruber, T. R. (1993) 'A translation approach to portable ontology specifications', *Knowledge Acquisition,* 5:199–220.

Guarino, N. (1998) Formal Onthology in Information Systems. *FIOS'98 – Proceedings of the First International Conference on Formal Ontology in Information Systems* (pp. 3–15). IOS Press.

Guha, R. & McCool, R. (2003) 'TAP: a Semantic Web platform', *Computer Networks*, 557–577.

Guha, R., McCool, R. & Miller, E. (2003) Semantic search. *Proceedings of the 12th international conference on World Wide Web* (pp. 700–709). ACM.

Hahn, U., Romacker, M. & Schulz, S. (2002) 'Creating Knowledge Repositories From Biomedical Reports: The MEDSYNDIKATE Text Mining System'. *Pac Symp Biocomput*, (pp. 338–349).

Hall, J., Nilsson, J., Nivre, J., Eryigit, G., Megyesi, B., Nilsson, M. & Saers, M. (2007) *Single Malt or Blended? A Study in Multilingual Parser Optimization*. (pp. 933–939). Association for Computational Linguistics.

Hobbs, J. R. (2002) 'Information extraction from biomedical text', *J Biomed Inform,* 35:260–264.

Jindal, V., Bawa, S. & Batra, S. (2014) 'A review of ranking approaches for semantic search on Web'. *Inf Process Manage,* 50(2): 416–425.

Kamath, S. S., Piraviperumal, D., Meena, G., Karkidholi, S. & Kumar, K. (2013) A semantic search engine for answering domain specific user queries. *Communications and Signal Processing (ICCSP), 2013 International Conference on* (pp. 1097–1101). IEEE.

Kaufmann, E., Bernstein, A. & Fischer, L. (2007) NLP-Reduce: A "naive" but

Domain-independent Natural Language Interface for Querying Ontologies. *ESCW'07 – Proceedings of the 6th International Semantic Web Conference.*

Kiss, T. & Strunk, J. (2006) 'Unsupervised multilingual sentence boundary detection', *Compu Linguist,* 32:485–525.

Kruse, P. M., Naujoks, A., Rosner, D. & Kunze, M. (2005) Clever search: A wordnet based wrapper for internet search engines. *arXiv preprint cs/0501086*.

Lee, L. (2004) "I'm sorry Dave, I'm afraid I can't do that": Linguistics, Statistics, and Natural Language Processing circa 2001. In C. O. Board, *Computer Science: Reflections on the Field, Reflections from the Field* (pp. 111–118). Washington DC: The National Academies Press.

Lei, Y., Uren, V. & Motta, E. (2006) Semsearch: A search engine for the semantic web. Proceedings of the 15th International Conference on *Managing Knowledge in a World of Networks* Berlin, Heidelberg (pp. 238–245). Berlin, Heidelberg: Springer-Verlag.

Maedche, A., Maedche, E., Neumann, G. & Staab, S. (2003) *Bootstrapping an Ontology-based Information Extraction System,* pp. 345–359. Heidelberg, Germany: Physica-Verlag GmbH.

Màrquez, L., Carreras, X., Litkowski, K. C. & Stevenson, S. (2008) 'Semantic role labeling: an introduction to the special issue', *Comput Linguist,* 34:145–159.

McGray, A. T., Sponsler, J. L., Brylawski, B. & Browne, A. (1987) 'The role of lexical knowledge in biomedical text understanding'. *SCAMC*, (pp. 103–107).

McNaught, J. & Black, W. (2006) Information extraction. In Ananiadou, S. & McNaught, J. *Text Mining for Biology and Biomedicine* (pp. 143–178). Norwood: Artech House.

Mendonça, R., Rosa, A. F., Oliveira, J. L. & Teixeira, A. J. (2012) Towards Ontology Based Health Information Search in Portuguese – A case study in Neurologic Diseases. *CISTI'2012 – 7th Iberian Conference on Information Systems and Technologies.*

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. & Miller, K. J. (1990) 'Introduction to wordnet: An on-line lexical database', *Int J Lexico*, 235–244.

Moldovan, D. I. & Mihalcea, R. (2000) 'Using wordnet and lexical operators to improve internet searches', *Internet Comput, IEEE,* 4:34–43.

Pakhomov, S. A. (2005) High throughput modularized NLP system for clinical text. *Proceedings of the ACL 2005 on Interactive poster and demonstration sessions* (pp. 25–28). New York, NY: Association for Computational Linguistics.

Rocha, C., Schwabe, D. & Aragao, M. P. (2004) A hybrid approach for searching in the semantic web. *Proceedings of the 13th international conference on World Wide Web* (pp. 374–383). ACM.

Rodrigues, M., Dias, G. P. & Teixeira, A. (2011) Ontology Driven Knowledge Extraction System with Application in e-Government. *Proc. of the 15th APIA Conference*, (pp. 760–774).

Schmid, H. (1994) Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of the International Conference on New Methods in Language Processing.*

Sirin, E. & Parsia, B. (2004) Pellet: An OWL DL Reasoner, In Haarslev, V. and Möller, R. (eds), International Workshop on Description Logics (DL'04), pp. 212–213. British Columbia, Canada: Whistler.

Suchanek, F. M., Ifrim, G. & Weikum, G. (2006) LEILA: Learning to Extract Information by Linguistic Analysis. *Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge* (pp. 18–25). New York, NY: Association for Computational Linguistics. Sydney, Australia

Suchanek, F. M., Kasneci, G. & Weikum, G. (2007) YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia. *WWW '07 – Proceedings of the 16th International World Wide Web Conference* (pp. 697–706). New York, NY: Association for Computational Linguistics.

USNLM. (2008) UMLS Knowledge Sources. United Stated National Library of Medicine.

Wang, L., Wang, J., Wang, M., Li, Y., Liang, Y. & Xu, D. (2012) 'Using internet search engines to obtain medical information: a comparative study'. *J Med Internet Res, 14*.

Wang, C., Xiong, M., Zhou, Q. & Yu, Y. (2007) PANTO: A Portable Natural Language Interface to Ontologies. *ESWC2007 – Proceedings of the 4th European Semantic Web Conference* (pp. 473–487). Berlin/Heidelberg: Springer.

Wimalasuriya, D. C. & Dou, D. (2010) 'Ontology-based information extraction: An introduction and a survey of current approaches', *J Inform Sci,* 36:306–323.

World Health Organization. (n.d.) *International Classification of Diseases*. Accessed on 01/27/2014, http://www.who.int/classifications/icd/en/

Wu, F., Hoffmann, R. & Weld, D. S. (2008) Information Extraction from Wikipedia: Moving Down the Long Tail. *KDD '08 – Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 731–739). New York, NY: Association for Computational Linguistics.

Yates, A., Banko, M., Broadhead, M., Cafarella, M. J., Etzioni, O. & Soderland, S. (2007) TextRunner: Open Information Extraction on the Web. *NAACL-HLT (Demonstrations) – Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 25–26). Morristown, NJ: Association for Computational Linguistics.

Yildiz, B. (2007) Ontology-driven Information Extraction. *PhD Thesis*. Vienna University of Technology.