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

1 http://arrowsmith.psych.uic.edu/arrowsmith_uic/index.html
2 http://ibmi3.mf.uni-lj.si/bitola/
3 http://www.nactem.ac.uk/medie/
4 http://www.nactem.ac.uk/EventMine/

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]

5 To support research within the IR community, a special program was erected by the National Institute of Standards and Technology (NIST) and U.S. Department of Defense back in 1992. The program was called The Text Retrieval Conference (TREC), which is part of the TIPSTER Text program. TREC consist of an ongoing series of workshops focusing on a list of different IR research areas or tracks. Its purpose is to support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies and to increase the speed of lab-to-product transfer of technology. The TREC test collections and evaluation software are available to the retrieval research community writ large, so organizations can evaluate their own retrieval systems at any time. TREC has successfully met its dual goals of improving the state-of-the-art in information retrieval and in facilitating technology transfer. Retrieval-system effectiveness approximately doubled in the first six years of TREC. In TREC-2003 there was a retrieval track dedicated to Genomics, and in 2004 this track was centered on tagging genes and proteins in relevant documents: http://trec.nist.gov/.

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

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

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.

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

6 https://www.nlm.nih.gov/pubs/factsheets/dif_med_pub.html

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).

7 http://opennlp.apache.org/

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-based approaches 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

8 http://software.ucv.ro/~cmihaescu/ro/teaching/AIR/docs/Lab1-Algorithms%20for%20Information%20Retrieval.%20Introduction.pdf

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 $e_i = \left|f_i - y_i\right|$, 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

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)

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.

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]

9 https://plot.ly
10 http://services.alphaworks.ibm.com/manyeyes/
11 http://www.grapheur.com/
12 http://www.visumap.net/
13 http://www.sas.com

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

14 http://www.netowl.com

15 http://www-01.ibm.com/common/ssi/cgi-bin

16 http://www.cs.waikato.ac.nz/ml/weka

17 http://zope.bioinfo.cnio.es/bionlp_tools/get_all_bionlp_tools_out?SUBMIT#equal#Submit+Query

18 http://bioinformatics.ua.pt/becas/#!/

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

19 http://www.nactem.ac.uk/Kleio/

20 http://www.ncbi.nlm.nih.gov/CBBresearch/Wilbur/IRET/PIE/

21 http://www.nactem.ac.uk/tsujii/medie/

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.

22 http://www.nextbio.com/b/nextbioCorp.nb

23 http://www.neuinfo.org/

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.

24 http://www.diseasesdatabase.com
25 https://dynamed.ebscohost.com
26 http://www.gpnotebook.co.uk/homepage.cfm
27 http://www.icd9data.com
28 http://www.lumrix.net/icd-9.php
29 http://icd9cm.chrisendres.com
30 http://medind.nic.in/imvw
31 http://openmed.nic.in/
32 http://www.silverplatter.com
33 http://www.medicine.ox.ac.uk/bandolier/
34 http://www.mcmaster.ca/Cochrane/Cochrane/revabstr/abidx.htm
35 http://www.enb.org.uk/hcd.htm

- 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.

36 http://www.popline.org

37 https://clinicalinformatics.stanford.edu/

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).