Chapter 14

BIOMEDICAL TEXT MINING: A SURVEY OF RECENT PROGRESS

Matthew S. Simpson
Lister Hill National Center for Biomedical Communications
United States National Library of Medicine, National Institutes of Health
simpsonmatt@mail.nih.gov

Dina Demner-Fushman
Lister Hill National Center for Biomedical Communications
United States National Library of Medicine, National Institutes of Health
ddemner@mail.nih.gov

Abstract The biomedical community makes extensive use of text mining technology. In the past several years, enormous progress has been made in developing tools and methods, and the community has been witness to some exciting developments. Although the state of the community is regularly reviewed, the sheer volume of work related to biomedical text mining and the rapid pace in which progress continues to be made make this a worthwhile, if not necessary, endeavor. This chapter provides a brief overview of the current state of text mining in the biomedical domain. Emphasis is placed on the resources and tools available to biomedical researchers and practitioners, as well as the major text mining tasks of interest to the community. These tasks include the recognition of explicit facts from biomedical literature, the discovery of previously unknown or implicit facts, document summarization, and question answering. For each topic, its basic challenges and methods are outlined and recent and influential work is reviewed.

Keywords: Biomedical information extraction, named entity recognition, relations, events, summarization, question answering, literature-based discovery

1. Introduction

The state of biomedical text mining is reviewed relatively regularly. The recent surveys [238, 237], special journal issues [85, 29], and books [12] in this area indicate that general-purpose text and data mining tools are not well-suited for the biomedical domain because it is highly specialized.

Despite the restricted nature of the domain, biomedical text mining is of interest not only to researchers but to the general public as well (perhaps unbeknownst to them). The recent biomedical advances that have prevented or altered the course of many diseases are undoubtedly valued by all. Progress in biomedicine is attributable to advances in the understanding of disease mechanisms and the societal and commercial value of researching these mechanisms as well as the approaches for the prevention and cure of diseases.

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 and explicit and implicit associations among facts.

Due to the specific goals of biomedical text mining, biologists and clinicians are better positioned to define useful text mining tasks. Cohen and Hunter [33] note that the most fruitful approaches to biomedical text mining will combine the efforts and leverage the abilities of both biologists and computational linguists. Biologists and clinicians will leverage their ability to focus on specific tasks and experience in using the unparalleled publicly available domain-specific knowledge sources whereas text mining specialists will provide system components and design and evaluate methods.

The sheer size of the so-called bibliome (the entirety of the texts relevant to biology and medicine) dictates a stepwise approach to biomedical text mining. The goal of the first step is to reduce the set of text documents to be mined. This reduction is most commonly achieved using domain-specific information retrieval approaches, as described in *Information Retrieval: A Health and Biomedical Perspective* [65]. Alternatively, document sets can be selected using clustering and classification [98, 177, 22]. As discussed later in this chapter, the meaning and grammar of biomedical texts are so intertwined that all surveys dedicate a section to natural language preprocessing and grammatical analysis. However, this chapter presents these methods (e.g., tokenization, part-of-speech tagging, parsing, etc.) as needed to describe the reviewed text mining approaches.

This survey of recent advances in biomedical text mining begins with a discussion of the resources available for mining the biomedical literature. It then proceeds to describe the basic tasks of named entity recognition and relation and event extraction. The more complex tasks of summarization, question answering, and literature based discovery are described thereafter. The chapter concludes with a discussion of open tasks and potentially high-impact avenues for further development of the domain.

2. Resources for Biomedical Text Mining

The primary resource for biomedical text mining is obviously text, and this section introduces some widely-used text collections in the biomedical domain. Although text mining does not require the use of specialized or annotated corpora, manually annotated collections are often more useful than the original texts alone. For example, the original conception of literature-based discovery [189] was facilitated by the use of Medical Subject Headings (MeSH®), which are controlled vocabulary terms added to bibliographic citations during the process of MEDLINE® indexing. With the growth of publicly available annotated collections, the biomedical language processing community has begun focusing on common interchangeable annotation formats, guidelines, and standards, which this section also discusses. After describing these resources, the section concludes with a description of equally important lexical and knowledge-based repositories, widely-used biomedical text mining tools and frameworks, and registries that provide overviews and links to text collections and other resources.

2.1 Corpora

Whether text mining is viewed in the strict sense of discovery or in the broader sense that includes all text processing and retrieval steps leading towards discovery, MEDLINE was the first—and remains the primary—resource in biomedical text mining. The MEDLINE database contains bibliographic references to journal articles in the life sciences with a concentration on biomedicine, and it is maintained by the U. S. National Library of Medicine® (NLM®). The 2011 MEDLINE contains over 18 million references published from 1946 to the present in over 5,500 journals worldwide.

Abstracts of biomedical literature can be obtained in a variety of different ways. For text mining purposes, MEDLINE/PubMed® records can be downloaded using the Entrez Programming Utilities [131]. Alternatively, subsets of MEDLINE citations can be obtained from the archives of community-wide evaluations that use MEDLINE, as well as individual research groups that share their annotations. Such collections include the historic OHSUMED [200] set containing all MEDLINE citations in 270 medical journals published over a five-year period (1987–1991) and a more recent set of TREC Genomics Track data [201] that contains ten years of MEDLINE citations (1994–2003). Stand-off annotations supporting information retrieval relevance, document classification, and question answering are available for portions of these collections. Whereas TREC collections provide access to MEDLINE spans over a given time period, other collections are task-oriented. For example, the GENIA corpus [90] contains 1,999 MEDLINE abstracts retrieved using the MeSH terms "human," "blood cells," and "transcription factors." The GENIA corpus is currently the most thoroughly annotated collection of MEDLINE abstracts. It is annotated for part-of-speech, syntax, coreference, biomedical concepts and events, cellular localization, disease-gene associations, and pathways. In addition, the GENIA corpus is one of the three constituents of the BioScope corpus [217], which provides GENIA MEDLINE abstracts, five full-text articles, and a collection of radiology reports annotated with negation and modality cues as well as scope. Other topically-annotated collections of MEDLINE abstracts include the earlier BioCreAtIve collections [69, 97] and the PennBioIE corpus [105, 106]. The PennBioIE corpus contains 1100 abstracts for cytochrome P-450 enzymes and 1157 oncology abstracts with annotations for paragraphs, sentences, tokens, parts-of-speech, syntax, and biomedical entities. Finally, the Collaborative Annotation of a Large Biomedical Corpus (CALBC) initiative [26] has proposed the creation of a "silver standard" corpus that contains MEDLINE abstracts that have been automatically annotated with biomedical entities by the initiative participants. This corpus has just recently become publicly available.

Being informative and undoubtedly useful for text mining, MEDLINE abstracts do not contain all the information presented in full-text articles. Some information (e.g., the exact settings of an experiment or the discussion of the results) is almost exclusively contained in the body of an article. The promise of a qualitative increase in the amount of useful information brought about several full-text collections. For example, the TREC Genomics Track dataset contains about 160,000 full-text articles from about 49 genomics-related journals, which were obtained in HTML format from the Highwire Press [66] electronic distribution of the journals. Another collection of full-text articles annotated with relevance to patients' case descriptions was developed in the ImageCLEF evaluations [127, 84]. The Colorado Richly Annotated Full Text Corpus [38] adds to the growing body of semantically and syntactically annotated full text collections (including the full-text portion of the BioScope collection mentioned above). Finally, the largest publicly available source of original, full-text articles is the Open Access subset of PubMed Central [154].

With the growing interest in clinical text mining and biosurveillance, several public collections of clinical text have recently become available. These collections include reports in the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC II) database [171], the Pittsburgh collection of clinical reports [211], and the annotated i2b2 collections [214, 213, 215, 216]. Several recent studies used the Web (i.e, Twitter and health-related blogs and community sites) as a corpus, but it is not clear if the collections created for these studies are publicly available or not.

2.2 Annotation

The annotation of biomedical text adds information to a document collection that can later be exploited for text mining purposes. In general, document annotation in the biomedical domain follows the principles set forth in open-domain natural language processing (NLP) by adding annotations at multiple levels of linguistic analysis. The various aspects involve grammatical (including morphological and syntactical), semantic, and pragmatic annotations [103]. Grammar and meaning are so intertwined that most annotation efforts combine the two. For example, corpus creators might decide to annotate named entities of interest only in noun phrases. As an alternative, Wilbur et al. [222] focus on annotating the “information-bearing fragments within scientific text” without specifying any grammar restrictions. The authors define the following five annotation axes: Focus, Polarity, Certainty, Evidence, and Direction. These classes are primarily used at the sentence level, and sentences may be broken as needed if a change in one of the annotations aspects is detected. However, even meaning-centric annotations cannot be completely grammar-free. For example, one of the clues for annotating fragments as *Evidence* is a past tense verb indicating an observation or finding. The guidelines published by the authors [178] are a good starting point for developing other text-mining annotation guidelines in the biomedical domain.

There are three approaches to the annotation of biomedical text. These methods include (1) a complete manual annotation that is based on annotators’ knowledge; (2) an assisted annotation, in which the output of an annotation tool is manually corrected; and (3) an ontology-based annotation—either manual or assisted—in which only terms and relations present in an existing knowledge source are annotated. Each of these approaches has its strengths and weaknesses. For example, an assisted annotation is usually more consistent, but it may be biased. Similarly, an ontology-based annotation will likely be biased towards known facts. Having more than one annotator for each text document and having various annotator groups can compensate for such biases [15].

In addition to generic information extraction tools that can be used to assist in annotation (described below), several text mining tools have been developed to specifically support the annotation process. Examples of widely-used tools for annotating biomedical text include Knowtator [140] and eHOST (Extensible Human Oracle Suite of Tools) [48], the later of which is increasingly used for the annotation of clinical text. In order for such tools to be useful, they must be easy to use, support various annotation types, and allow collaborative annotation, among other factors [115, 47].

2.3 Knowledge Sources

The biomedical domain offers a rich set of knowledge sources supporting text mining applications. The Unified Medical Language System® (UMLS®) [111], a compendium of controlled vocabularies that is maintained by NLM, is the most comprehensive resource, unifying over 100 dictionaries, terminologies, and ontologies in its Metathesaurus. It also provides a semantic network that represents relations between Metathesaurus entries, a lexicon that contains lexicographic information about biomedical terms and common English words, and a set of lexical tools. Overall, NLM provides over 200 knowledge sources and tools that can be used for text mining [210]. Other sets of ontologies are maintained by collaborative effort in the OBO Foundry [143] and the National Center for Biomedical Ontology (NCBO) [129]. The NCBO ontologies are accessed and shared through BioPortal [130]. Other major centers that maintain specialized resources for biomedical text mining include the British National Centre for Text Mining [132] and the European Bioinformatics Institute [52].

In addition to these broad-coverage resources, the biomedical domain offers in-depth knowledge sources focused on specific subdomains of biomedicine. For example, the Pharmacogenomics Knowledge Base [152] is a collection of scientific publications annotated with primary genotype and phenotype data, gene variants, and gene-drug-disease relationships. The annotations are downloadable for individual research purposes. Another specialized source, the Neuroscience Information Framework [134], includes an ontology covering brain anatomy, cells, organisms, diseases, techniques, and other areas of neuroscience.

The best knowledge source for a given text mining task is determined by the nature of the problem at hand. For example, mining the scientific literature for relations between genes, diseases, and drugs first requires recognizing instances of these entities. To aid in this task, a researcher might rely on knowledge of the terms' corresponding semantic types in the UMLS or instead may chose to use individual knowledge sources, such as the Gene Ontology [16], SNOMED Clinical Terms® [188], or the FDA Approved Drug Products with Therapeutic Equivalence Evaluations (Orange Book) [209]. Approaches to the various text mining tasks in the biomedical domain make extensive use of the resources described in this section and sometimes derive meta-resources for a specific task. For example, Rinaldi et al. [163] define several entity types needed for mining the literature for protein interactions (protein/gene names, chemical compounds, cell lines, etc.) and then automatically aggregate terms extracted from curated resources such as the UMLS, Affymetrix identifiers for micro array probes, organism databases, and others into a list of 2,347,734 terms.

2.4 Supporting Tools

The variety and purpose of the tools supporting biomedical text mining echoes that of the knowledge sources described above. The following discussion of text mining tools omits applications described in recent surveys and instead focuses on the basic, widely used tools for identifying named entities and relations and the platforms that allow building text mining pipelines.

The most widely used tool for named entity recognition that is based upon the UMLS is MetaMap [14]. MetaMap is a highly configurable application that identifies UMLS Metathesaurus concepts in free text. Because MetaMap provides a wide range of configuration options and relies on the entire UMLS Metathesaurus, it is not easy to determine the best configuration for a given task. However, exploring the options using the interactive MetaMap website may aid with such choices. MetaMap, which was provided as service until recently, is now open source and available for download. Two statistical tools widely used for biological named entity recognition are ABNER [176] and BANNER [101]. Both ABNER and BANNER are based on conditional random fields and rely on a wide array of features. Unlike ABNER, BANNER avoids semantic features, but it uses syntactic features. Both systems exploit such domain-specific language characteristics as capitalization, word shapes, prefixes, suffixes, and Greek letters.

Tools for relation extraction are not yet as readily accessible as entity recognition tools. Kabiljo et al. [83] compared available tools for identifying biomedical relations (AkanePPI, Whatizit, and OpenDMAP) to a simple, regular expression-based approach and found that the simple approach performed surprisingly well. The authors conclude that high recall (around 90%) is achievable for extracting gene-protein relations when the available tools are combined.

A recent trend in tool development and use is the assemblage of pipelines based on open-source frameworks, such as the Generalized Architecture for Text Engineering (GATE) [39] and the Unstructured Information Management Architecture (UIMA) [54]. The most mature system for clinical text processing (ranging from identifying patients' problems to events) is MedLEE [58]. Descriptions of other systems and clinical text mining tasks can be found in a recent review [41].

This section has presented only a snapshot of open-domain biomedical text mining resources. By its nature, the information contained herein will become dated sooner than the other material presented in this chapter. To compensate for the rapid progress of research related to biomedical text mining, many researchers maintain websites with links to useful resources (e.g., BioNLP [19]). Realizing that this task is too time consuming for individual researchers, the U. S. Department of Veterans Affairs and NLM provide a registry of biomedical text mining tools, known as ORBIT, which is maintained by the research community [144].

3. Information Extraction

A goal of many biomedical text mining tasks is the identification of explicitly stated facts. *Information extraction* refers to the process by which structured facts are automatically derived from unstructured or semi-structured text. In the biomedical domain, unstructured text commonly includes scientific articles appearing in the biomedical literature as well as clinical narratives found in electronic health records or other clinical information systems. Although the information extracted from these sources can be the target of information retrieval systems, information extraction is often performed as an initial processing step for other biomedical text ming applications (Sections 4–6).

Biomedical information extraction technology has undergone rapid development in recent years, spurred in part by community-wide evaluations that have been focussed specifically on text mining within the biomedical domain. Some examples of recent evaluation forums include BioCreAtivE [69, 97], BioNLP [89, 88], i2b2 [214, 213, 215, 216], JNLPBA [91], and LLL [133] shared tasks. The strong interest in community-wide evaluation efforts such as these is reflective of the growing volume of unstructured biomedical text available electronically in databases such as MEDLINE or in clinical information systems.

Three major subtasks of information extraction are particularly relevant for processing biomedical text. First, named entity recognition is a task that seeks to identify and classify biomedical entities into predefined categories such as the names of proteins, genes, or diseases. Often, extracted entities are normalized to canonical, unambiguous representations with the aid of ontological resources and further classified into semantic categories. The second subtask of information extraction relevant to the biomedical domain is relation extraction, which aims to detect binary relationships among named entities. Examples include gene-disease relationships, protein-protein interactions, and medical problem-treatment relationships. Finally, the third major subtask, event extraction, seeks to identify highly complex relations among extracted entities. Events relevant to the biomedical domain include, for example, gene expression and regulation and protein binding.

Although each of these subtasks are distinct in the type of information they aim to extract, they achieve their goals by employing similar methods, which include machine learning, statistical analysis and other techniques of natural language processing. Challenges and approaches to the subtasks of biomedical information extraction are discussed below.

3.1 Named Entity Recognition

Biomedical Named Entity Recognition (NER) refers to the task of automatically identifying occurrences of biological or medical terms in unstructured text. Common entities of interest include gene and protein names, medical problems and treatments, drug names and their dosages, and other semantically well-defined data classifiable within the biomedical domain [104]. Although commonly discussed as a single task, NER is typically a three-step process that involves determining an entity's substring boundaries within the text, assigning the entity to a predefined class or category, and selecting the preferred name or unique identifier of the concept that the entity names. This last subtask, entity normalization, is sometimes addressed as a separate problem from NER, but it is briefly discussed here in the context of describing the many issues that make NER a challenging task in the biomedical domain.

NER is particularly challenging for biomedical text due to a variety of reasons. The most basic obstacle results from the dynamic nature of scientific discovery. In the biomedical domain, there exists a vast amount of semantically relevant entities that is constantly and rapidly increasing as new scientific discoveries are made [226]. This ever-growing list of relevant terms is problematic for NER systems that rely only on a dictionary of known terms or other curated resources to identify named entities since these resources can never be complete as long as scientific progress continues.

Another challenge to biomedical NER is synonymy. In biomedical literature, the same concept may be expressed using different words. For example, “heart attack” and “myocardial infarction” refer to the same medical problem so an NER system should recognize these terms as instances of the same concept, despite being expressed differently. When many synonyms for a particular concept are in use, it becomes difficult to integrate knowledge from multiple sources without a comprehensive synonymy resource such as the UMLS Metathesaurus or Gene Ontology. However, given the rapidly increasing number of biomedical entities, these resources are unlikely to be complete at any given moment, resulting in some synonymy relationships that may not be captured.

Finally, the abundant use of acronyms and abbreviations in biomedical literature make it difficult to automatically identify the concepts to which these terms refer. Often, successful acronym and abbreviation resolution depends greatly on the context in which the terms appear since the same term can refer to different concepts. For example, the abbreviation *RA* can refer to “right atrium,” “rheumatoid arthritis,” “refractory anemia,” “renal artery,” or one of several other concepts [148]. To address the challenges associated with acronyms, abbreviations, and synonymy, NER systems typically perform some form of entity normalization.

Entity normalization is a subtask of NER and refers to the process of mapping entity occurrences to their canonical, preferred names. Although a challenging task itself, entity normalization can help resolve issues resulting from synonymous terms and ambiguous acronyms and abbreviations by associating these entities with unique, unambiguous representations. Often, since there may not be community-wide agreement on the preferred name for a given entity, the goal of entity normalization is to map an entity instance to the unique identifier of a concept in a terminology resource. In general, entity normalization requires the existence of such terminology resources, though they may be incomplete. Since normalization is such a crucial component of many NER systems, it is often an implied processing step after identifying entity boundaries and assigning them to a category. However, the entity normalization subtask may be evaluated independently of these subtasks, as was the case in recent BioCreAtivE shared task evaluations [67, 126].

For NER systems that analyze large amounts of biomedical text, it is important to consider the quality that can be expected of the methods being utilized. Typically, the performance of NER systems is measured in terms of precision, recall, and F-score. However, a variety of issues make these measurements difficult to reliably obtain and compare.

One issue is the availability of large, high-quality annotated corpora to serve as the ground truth on which to base NER system evaluations. The ground truth corpora must be large enough to allow the extrapolation of experimental results to large text collections, such as the entirety of MEDLINE, and the annotations should exhibit high inter-annotator agreement and reflect expert-level judgement. However, while the size of a ground truth data set is crucial, annotation errors do not necessarily pose an insurmountable problem to system evaluation, especially if the data set is sufficiently large. For example, Uzuner et al. [216] demonstrated that errors in the ground truth for a recent i2b2 shared task evaluation could affect the relative performance of competing NER systems by 0.05% at most.

Another issue to consider when evaluating NER systems is how to define the boundaries of a correctly identified entity. A strict evaluation requires both the left and right boundaries of an extracted entity to exactly match those of the ground truth annotations while a loose evaluation requires only that the extracted entity boundaries overlap those of the annotations [104]. Olsson et al. [142] showed that the choice of a strict or loose evaluation affects the relative performance of NER systems and suggested several scoring criteria for different application needs.

Recent community-wide evaluations have demonstrated that NER systems are typically capable of achieving favorable results. For example, the best performing systems achieved F-scores of 0.83 and 0.87 for the first [226] and second [187] BioCreAtIve gene mention recognition tasks, 0.85 for the i2b2 concept extraction task [216], and 0.73 for the JNLPBA bio-entity recognition task [91]. Although NER systems may be tailored for a particular information extraction task, their primary methods can broadly be grouped as following one of several basic approaches, which are discussed below.

Dictionary-based methods, one of the most basic biomedical NER approaches, utilize comprehensive lists of biomedical terms in order to identify entity occurrences in text. Such systems determine whether a word or group of words selected from the text exactly matches a term from some biomedical resource. When used as stand-alone methods, dictionary-based approaches generally exhibit reasonably high precision, but they suffer from poor recall due to the existence of spelling mistakes and morphological variants [207]. However, low precision is also possible due to homonymy [68]. For example, many gene names and abbreviations (e.g, "an," "by," and "can") share lexical representations with common English words [99]. For these reasons, some form of inexact string matching is commonly utilized to improve the precision and recall of dictionary-based approaches. Some methods improve performance by first generating spelling variants for the terms in a biomedical resource, and then by appending these additional terms to the underlying word lists [205, 204]. The methods are then able perform exact matching using the augmented resource. Other methods utilize algorithms such as BLAST® [10, 11] to perform approximate string matching instead of exact matching [100]. Despite these improvements, dictionary-based methods are most often used in conjunction with more advanced NER approaches.

Another approach to NER is to define rules that describe the composition patterns of named biomedical entities and their context. Examples of rule-based approaches include the EMPathIE and PASTA systems [78, 61], which use context free grammars that recognize enzyme interactions and protein structures. Other systems utilize pattern-based rules that exploit the orthographic and lexical characteristics of targeted entity classes in order to recognize protein [59] and chemical [128] names. These simpler methods may be improved by additionally considering contextual information [70] and the results of syntactic parsing for determining entity boundaries [57]. However, while rule-based approaches typically achieve better performance than dictionary-based approaches, manual generation of the required rules is a time-consuming process, and, since the rules are usually very specific in order to achieve high precision, they are difficult to extend to other entity classes.

It is increasingly common for NER approaches to rely on statistical methods instead of, or in combination with, dictionary- and rule-based approaches. Unlike the previously described approaches, statistical methods typically rely on some form of machine learning algorithm to identify biomedical entities. While supervised machine learning approaches must be trained with observations taken from large annotated corpora, recent work has investigated the automatic generation of training data for the NER task through the use of bootstrapping and other semi-supervised statistical techniques [218, 125, 212]. Common statistical methods used for NER can be grouped as either classification- or sequence-based approaches.

Classification-based approaches transform the NER task into a classification problem, which can either be applied to individual words or groups of words. Common classifiers used for biomedical NER include

Naïve Bayes [139] and Support Vector Machine (SVM) [86, 118, 196, 224] classifiers. Although it is possible to classify multi-word phrases, a popular approach follows the BIO tagging scheme [157], where individual tokens are classified as being at the beginning (B) of an entity, inside (I) the boundaries of an entity, or outside (O) the boundaries of an entity. However, despite its success, this tagging scheme can be problematic if entity boundaries overlap, and several authors have addressed the problem of recognizing nested biomedical entities [62, 8]. The performance of classification-based approaches is highly dependent on the choice of features used for training, and many authors have explored various feature combinations. For example, Kazama et al. [86] and Mitsumori et al. [118], consider morpho-syntactic properties of named entities, Takeuchi and Collier [196] consider orthographic and head-noun features, and Yamamoto et al. [224] explore a variety of features encompassing boundary, morpho-lexical, and syntactic properties as well as a dictionary-based feature that indicates whether a word appears in a biomedical resource. Given the sensitivity of classification-based approaches to the choice of features, automatic feature selection is an important consideration. Hakenberg et al. [63] perform a systematic evaluation of common features and discuss their influence on the predictive quality of classification-based NER systems.

Unlike classification-based approaches, sequence-based NER systems consider complete sequences of words instead of only individual words or phrases. They are trained on tagged corpora and aim to predict the mostly likely tags for a given sequence of observations. A common statistical framework used for biomedical NER is the Hidden Markov Model (HMM) [36, 179, 124, 93]. Methods based on the Maximum Entropy Markov Model are also common [55, 37]. However, Conditional Random Fields (CRF) [141, 175] are often demonstrated to be superior statistical frameworks for biomedical NER. For example, CRFs were utilized by the best performing system on the i2b2 medical concept extraction task [216] and by highly ranked systems on the BioCreAtIve gene mention recognition tasks [226, 187] and the JNLPBA bio-entity recognition task [91]. Like other statistical methods, sequence-based approaches can be trained on a variety of features including orthographic features [36, 124], prefix and suffix information [179], and part-of-speech tag sets augmented to include tags for entity classes [93].

Many approaches do not just utilize a single method for performing biomedical NER and instead rely on multiple techniques and various resources. These hybrid approaches are often quite successful at combining dictionary- or rule-based approaches with statistical methods. As evidence of the advantages of hybrid approaches, Abacha et al. [2] compared the performance of common rule-based and statistical approaches to medical entity recognition and concluded that hybrid approaches utilizing machine learning and domain knowledge perform best. There are numerous hybrid biomedical NER systems. For example, Sasaki et al. [173] use a dictionary-based approach to identify known protein names in parallel with part-of-speech tagging. They then use a CRF-based approach to reduce the number of false positives and false negatives in the resulting tagged sequence. Other methods create meta-learners from multiple statistical methods. For example, Zhou et al. [236] utilize a meta-learner composed of two HMMs trained on different corpora whose outputs are combined with one SVM to recognize protein and gene names. Similarly, Mika and Rost [117] compose a meta-learner to recognize protein names from three SVMs trained on different copora and feature sets whose outputs are then combined with a fourth SVM. Finally, Cai and Cheng [25] present an approach to biomedical NER that utilizes three different classifiers to improve the generalization ability of the system.

A more thorough analysis of NER approaches in the biomedical domain can be found in the several literature surveys dedicated to the subject [99, 104].

3.2 Relation Extraction

Most information extraction tasks in the biomedical domain go beyond simply identifying named entities and, in addition, involve determining relationships among those entities. In their simplest form, associations among biomedical entities are binary, involving only the pair-wise relations between two entities. However, biomedical relationships can involve more than just two entities, and these complex associations are discussed later with the event extraction task. The goal of the relation extraction task, therefore, is to identify occurrences of particular types of relationships between pairs of given entities. Although common entity classes (e.g., genes or drugs) are generally quite specific, the types of identified relationships may be broad, including any type of biomedical association, or they may be specific, for example, by characterizing only gene regulatory associations.

A variety of biomedical relations have been the subject of information extraction tasks in the literature. In the current genomic era, much of this work has focussed on automatically extracting interactions between genes and proteins. In particular, because of its critical role in understanding biological processes, Protein-Protein Interaction (PPI) has been one of the most widely researched topics in biomedical information extraction. Other associations of interest include interactions between proteins and point mutations [102], proteins and their binding sites [28], genes and diseases [31], and genes and phenotypic context [113]. In the clinical domain, relationships between patients' presented medical problems and the tests or treatments they may undergo [216] is an increasingly important type of relation, especially considering the growing prominence of electronic health record systems.

Biomedical relation extraction faces many of the same challenges as NER, including the creation of high quality annotated corpora for training and evaluating relation extraction systems. Compared with the annotation of named entities, the annotation of relations is considerably more complicated since relations are generally expressed as discontinuous spans of text and the types of relations considered are usually application-specific [13]. Additionally, since there is often little consensus regarding how to best annotate given types of relations, the resulting resources are largely incompatible, and, as a result, the quality of the methods utilizing these resources is difficult to evaluate. For example, Pyysalo et al. [155] performed a comparative analysis of five PPI corpora and found that the performance of state-of-the-art PPI extraction systems, measured in terms of F-score, varied on average by 19 percentage points and by as much as 30 percentage points on the evaluated corpora. Participation in community-wide evaluations that are dedicated to the relation extraction task is indispensable for obtaining annotated corpora.

Relation extraction tasks have been a component of several recent evaluation forums, and these tasks include the LLL genic interaction challenge [133], the BioCreAtIve PPI extraction task [96], and the i2b2 relation extraction task [216]. The purpose of the LLL challenge was to extract protein and gene relationships from abstracts contained in MEDLINE, and the best-performing system achieved an F-score of 0.54 identifying these associations. The BioCreAtIve task consisted of four subtasks related to PPI extraction. These challenges included the classification of PubMed abstracts as to whether they were relevant for PPI annotation, the identification of binary protein-protein interactions from full-text articles, the extraction of protein interaction methods, and the retrieval of textual evidence describing the interactions. The best-performing system achieved a precision of 0.37 at recall 0.33 for extracting binary PPI relations. Finally, the aim of the i2b2 relation extraction challenge was to identify medical problem-treatment, problem-test, and problem-problem relationships in clinical notes. Participants were tasked, for example, with determining whether two co-occurring problem and treatment concepts were related, and if so, whether the patient's treatment improved, worsened, or caused the medical problem. The best-performing system on the i2b2 relation extraction challenged achieved an F-score of 0.74. Like the forums dedicated to evaluating the NER task, community-wide evaluations such as these have been instrumental in the development and evolution of relation extraction approaches.

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. Some recent approaches to the relation extraction task are described below. An accounting of additional methods can be found in other biomedical text mining surveys that cover the relation extraction task [13, 32, 238].

The simplest method of identifying relations between biomedical entities is to collect instances where the entities co-occur. If the entities are repeatedly mentioned together, then there is a greater chance that they may be related in some way, although the type and direction of this relation typically cannot be determined by co-occurrence statistics alone. For example, Chen et al. [30] apply co-occurrence statistics to compute the degree of association between diseases and drugs extracted from clinical records and biomedical literature. Co-occurrence approaches commonly exhibit high recall and low precision.

Rule-based approaches describe the linguistic patterns exhibited by particular relations. Unlike the systems based on term co-occurrences, rule-based approaches typically demonstrate high precision and low recall. The rules used for relation extraction can be manually defined by domain experts [172], or they can be derived from annotated copora by machine learning algorithms [64].

Classification-based approaches are also commonly used to identify relations, particularly those involving medical entities. Roberts et al. [168] describe a supervised machine learning system, trained on shallow features extracted from oncology reports, that detects various clinical relationships in patient narratives. Similarly, Rink et al. [167] describe a system that discovers relations between medical problems, treatments, and tests mentioned in electronic medical records. The system relies on supervised machine learning and lexical, syntactic, and semantic context features. Bundschus et al. [23] utilize CRFs to identify and classify relations between diseases and treatments extracted from PubMed abstracts and relations between genes and diseases in the human GeneRIF database. Finally, Abach and Zweigenbaum [1] describe a hybrid approach that utilizes patterns developed by domain experts as well as SVM classification to extract relations that occur between diseases and treatments in medical texts.

An important advance in the evolution of relation extraction methods has been the consideration of syntactic structures. In particular, dependency parsing is capable of producing informative syntactic descriptions of biomedical text, in the form of dependency trees or graphs, which encode grammatical relations between phrases or words. Fundel et al. [60] produce dependency trees from MEDLINE abstracts. Their system then applies three relation extraction rules to the syntactic structures in order to identify gene and protein associations. Similarly, Rinaldi et al. [164] combine syntactic patterns obtained from dependency tree structures in order to support querying the biomedical literature for interactions between genes and proteins. Miyao et al. [121] perform deep parsing to annotate predicate-argument structures in MEDLINE abstracts. Their system then relies on the structural matching of the semantic annotations to identify and retrieve relational concepts. In other work, Miyao et. al [122] evaluate various parsers and their output representations on their ability to improve accuracy when used as a component of a PPI extraction system.

With the growing availability of large corpora containing relational annotations, many approaches utilize machine learning algorithms to extract useful information from syntactic structures rather than applying manually derived patterns. In the context of kernel-based machine learning, several authors have proposed kernels capable of measuring the similarity between syntactic parse trees or graphs. Airola et al. [7] describe an all-paths graph kernel for computing the similarity between dependency graphs. The kernel function is then used in training a least squares support vector machine to identify protein-protein interactions. Kim et al. [92] suggest four genic relation extraction kernels defined on the shortest syntactic dependency path between two named entities. Finally, Miwa et al. [120] describe a framework for combining the outputs of multiple kernels and syntactic parsers to extract protein-protein interactions.

Syntactic analysis is often complemented by semantic role labeling, a natural language processing technique that identifies the semantic roles of words or phrases in sentences and expresses them as predicate-argument structures. Tsai et al. [202] construct a role labeling system that uses a maximum entropy machine learning model to extract biomedical relations from a prepared portion of the GENIA corpus. As discussed below, the annotation of semantic roles for named biomedical entities has enabled the extraction of a variety of complex entity associations.

3.3 Event Extraction

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 [13].

Another important distinction between the extraction of binary relations and complex events is that events can be nested, with one event functioning as a participant of another event. For example, in the sentence "RFLAT-1 activates RANTES gene expression" two events are present [13]. One event is indicated by the nominalized verb *expression* whose theme is *RANTES*, a gene, and the other event is indicated by the verb *activates* whose cause is *RFLAT-1*, a protein, and whose theme is the gene expression event itself. Thus, event representations, unlike binary relations, are capable of capturing many different types of associations with an arbitrary number of entities and events related by a variety of semantic roles.

Due to the complexity of biomedical events, effective event extraction typically requires a thorough analysis of sentence structure. Event extraction is particularly aided by the use of semantic processing and deep parsing techniques, which are capable of analyzing both the syntactic and semantic structure of biomedical text. Dependency parsing is an especially useful technique for capturing semantics such as predicate-argument relationships, which have been shown to be an effective representation for event extraction [219]. Despite the complexity of the task, event extraction has broad applicability in the biomedical domain, and it is increasingly being used for the annotation of biomedical pathways, Gene Ontology annotation, and the enrichment of biological databases.

The growing interest in event extraction has largely been driven by the introduction, mostly in the domain of systems biology, of corpora containing the annotations necessary for the training and evaluation of statistical event extraction methods. The BioInfer corpus [156] was the first publicly available corpus in the biomedical domain to incorporate event annotations. Other annotated event corpora include the GENIA Event Corpus [92] and the Gene Regulation Event Corpus [198]. Notably, the GENIA corpus remains one of the most widely used resources in biomedical text mining, and the data for the BioNLP shared tasks on event extraction [89, 88] were prepared based on this resource.

The BioNLP '09 shared task [89] was the first-of-its-kind community-wide evaluation of event extraction methods. The primary challenge was to extract event types related to protein biology from MEDLINE abstracts. Targeted event types included, among others, gene expression, transcription, localization, binging, and regulation. The binding event type was more complex than the others since it required the detection of an arbitrary number of arguments, and the regulation event types were notable for allowing other events to act as their cause or theme. The best-performing system obtained an F-score of 0.52 on the primary event extraction task. The BioNLP '11 shared task [88] repeated the evaluation from the previous meeting, but also included additional tasks targeting event types in other subdomains of biology. On the subtask comparable with that of that of the first meeting, the best-performing system achieved an F-score of 0.57, which demonstrated a significant improvement in the community. Successful systems at the BioNLP shared task meetings relied on a variety of techniques including machine learning, Markov logic networks, and dependency parsing. Several approaches to biomedical event extraction are described below.

Most event extraction systems follow a pipelined approach that divides the task into a sequence of three stages. Fist, the systems predict a candidate set of event trigger words. Trigger words are often the verbs or nominalized verbs that indicate a particular event type, such as "phosphorylation," "activates," or "inhibits." Then, the systems seek to determine whether any recognized named entities or trigger words are instantiations of event arguments. The final stage in the process is a semantic post-processing step that attaches arguments to event triggers following constraints on the type and number arguments allowable for a given event type.

This basic architecture is a common approach to the event extraction task. Björne et al. [21] describe the best-performing system on the BioNLP '09 event extraction task. Their method trains separate multi-class SVMs for detecting event triggers and arguments using an extensive set of features, especially those derived from dependency parse graphs. Their system then uses a rule-based approach for attaching arguments to their corresponding events. This approach has been combined with BANNER to perform event extraction on an unlabeled subset of citations from PubMed [20]. Miwa et al. [119] describe an event extraction approach similar to that of Björne et al., but instead of relying on a rule-based approach to attach event participants to trigger words, they obtain an improvement by utilizing a classifier and additional features for this step. Buyko et al. [24] describe a system that relies on a dictionary-based approach to identify event triggers and an ensemble of feature- and kernel-based classifiers trained using “trimmed” dependency graphs to identify event participants. Kilicoglu and Bergler [87] also use a dictionary-based approach to identify event riggers, but they develop rules based on syntactic dependency paths to detect event participants. Finally, Cohen et al. [34] describe a pattern-based approach to event extraction that utilizes the OpenDMAP system [79] to define entity and event types as well as the constraints on event arguments.

Recently, joint prediction approaches have been proposed that seek to overcome the problem of cascading errors, which some of the above approaches allow. For example, by separating the event trigger and argument detection tasks, a system may not correctly extract an event if it fails to detect a trigger word in the first stage of the process. Poon and Vanderwende [153] propose a method based on Markov logic networks that jointly predicts events and arguments. For each word, the system predicts whether it is an event trigger word, and for each syntactic dependency edge, the system predicts whether it is an argument path leading to an event theme or cause. Additionally, Riedel and McCallum [161] propose a family of three joint prediction models based on Markov logic that are less computationally complex than previous work [160] and lead to better event extraction results.

4. Summarization

Information extraction techniques are often utilized as a first step in other biomedical text mining tasks. One such task is the automatic summarization of biomedical documents. *Automatic summarization* refers to the process by which the salient aspects of one or more documents is identified and presented succinctly and coherently. Due to the enormous growth of unstructured information in the form of scientific articles and electronic health records, a means for clinicians and researches to quickly and reliably assimilate knowledge from a multitude of biomedical sources is desirable. Automatic summarization is one approach to determine and make accessible the important information contained in an increasingly large and diverse volume of biomedical text.

In the biomedical domain, document summaries are commonly application-oriented, and can serve a variety of purposes. Summaries may be either a generic assimilation of facts or they may be targeted [3]. Generic summaries consider all the information contained in a document or set of documents while targeted summaries aim to satisfy a specific information need, which is usually presented to a system in the form of a query. For example, a targeted summary of the biomedical literature might seek to determine the best treatments for a given disease [56], whereas a generic summary might aim to extract from articles key sentences related to results or conclusions [169]. Additionally, a summary is considered indicative if its purpose is to inform a reader of the contents of a document or set of documents, or it is informative if its purpose is to supplant those contents in terms of information coverage [3].

Depending on their purpose, several different types of document summaries can be produced. Single-document summaries seek to summarize the contents of individual sources, whereas multi-document summaries consider the information contained in a collection of sources [3]. Often, document clustering is utilized when generating multi-document summaries in order to produce a topical account of a particular group of documents. Summaries may also be extractive or abstractive [3]. Extractive summaries are created by identifying the salient textual components of documents (e.g., their important sentences or paragraphs) and then presenting this information as the summary. The representative textual components are determined by statistical methods that rank them according to relevance or by graph-based methods that organize them according to their similarity. Alternatively, abstractive summaries are created by structuring document information in a way that can be processed by a natural language generation system to produce the summary. Salient information is typically generated through prior knowledge of the documents' structure or by utilizing ontological resources to produce semantic representations of the documents.

Considering both the various types of summaries that may be generated and their intended applications, the evaluation of summarization techniques within the biomedical domain is a challenging issue. This difficulty is due, in part, to the subjective aspect of determining whether a summary is of "good" quality or not. Existing evaluation criteria consider the intrinsic aspects of a summary, such as its coherence, conciseness, grammaticality, and readability. Other extrinsic evaluation criteria measure, for example, whether a reader is able to comprehend the content of a summary [3]. However, manual evaluations of summaries are time-consuming and expensive to perform. A popular automatic summary evaluation methodology is ROUGE [107]. ROUGE is an acronym for Recall-Oriented Understudy for Gisting Evaluation, and it determines the quality of an automatically generated summary by computing statistics based on n-gram co-occurrences and common subsequences between it and ideal human-produced summaries. ROUGE has been shown to correlate well with human evaluations of single-document summaries.

A related method is based on the Jensen-Shannon divergence of distributions between an automatically generated summary and reference summaries and is more effective for the multiple document summarization task [108].

Recent biomedical text summarization techniques have been shown to be effective tools for assimilating information from a diverse collection of sources. While most approaches in the biomedical domain aim to produce targeted or topic-specific summaries, the types of generated summaries are generally more diverse and include both single- and multi-document summaries as well as extractive and abstractive summaries. However, given the rapidly expanding volume of published biomedical literature, multi-document summaries are increasingly viewed as important. Examples of recent text summarization approaches and their applications are described below.

One of the most basic approaches to biomedical text summarization involves the classification of individual sentences into a given set of categories. These categories may be specific to the biomedical domain, but they are often representative of the general rhetorical categories commonly encountered in scientific literature. Agarwal and Yu [4] train a Naïve Bayes classifier to classify sentences in full-text biomedical articles as being related to the introduction, methods, results, and discussion rhetorical categories. Their system achieves an overall annotation agreement of 0.76 kappa with human annotators. Ruch et al. [169] describe a similar approach that classifies sentences in MEDLINE abstracts as being related to an article's purpose, methods, results, or conclusions. Finally, Demner-Fushman and Lin [45] produce extractive summaries for clinical information needs by extracting sentences from MEDLINE abstracts relating to the outcomes of a clinical study.

While some of the above approaches apply generic summarization methods to biomedical articles, most applications are targeted and seek a concise description of a specific type of information. Since the understanding of gene regulation and expression is crucial in current biomedical research, a variety of targeted methods have been proposed to generate multi-document gene summaries. Ling et al. [112] propose a method for generating abstractive multi-document gene summaries from biomedical literature. Their two-stage approach to gene summarization first retrieves articles that mention a particular gene, and it then identifies text within those articles that pertains to several gene-related semantic categories, which include expression, sequence, and phenotypic information. Similarly, Yang et al. [225] describe an extractive approach to gene summarization that first clusters genes into functional groups based on their mentions in MEDLINE abstracts. Their system then presents summaries for each functional group by ranking and extracting sentences from the abstracts.

A challenge facing many automatic summarization techniques is the accurate semantic interpretation of the text. To address this issue, several summarization methods utilize domain knowledge in order to produce ontology-based document summaries. Reeve et al. [158] describes a single-document abstractive approach that utilizes MetaMap to map text to concepts in the UMLS Metathesaurus. Their approach then discovers strong thematic chains of UMLS semantic types and extracts the corresponding sentences. Yoo et al. [229] describe an approach to multi-document summarization that first clusters articles into topical groups and then produces summaries for each cluster. Their system uses a graph-based method for both document clustering and summarization that is enriched with concepts from the MeSH ontology. Morales et al. [123] describe a similar graph-based approach to single-document summarization that represents documents using UMLS concepts. Finally, Fiszman et al. [56] utilize SemRep [165] to produce multi-document summaries of MEDLINE citations according to disease-treatment relationships relevant to user-specified topics. Their approach has become an integral component of Semantic MEDLINE [166].

In addition to the text found in biomedical articles, the figures they contain also convey essential information. However biomedical images are seldom self-evident, and much of the information required for their comprehension is found elsewhere in an article. Figure captions, article titles and abstracts, and snippets of text from within the bodies of articles all contribute to image understanding [230]. Given that figures are a crucial source of information in the biomedical literature, many methods seek to incorporate image-related text into document summaries. However, since the number of such approaches is so large, and their methods are diverse, a full accounting of the use of image-related text in bioinformatics warrants a separate review.

A few representative examples of figure summarization and the use figure captions for producing document summaries include the following. Similar to their approach for full-text summarization, Agarwal and Yu [5] produce figure summaries consisting of one sentence each from an article's introduction, methods, results, and discussion rhetorical categories. Yu and Lee [232] produce figure summaries by extracting sentences from article abstracts that are similar to figure captions, and Simpson et al. [181] utilize image-related text to produce full-text summaries in support of case-based article retrieval.

Several user-oriented systems have been developed for supporting biomedical document summarization. PERSIVAL [116, 49] is a clinical system that seeks to provide access to medical literature and consumer health information. For clinicians, the system produces targeted, multi-document summaries containing sentences, extracted from full-text biomedical articles, that relate to experimental results. For users of the system that are patients, PERSIVAL provides indicative summaries of information that is commonly repeated across a set of consumer health documents. Anne O'Tate [184] is another user-oriented system capable of producing summaries of biomedical literature. Anne O'Tate is a web-based tool that provides navigable, extractive multi-document summaries of article citations retrieved by PubMed. The tool presents import words and authors mentioned in the results and can cluster the retrieved citations by topic.

5. Question Answering

Another biomedical text mining task that builds upon information extraction techniques is question answering. Unlike traditional information retrieval, where a set of potentially relevant documents is returned for a given query, *question answering* refers to the process of providing direct and precise answers to natural language questions. Like automatic summarization, question answering is a task directed towards aiding researchers and health care professionals in managing the continuous growth of information in the biomedical domain. Since question answering requires the use of complex natural language processing techniques in order to produce accurate responses, question answering systems are often regarded as the next generation of search engines.

The basic processing steps required of a question answering system are well-understood. The input to such a system is natural language text. A question processing stage uses linguistic analysis and question classification techniques to determine the type of question being posed to the system and the type of response it should generate. It then constructs a query from the input text to be fed into a document processing stage. In the document processing stage, the system inputs the query into a search engine, which retrieves a set of documents, and from these documents, extracts relevant passages or snippets of text as potential answers. An answer processing stage ranks the candidate answers according to the degree to which they match the expected answer type that was determined in the question processing stage. The output of a question answering system is the top-ranked answer.

Several characteristics of this process distinguish question answering in the biomedical domain from general, open-domain question answering systems. First, biomedical question answering is both challenged and advantaged by a prominent use of domain-specific terminology. Although terminological variations and synonymy make text mining difficult in general for the biomedical domain (Section 3), question answering systems may benefit from the specificity and limited scope of potential questions that a domain-specific terminology provides. Second, the multitude of domain-specific corpora and the tools and methods required for exploiting the semantic information they contain (Section 2) allow for deep question processing. Lastly, agreement on domain-specific structures in which to organize questions—especially clinical questions—allows for answer processing strategies that can be tailored to specific question types.

Due to the unique characteristics of biomedicine as an application domain for question answering, recently proposed systems have increasingly sought to incorporate deep semantic knowledge throughout their processing stages in order to produce more precise responses. The remainder of the discussion in this section surveys biomedical question answering techniques, and organizes the methods according to the recent review by Athenikos and Han [17], in which the authors classify biomedical "semantic knowledge-based" systems into semantics-based, inference-based, and logic-based approaches. Semantics-based approaches produce answers to biomedical questions by exploiting the semantic metadata encoded in structured knowledge resources and ontologies; inference-based approaches derive responses by exploiting extracted semantic relationships, and logic-based approaches utilize explicit logical forms and theorem proving techniques to produce answers. The approaches can further be divided into those that support medical question answering and those that support biological question answering.

5.1 Medical Question Answering

A dominant theme of work related to medical (or clinical) question answering is the use of the evidence-based medicine framework. Evidence-based medicine [170] seeks to apply the best information garnered from scientific inquiry to clinical decision making. For determining the best available evidence supporting an answer to a given clinical question, the evidence-based paradigm suggests questions be structured according to the PICO [159] format. PICO is an acronym for Patient/Problem, Intervention, Comparison, and Outcome. Clinical questions containing elements that pertain to each of these semantic roles are considered well-formed. In addition to the structure of clinical questions, taxonomies of questions in the evidence-based framework have also been proposed. Ely et al. [50] describe a generic taxonomy for clinical questions that distinguishes among questions that are potentially answerable and those that are not. The authors claim that questions involving a search for evidence are among the answerable ones.

The first step towards answering a clinical question is processing the question so as to determine the type of answer to produce. Several authors in the medical domain have investigated question classification as a means of analyzing and filtering clinical questions. Huang et al. [77] describe a manual classification of primary care clinical questions as a means to evaluate the effectives of the PICO framework. The authors conclude that PICO is a useful organizing structure for clinical questions, but they suggest it is less suitable for questions that do not involve therapy elements. Additionally, Yu et al. [234, 235, 231] investigate various machine learning approaches for question filtering that automatically determine whether a clinical question is answerable according to the evidence taxonomy proposed by Ely et al., which was described above.

Most approaches to medical question answering in some way make use of domain-specific semantic knowledge for information extraction and retrieval. Jacqumart et al. [40, 81] describe a semantics-based approach for the development of a French-language medical question answering system. Their approach is notable for the use of pattern-based semantic models of medical questions and the use of UMLS concepts, semantic types and relations for identifying named entities and extracting answers. Niu et al. [135, 136, 138, 137] propose a PICO-based question answering approach within the EPoCare system. Their methods locate potential answers by identifying, in both the question and answer texts, semantic roles that correspond to the four elements of the PICO framework. The semantic roles identified in the question are then compared with those identified in candidate answers to select a response. Similarly, Demner-Fushman et al [45, 44, 42, 109, 43] propose an approach to clinical question answering based on the semantic unification of a query PICO frame with those of candidate answers. Making extensive us of MetaMap and SemRep, the authors describe semantic knowledge extractors for identifying PICO elements in medical texts, a semantic matcher for scoring and ranking MEDLINE citations according to a query PICO frame, and an answer generator for extracting answers from the scored citations. Weiming et al. [221] describe a question answering approach that represents questions and documents using UMLS concepts, semantic types, and semantic relations. Their approach is notable for incorporating a semantic clustering phase into the answer processing stage so as to organize potential answers according to their hierarchical relationships in the UMLS Metathesaurus. Finally AskHERMES [27] is an online clinical question answering system capable of processing long and complex questions. The system uses machine learning techniques with a variety of lexical, syntactic, and UMLS-derived features to classify questions and topically group and rank candidate answers. A preliminary version of AskHERMES, known as the MedQA [233] system, was a non-semantic-knowledge-based approach capable of answering definitional questions.

Few approaches to question answering in the medical domain are inference- or logic-based. Terol et al. [197] describe an approach based on comparing the formal logic forms derived from a natural language question with those of candidate answers. Their technique utilizes a pattern-based method for question classification, and it identifies medical entities in both questions and answers based on UMLS concepts and semantic types.

5.2 Biological Question Answering

Whereas evidence-based medicine provides a means to structure clinical questions and answers, work in the biological domain has yet to adopt such a prominent framework. However, systems targeting the biological domain still follow the general architecture of questioned answering systems outlined previously. A review of recent work related to biological question answering is presented below.

Like their use for medical question answering, semantics-based approaches are also commonly employed for answering questions in the biological domain. Takahashi et al. [195] describe an approach that utilizes the UMLS Metathesaurus and other biological dictionaries and thesauri for analyzing questions and generating queries. Their system then uses semantic information of terms selected from the retrieved documents to assimilate and rank candidate answers. Lin et al. [110] propose a system for answering questions about biomolecular events, including interactions between genes and proteins (Section 3). Their approach involves the use of semantic role labeling for extracting predicate-argument structures and the use of semantic features for ranking candidate answers. The system provides answer responses in the form of biomedical named entities. Finally, the BioSquash [180] system is a targeted, multi-document, semantic graph-based summarization system oriented towards answering biological questions.

Like the use inference- and logic-based methods for medical question answering, few approaches in the biological domain make use of these techniques. Kontus et al. [94, 95] describe the AROMA inference-based system for biological question answering. AROMA extracts rhetorical and causal relationships from multiple biological texts, combines the extracted text with manually entered domain knowledge, and encodes this information as Prolog facts. The system generates answers to questions by applying inference rules over the encoded facts. Rinaldi et al. [162] describe a logic-based approach to question answering in the genomics domain. Using deep linguistic and terminological information, the system derives logical forms for text taken from documents in the GENIA corpus and a subset of full-text documents indexed in MEDLINE. Natural language questions are processed with the same mechanism, and the system derives an answer using a theorem proving process.

6. Literature-Based Discovery

While the extraction of explicit relations and events among biomedical entities can be used to produce rich document summaries and enable complex question answering systems, an exciting use of these methods aims to uncover relationships that are not present in the text, but that can be inferred from other information. *Literature-based discovery* refers to the task of utilizing scientific literature to uncover "hidden," previously unknown or neglected relationships between existing knowledge. The goal of discovering these implicit relationships is to identify relations worthy of further scientific investigation or to find evidence supporting suspected relations.

As a technique useful for biomedical text mining, literature-based discovery was pioneered by the work of Swanson in the 1980s. Swanson suggested that novel information could be uncovered by systematically reviewing "complementary but disjoint" bodies of literature [192]. In what has become the prototypical example of literature-based discovery, Swanson linked fish oil, a substance widely-understood to have potential cardiovascular benefits, with Raynaud's syndrome, a vasospastic disorder causing the narrowing of blood vessels [189]. The discovery suggests fish oil supplements may help to control the symptoms of Raynaud's syndrome. To further demonstrate the feasibility of his ideas, Swanson later found evidence for relationships between migraine and magnesium [190], somatomedin C and arginine [191], and viruses and their potential use as biological weapons [194].

The basic premise of Swanson's approach is that there exists two scientific communities that do not communicate. A portion of the knowledge in one community may be related to or complement knowledge in the other one, but this relationship is unknown to either community. For example, suppose a scientific community has researched the relationship between a medical finding or characteristic B and a disease C. Further suppose that a separate community has studied the affects of substance A on characteristic B. The use of literature-based discovery techniques may suggest an A-C relationship, indicating in this example that substance A may potentially treat disease C.

Weeber et al. [220] distinguish between two modes of discovery. A "closed discovery," or hypothesis testing study, begins with known A- and C-terms. Thus, the discovery concerns finding novel B-terms that may explain the observed A-C association or hypothesis. On the other hand, an "open discovery," or hypothesis generation study, begins with known A-B associations in one domain and seeks to discover B-C relations in another domain, thereby suggesting or generating a potential A-C association.

Since the pioneering work of Swanson, literature-based discovery techniques have seen widespread use. Existing approaches can be grouped by the way in which they identify potentially novel relationships. There are those that depend exclusively on the co-occurrence of terms or concepts, those that make use of semantic information to inform the processing of co-occurring terms, and those that construct interaction networks of individual relations whose paths can reveal hidden associations. Some recent methods following these general approaches are reviewed below. Unlike other text mining tasks, measuring the performance of literature-based discovery tools is not straightforward, and a discussion of system evaluation follows as well.

Co-occurrence-based methods are among the simplest, although less precise, approaches to literature-based discovery. Like the most basic approaches to the relation extraction task (Section 3), these methods seek to identify terms that frequently occur together. However, whereas approaches to relation extraction identify first-order term co-occurrences, approaches to literature-based discovery explore second-order co-occurrences—the shared co-occurrences of two given biomedical entities [238].

Most of the earliest approaches to literature-based discovery and many modern approaches rely on entity co-occurrence statistics. The Arrowsmith [193, 185, 182, 183, 199, 186] two-node search tool implements Swanson's original approach to find biologically meaningful links between two sets of articles in PubMed using title words and phrases. Recent work related to this project has developed a method to estimate and rank the relevance of associations. BITOLA [74, 73, 72, 75] is a similar literature-based discovery system, but instead of identifying relations using title words, it represents documents using their MeSH terms and recognized gene symbols. Additionally, BITOLA uses association rules [6] as a measure of concept relatedness instead of word frequencies. LitLinker [227] also utilizes MeSH terms; however, it uses a statistical approach based on the background distribution of term probabilities to identify correlated concepts. Jelier et al. [82] describe a system that identifies functional associations between genes and other biomedical concepts. Their approach measures the strength of association of co-occurring concepts using a log likelihood ratio. RaJoLink [151] provides semi-automated suggestions for links between two sets of articles based on rare terms identified in the literature. FACTA+ [203] uses an information theoretic score to rank indirectly associated concepts. It identifies explicit associations among biomedical entities using methods inherited from an earlier version of the system [206]. Finally, unlike other literature-based discovery methods that rely on associations explicit in scientific literature, Benton et al. [18] use a corpus of posts to Internet breast cancer message boards to discover adverse drug effects.

Because systems relying solely on co-occurrence statistics tend to produce a large number of spurious relations, recent approaches increasingly rely on semantic information to identify hidden relations or augment the processing of co-occurring entities. Hristovski et al. [71] describe an improvement to BITOLA that uses the semantic predications produced by SemRep and BioMedLEE [114] to enable users to eliminate uninteresting or incorrect relations. A similar approach is used in the EpiphaNet system [35], an interactive visualisation tool for exploring associations between concepts found in MEDLINE. EpiphaNet makes extensive use of MetaMap and SemRep for identifying explicit relations. Other systems, including Weeber et al.'s DAD-system [220], filter candidate relations based on the UMLS semantic type of identified B-terms. Recall that for hypothesis generation, B-terms are used to uncover hidden A-C relations from explicit A-B and B-C associations. Hu et al. [76] describe a literature-based discovery method that uses association rules as a measure of concept relatedness but also filters potential relations using UMLS semantic types.

Another approach to discovering hidden relationships among biomedical entities involves the construction of interaction networks whose paths can reveal indirect associations. Seki and Mostafa [174] build an inference network [208] to predict implicit gene-disease associations. Genes and diseases are connected within the graph by intermediary nodes representing gene functions and phenotypes. Similarly, Özgür et al. [145, 146] build a gene-interaction network by collecting an initial set of known disease-related genes from biomedical texts using dependency parsing and SVMs. They then use network centrality metrics to predict gene-disease associations. Finally, Palakal et al. [149] describe BioMap, a directed graph that is constructed from explicit relationships between biomedical entities identified within text. Users are able to query the graph to uncover implicit associations among the entities.

Due to the nature of uncovering novel information, there is no ground truth available for evaluating literature-based discovery systems, and comparing the relative performance of alternative approaches is difficult. A common method for evaluating an automatic discovery technique is to use the system to replicate known discoveries, such as Swanson's linking of Raynaud's syndrome with fish oil or migraine with magnesium [220]. However, Yetisgen-Yildiz and Pratt [228] suggest this approach is uninformative of the overall performance of a system. They describe an alternative methodology that divides the abstracts in MEDLINE into two sets: those that were published before a given cut-off date, and those that were published after this date. Literature-based discovery methods are then applied to the older set of abstracts as hypotheses generating systems and to the newer set as hypotheses testing systems, using the generated associations from the older set as input. The performance of a system can then be quantified using standard information retrieval evaluation methods.

7. Conclusion

The past several years have seen some exciting developments in biomedical text mining. Progress was made in (1) defining and attempting more challenging tasks, such as event extraction and clinical text mining; (2) increasing the public availability of and community investment in resources, such as the MIMIC II database and the ORBIT registry; and (3) development and use of common frameworks, such as UIMA.

It is interesting to compare the development of the field to the desirable directions outlined by the leading researchers in 2008 [9]. At that time, the researchers were asked about the importance of text mining for biology, the utility of the text mining systems, and future directions.

The first suggested avenue for future research was *fusing literature and biological databases through text mining*. Understandably, this requires engaging the publishers of scientific literature and realizing potentially additional efforts by the publications' authors. To that end, Elsevier is piloting a tool, Reflect-Network [147], developed in partnership with the European Molecular Biology Laboratory and the Novo Nordisk Foundation Center for Protein Research. Reflect tags proteins and chemicals in documents and generates a graphical representation displaying interactions between entities and additional details about them.

The second proposed research direction was *interactivity and user interfaces*. This direction requires identifying more potential user groups and tasks. Progress was made in developing tools for database curation [223, 150]; however, more research is still needed in identifying user groups and tasks in parallel with tool development for known users.

The authors noted that success in the third direction, *tool scalability and integration into workflows*, depends on commonly accepted and used stable standards for the exchange and integration of information derived from text mining. Despite major initiatives towards seamless data exchange and interoperability (e.g., the i2b2 hive [80] or the eMERGE Network [51]) and pilot applications being included into workflows (e.g., NLM InfoBot [46]), this direction remains challenging. The efforts needed to make a system scalable and capable of handling real-time workflow interactions were recently demonstrated in the IBM DeepQA project [53].

The last direction, *development of text mining resources*, is an ongoing activity. Existing lexicons, standards, and ontologies are maintained—and new resources and community-wide evaluations emerge—following the progress in biology and medicine.

Acknowledgements

M. S. S. was supported in part by an appointment to the NLM Research Participation Program sponsored by the U. S. National Library of Medicine and administered by the Oak Ridge Institute for Science and Education.

References

[1] A. B. Abacha and P. Zweigenbaum. A hybrid approach for the extraction of semantic relations from MEDLINE abstracts. In A. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 6609 of *Lecture Notes in Computer Science*, pages 139–150. Springer Berlin / Heidelberg, 2011.

[2] A. B. Abacha and P. Zweigenbaum. Medical entity recognition: A comparison of semantic and statistical methods. In *Proceedings of BioNLP 2011 Workshop*, pages 56–64, 2011.

[3] S. Afantenos, V. Karkaletsis, and P. Stamatopoulos. Summarization from medical documents: A survey. *Artificial Intelligence in Medicine*, 33(2):157–177, 2005.

[4] S. Agarwal and H. Yu. Automatically classifying sentences in full-text biomedical articles into introduction, methods, results and discussion. *Bioinformatics*, 25(23):3174–3180, 2009.

[5] S. Agarwal and H. Yu. FigSum: Automatically generating structured text summaries for figures in biomedical literature. In *AMIA*

Annual Symposium Proceedings, pages 6–10, 2009.

[6] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo. Fast discovery of association rules. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 307–328. American Association for Artificial Intelligence, 1996.

[7] A. Airola, S. Pyysalo, J. Bjorne, T. Pahikkala, F. Ginter, and T. Salakoski. All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. *BMC Bioinformatics*, 9(Suppl 11):S2, 2008.

[8] B. Alex, B. Haddow, and C. Grover. Recognising nested named entities in biomedical text. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 65–72, 2007.

[9] R. B. Altman, C. M. Bergman, J. Blake, C. Blaschke, A. Cohen, F. Gannon, L. Grivell, U. Hahn, W. Hersh, L. Hirschman, L. J. Jensen, M. Krallinger, B. Mons, S. I. O'Donoghue, M. C. Peitsch, D. Rebholz-Schuhmann, H. Shatkay, and A. Valencia. Text mining for biology - the way forward: opinions from leading scientists. *Genome Biology*, 9(Suppl 2):S7, 2008.

[10] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

[11] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.

[12] S. Ananiadou and J. Mcnaught. *Text Mining for Biology And Biomedicine*. Artech House, Inc., 2005.

[13] S. Ananiadou, S. Pyysalo, J. Tsujii, and D. B. Kell. Event extraction for systems biology by text mining the literature. *Trends in Biotechnology*, 28(7):381–390, 2010.

[14] A. R. Aronson and F.-M. Lang. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010.

[15] R. Artstein and M. Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008.

[16] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cheryy, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis,

J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: Tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

[17] S. J. Athenikos and H. Han. Biomedical question answering: A survey. *Computer Methods and Programs in Biomedicine*, 99(1):1–24, 2010.

[18] B. Benton, L. Ungar, S. Hill, S. Hennessy, J. Mao, A. Chung, C. E. Leonard, and J. H. Holmes. Identifying potential adverse effects using the web: A new approach to medical hypothesis generation. In Press, 2011.

[19] BioNLP. `http://www.bionlp.org/`.

[20] J. Björne, F. Ginter, S. Pyysalo, J. Tsujii, and T. Salakoski. Complex event extraction at PubMed scale. *Bioinformatics*, 26(12):i382–i390, 2010.

[21] J. Björne, J. Heimonen, F. Ginter, A. Airola, T. Pahikkala, and T. Salakoski. Extracting complex biological events with rich graph-based feature sets. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, pages 10–18, 2009.

[22] K. W. Boyack, D. Newman, R. J. Duhon, R. Klavans, M. Patek, J. R. Biberstine, B. Schijvenaars, A. Skupin, N. Ma, and K. Borner. Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches. *PLoS ONE*, 6(3):e18029, 2011.

[23] M. Bundschus, M. Dejori, M. Stetter, V. Tresp, and H.-P. Kriegel. Extraction of semantic biomedical relations from text using conditional random fields. *BMC Bioinformatics*, 9(1):207, 2008.

[24] E. Buyko, E. Faessler, J. Wermter, and U. Hahn. Event extraction from trimmed dependency graphs. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, pages 19–27, 2009.

[25] Y. Cai and X. Cheng. Biomedical named entity recognition with tri-training learning. In *Proceedings of the 2009 2nd International Conference on Biomedical Engineering and Informatics*, pages 1–5, 2009.

[26] CALBC challenge. `http://www.calbc.eu/`.

[27] Y. Cao, F. Liu, P. Simpson, L. Antieau, A. Bennett, J. J. Cimino, J. Ely, and H. Yu. AskHERMES: An online question answering system for complex clinical questions. *Journal of Biomedical Informatics*, 44(2):277–288, 2011.

[28] D. T.-H. Chang, Y.-Z. Weng, J.-H. Lin, M.-J. Hwang, and Y.-J. Oyang. Protemot: Prediction of protein binding sites with automatically extracted geometrical templates. *Nucleic Acids Research*, 34(suppl 2):W303–W309, 2006.

[29] W. W. Chapman and K. B. Cohen. Current issues in biomedical text mining and natural language processing. *Journal of Biomedical Informatics*, 42(5):757–759, 2009.

[30] E. S. Chen, G. Hripcsak, H. Xu, M. Markatou, and C. Friedman. Automated acquisition of disease-drug knowledge from biomedical and clinical documents: An initial study. *Journal of the American Medical Informatics Association*, 15(1):87–98, 2008.

[31] H. W. Chun, Y. Tsuruoka, J. D. Kim, R. Shiba, N. Nagata, T. Hishiki, and J. Tsujii. Extraction of gene-disease relations from MEDLINE using domain dictionaries and machine learning. In *Pacific Symposium on Biocomputing*, pages 4–15, 2006.

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

[33] K. B. Cohen and L. Hunter. Getting started in text mining. *PLoS Computational Biology*, 4(1):e20, 2008.

[34] K. B. Cohen, K. Verspoor, H. L. Johnson, C. Roeder, P. V. Ogren, W. A. Baumgartner, Jr., E. White, H. Tipney, and L. Hunter. High-precision biological event extraction with a concept recognizer. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, pages 50–58, 2009.

[35] T. Cohen, G. K. Whitfield, R. W. Schvaneveldt, K. Mukund, and T. Rindflesch. EpiphaNet: An interactive tool to support biomedical discoveries. *Journal of Biomedical Discovery and Collaboration*, 5:21–49, 2010.

[36] N. Collier, C. Nobata, and J.-i. Tsujii. Extracting the names of genes and gene products with a hidden Markov model. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 1*, pages 201–207, 2000.

[37] P. Corbett and A. Copestake. Cascaded classifiers for confidence-based chemical named entity recognition. *BMC Bioinformatics*, 9(Suppl 11):S4, 2008.

[38] CRAFT: The colorado richly annotated full text corpus. `http://bionlp-corpora.sourceforge.net/CRAFT/index.shtml`.

[39] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, I. Roberts, G. Gorrell, A. Funk, A. Roberts, D. Daml-

janovic, T. Heitz, M. A. Greenwood, H. Saggion, J. Petrak, Y. Li, and W. Peters. *Text Processing with GATE (Version 6)*. GATE, 2011.

[40] T. Delbecque, P. Jacquemart, and P. Zweigenbaum. Indexing UMLS semantic types for medical question-answering. In R. Engelbrecht, A. Geissbuhler, C. Lovis, and G. Mihalas, editors, *Connecting Medical Informatics and Bio-Informatics: Proceedings of MIE2005 - The XIXth International Congress of the European Federation for Medical Informatics*, pages 805–810. IOS Press, 2005.

[41] D. Demner-Fushman, W. W. Chapman, and C. J. McDonald. What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, 42(5):760–772, 2009.

[42] D. Demner-Fushman, B. Few, S. E. Hauser, and G. Thoma. Automatically identifying health outcome information in MEDLINE records. *Journal of the American Medical Informatics Association*, 13(1):52–60, 2006.

[43] D. Demner-Fushman and J. Lin. Knowledge exraction for clinical question answering: Preliminary results. In *Proceedings of the AAAI 2005 Workshop on Question Ansering in Restricted Domains*, 2005.

[44] D. Demner-Fushman and J. Lin. Answer extraction, semantic clustering, and extractive summarization for clinical question answering. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 841–848, 2006.

[45] D. Demner-Fushman and J. Lin. Answering clinical questions with knowledge-based and statistical techniques. *Computational Linguistics*, 33(1):63–103, 2007.

[46] D. Demner-Fushman, C. Seckman, C. Fisher, S. E. Hauser, J. Clayton, and G. R. . Thoma. A prototype system to support evidence-based practice. In *AMIA Annual Symposium Proceedings*, pages 151–155, 2008.

[47] S. Dipper, M. Götze, and M. Stede. Simple annotation tools for complex annotation tasks: An evaluation. In *Proceedings of the LREC Workshop on XML-Based Richly Annotated Corpora*, pages 54–62, 2004.

[48] eHOST: The extensible human oracle suite of tools. `http://code.google.com/p/ehost/`.

[49] N. Elhadad, M.-Y. Kan, J. L. Klavans, and K. R. McKeown. Customization in a unified framework for summarizing medical literature. *Artificial Intelligence in Medicine*, 33(2):179–198, 2005.

[50] J. W. Ely, J. A. Osheroff, M. H. Ebell, M. L. Chambliss, D. C. Vinson, J. J. Stevermer, and E. A. Pifer. Obstacles to answering doctors' questions about patient care with evidence: qualitative study. *British Medical Journal*, 324(7339):710, 2002.

[51] Electronic medical records and genomics. `https://www.mc.vanderbilt.edu/victr/dcc/projects/acc/index.php/Main_Page`.

[52] European bioinformatics institute. `http://www.ebi.ac.uk/`.

[53] D. Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. A. Kalyanpur, A. Lally, J. W. Murdock, E. Nyberg, J. Prager, N. Schlaefer, and C. Welty. Building Watson: An overview of the DeepQA project. *AI Magazine*, 31(3):59–79, 2010.

[54] D. Ferrucci and A. Lally. UIMA: An architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3-4):327–348, 2004.

[55] J. Finkel, S. Dingare, H. Nguyen, M. Nissim, C. Manning, and G. Sinclair. Exploiting context for biomedical entity recognition: From syntax to the web. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pages 88–91, 2004.

[56] M. Fiszman, D. Demner-Fushman, H. Kilicoglu, and T. C. Rindflesch. Automatic summarization of MEDLINE citations for evidence-based medical treatment: A topic-oriented evaluation. *Journal of Biomedical Informatics*, 42(5):801–813, 2009.

[57] K. Franzén, G. Eriksson, F. Olsson, L. Asker, P. Lidén, and J. Cöster. Protein names and how to find them. *International Journal of Medical Informatics*, 67(1-3):49–61, 2002.

[58] C. Friedman, G. Hripcsak, L. Shagina, and H. Liu. Arepresenting information in patient reports using natural language processing and the extensible markup language. *Journal of the American Medical Informatics Association*, 6:76–87, 1999.

[59] K. Fukuda, A. Tamura, T. Tsunoda, and T. Takagi. Toward information extraction: Identifying protein names from biological papers. In *Pacific Symposium on Biocomputing*, pages 707–718, 1998.

[60] K. Fundel, R. Küffner, and R. Zimmer. RelEx—relation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371, 2007.

[61] R. Gaizauskas, G. Demetriou, P. J. Artymiuk, and P. Willett. Protein structures and information extraction from biological texts: The PASTA system. *Bioinformatics*, 19(1):135–143, 2003.

[62] B. Gu. Recognizing nested named entities in GENIA corpus. In *Proceedings of the Workshop on Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis*, pages 112–113, 2006.

[63] J. Hakenberg, S. Bickel, C. Plake, U. Brefeld, H. Zahn, L. Faulstich, U. Leser, and T. Scheffer. Systematic feature evaluation for gene name recognition. *BMC Bioinformatics*, 6(Suppl 1):S9, 2005.

[64] J. Hakenberg, C. Plake, and U. Leser. LLL'05 challenge: Genic interaction extraction - identification of language patterns based on alignment and finite state automata. In *In Proceedings of the ICML 2005 Workshop on Learning Language in Logic*, pages 38–45, 2005.

[65] W. Hersh. *Information Retrieval: A Health and Biomedical Perspective.* Health Informatics. Springer, third edition, 2005.

[66] HighWire press. `http://highwire.org/`.

[67] L. Hirschman, M. Colosimo, A. Morgan, and A. Yeh. Overview of BioCreAtIvE task 1B: Normalized gene lists. *BMC Bioinformatics*, 6(Suppl 1):S11, 2005.

[68] L. Hirschman, A. A. Morgan, and A. S. Yeh. Rutabaga by any other name: Extracting biological names. *Journal of Biomedical Informatics*, 35(4):247–259, 2002.

[69] L. Hirschman, A. Yeh, C. Blaschke, and A. Valencia. Overview of BioCreAtIvE: Critical assessment of information extraction for biology. *BMC Bioinformatics*, 6(Suppl 1):S1, 2005.

[70] W.-J. Hou and H.-H. Chen. Enhancing performance of protein name recognizers using collocation. In *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine - Volume 13*, pages 25–32, 2003.

[71] D. Hristovski, C. Friedman, T. C. Rindflesch, and B. Peterlin. Exploiting semantic relations for literature-based discovery. In *AMIA Anual Symposium Proceedings*, pages 349–353, 2006.

[72] D. Hristovski, B. Peterlin, S. Džeroski, and J. Stare. Literature-based discovery support system and its application to disease gene identification. In S. Džeroski and L. Todorovski, editors, *Computational Discovery of Scientific Knowledge*, volume 4660 of *Lecture Notes in Computer Science*, pages 307–326. Springer Berlin / Heidelberg, 2007.

[73] D. Hristovski, B. Peterlin, J. A. Mitchell, and S. M. Humphrey. Improving literature-based discovery support by genetic knowledge integration. *Studies in Health Technogy and Informatics*, 95:68–73, 2003.

[74] D. Hristovski, B. Peterlin, J. A. Mitchell, and S. M. Humphrey. Using literature-based discovery to identify disease candidate genes. *International Journal of Medical Informatics*, 74(2-4):289–298, 2005.

[75] D. Hristovski, J. Stare, B. Peterlin, and S. Džeroski. Supporting discovery in medicine by association rule mining in MEDLINE and UMLS. In V. L. Patel, R. Rogers, and R. Haux, editors, *Proceedings of the 10th World Congress on Medical Informatics*, volume 84/2001 of *Studies in Health Technology and Informatics*, pages 1344–1348. IOS Press, 2001.

[76] X. Hu, X. Zhang, I. Yoo, X. Wang, and J. Feng. Mining hidden connections among biomedical concepts from disjoint biomedical literature sets through semantic-based association rule. *International Journal of Intelligent Systems*, 25(2):207–223, 2010.

[77] X. Huang, J. Lin, and D. Demner-Fushman. Evaluation of PICO as a knowledge representation for clinical questions. In *AMIA Annual Symposium Proceedings*, pages 359–363, 2006.

[78] K. Humphreys, G. Demetriou, and R. Gaizauskas. Two applications of information extraction to biological science yournal articles: Enzyme interactions and protein structures. In *Pacific Symposium on Biocomputing*, pages 502–513, 2000.

[79] L. Hunter, Z. Lu, J. Firby, W. Baumgartner, H. Johnson, P. Ogren, and K. B. Cohen. OpenDMAP: An open source, ontology-driven concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and cell-type-specific gene expression. *BMC Bioinformatics*, 9(1):78, 2008.

[80] Informatics for integrating biology and the bedside. `https://www.i2b2.org/resrcs/hive.html`.

[81] P. Jacqumart and P. Zweigenbaum. Towards a medical question-answering system: A feasibility study. *Studies in Health Technology and Informatics*, 95:463–468, 2003.

[82] R. Jelier, G. Jenster, L. Dorssers, B. Wouters, P. Hendriksen, B. Mons, R. Delwel, and J. Kors. Text-derived concept profiles support assessment of DNA microarray data for acute myeloid leukemia and for androgen receptor stimulation. *BMC Bioinformatics*, 8(1):14, 2007.

[83] R. Kabiljo, A. B. Clegg, and A. J. Shepherd. A realistic assessment of methods for extracting gene/protein interactions from free text. *BMC Bioinformatics*, 10:233, 2008.

[84] J. Kalpathy-Cramer, H. Müler, S. Bedrick, I. Eggel, A. de Herrera, and T. Tsikrika. The CLEF 2011 medical image retrieval and classification tasks. In *CLEF 2011 Working Notes*, 2011.

[85] H. Karsten and H. Suominen. Mining of clinical and biomedical text and data. *International Journal of Medical Informatics*, 78(12):786–787, 2009.

[86] J. Kazama, T. Makino, Y. Ohta, and J. Tsujii. Tuning support vector machines for biomedical named entity recognition. In *Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain - Volume 3*, pages 1–8, 2002.

[87] H. Kilicoglu and S. Bergler. Syntactic dependency based heuristics for biological event extraction. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, pages 119–127, 2009.

[88] J.-D. Kim, T. Ohta, N. Nguyen, S. Pyysalo, R. Bossy, and J. Tsujii. Overview of BioNLP shared task 2011. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, pages 1–6, 2011.

[89] J.-D. Kim, T. Ohta, S. Pyysalo, Y. Kano, and J. Tsujii. Overview of BioNLP'09 shared task on event extraction. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, pages 1–9, 2009.

[90] J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. GENIA corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(Suppl 1):i180–i182, 2003.

[91] J.-D. Kim, T. Ohta, Y. Tsuruoka, Y. Tateisi, and N. Collier. Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pages 70–75, 2004.

[92] S. Kim, J. Yoon, and J. Yang. Kernel approaches for genic interaction extraction. *Bioinformatics*, 24(1):118–126, 2008.

[93] S. Kinoshita, K. B. Cohen, P. Ogren, and L. Hunter. BioCreAtIvE task 1A: Entity identification with a stochastic tagger. *BMC Bioinformatics*, 6(Suppl 1):S4, 2005.

[94] J. Kontos, J. Lekakis, I. Malagardi, and J. Peros. Grammars for question answering systems based on intelligent text mining in biomedicine. In *Proceedings of the 7th Hellenic Europeoan Conference on Computer Mathematics and its Applications*, 2005.

[95] J. Kontos, I. Malagardi, and J. Peros. Question answering and rhetoric analysis of biomedical texts in the AROMA system. In *Proceedings of the 7th Hellenic Europeoan Conference on Computer Mathematics and its Applications*, 2005.

[96] M. Krallinger, F. Leitner, C. Rodriguez-Penagos, and A. Valencia. Overview of the protein-protein interaction annotation extraction task of BioCreAtIve II. *Genome Biology*, 9(Suppl 2):S4, 2008.

[97] M. Krallinger, A. Morgan, L. Smith, F. Leitner, L. Tanabe, J. Wilbur, L. Hirschman, and A. Valencia. Evaluation of text-mining systems for biology: Overview of the second BioCreAtIvE community challenge. *Genome Biology*, 9(Suppl 2):S1, 2008.

[98] M. Krallinger, A. Valencia, and L. Hirschman. Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome biology*, 9(Suppl 2):S8, 2008.

[99] M. Krauthammer and G. Nenadic. Term identification in the biomedical literature. *Journal of Biomedical Informatics*, 37(6):512–526, 2004.

[100] M. Krauthammer, A. Rzhetsky, P. Morozov, and C. Friedman. Using BLAST for identifying gene and protein names in journal articles. *Gene*, 259(1-2):245–252, 2000.

[101] R. Leaman and G. Gonzalez. BANNER: An executable survey of advances in biomedical named entity recognition. In *Pacific Symposium on Biocomputing*, pages 652–663, 2008.

[102] L. C. Lee, F. Horn, and F. E. Cohen. Automatic extraction of protein point mutations using a graph bigram association. *PLoS Computational Biology*, 3(2):e16, 2007.

[103] G. Leech. Adding linguistic annotation. In M. Wynne, editor, *Developing Linguistic Corpora: A Guide to Good Practice*, pages 17–29. Oxbow Books, 2005.

[104] U. Leser and J. Hakenberg. What makes a gene name? named entity recognition in the biomedical literature. *Briefings in Bioinformatics*, 6(4):357–369, 2005.

[105] M. Liberman, M. Mandel, and GlaxoSmithKline Pharmaceuticals R&D. PennBioIE CYP 1.0, 2008.

[106] M. Liberman, M. Mandel, and P. White. PennBioIE Oncology 1.0, 2008.

[107] C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out*, 2004.

[108] C.-Y. Lin, G. Cao, J. Gao, and J.-Y. Nie. An information-theoretic approach to automatic evaluation of summaries. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 463–470, 2006.

[109] J. Lin and D. Demner-Fushman. The role of knowledge in conceptual retrieval: A study in the domain of clinical medicine. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 99–106, 2006.

[110] R. T. K. Lin, J. Liang-Te Chiu, H.-J. Dai, M.-Y. Day, R. T.-H. Tsai, and W.-L. Hsu. Biological question answering with syntactic and semantic feature matching and an improved mean reciprocal ranking measurement. In *Proceedings of the 2008 IEEE International Conference on Information Reuse and Integration*, pages 184–189, 2008.

[111] D. A. Lindberg, B. L. Humphreys, and A. T. McCray. The unified medical language system. *Methods of Information in Medicine*, 32(4):281–291, 1993.

[112] X. Ling, J. Jiang, X. He, Q. Mei, C. Zhai, and B. Schatz. Generating gene summaries from biomedical literature: A study of semi-structured summarization. *Information Processing & Management*, 43(6):1777–1791, 2007.

[113] Y. Lussier, T. Borlawsky, D. Rappaport, Y. Liu, and C. Friedman. PheneGo: Assigning phenotypic context to gene ontology annotations with natural language processing. In *Pacific Symposium on Biocomputing*, pages 64–75, 2006.

[114] Y. Lussier, T. Borlawsky, D. Rappaport, Y. Liu, and C. Friedman. PhenoGo: Assigning phenotypic context to Gene Ontology annotations with natural language processing. In *Pacific Symposium on Biocomputing*, pages 64–75, 2006.

[115] D. Maynard. D1.2.2.1.3 benchmarking of annotation tools, 2007. `http://knowledgeweb.semanticweb.org/semanticportal/deliverables/D1.2.2.1.3.pdf`.

[116] K. R. McKeown, S.-F. Chang, J. Cimino, S. K. Feiner, C. Friedman, L. Gravano, V. Hatzivassiloglou, S. Johnson, D. A. Jordan, J. L. Klavans, A. Kushniruk, V. Patel, and S. Teufel. PERSIVAL, a system for personalized search and summarization over multimedia healthcare information. In *Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 331–340, 2001.

[117] S. Mika and B. Rost. Protein names precisely peeled off free text. *Bioinformatics*, 20(suppl 1):i241–i247, 2004.

[118] T. Mitsumori, S. Fation, M. Murata, K. Doi, and H. Doi. Gene/protein name recognition based on support vector machine using dictionary as features. *BMC Bioinformatics*, 6(Suppl 1):S8, 2005.

[119] M. Miwa, R. Sætre, and J.-D. Kim. Event extraction with complex event classification using rich features. *Journal of Bioinformatics and Computational Biology*, 8(1):131–146, 2010.

[120] M. Miwa, R. Sætre, Y. Miyao, and J. Tsujii. Protein-protein interaction extraction by leveraging multiple kernels and parsers. *International Journal of Medical Informatics*, 78(12):e39–e46, 2009.

[121] Y. Miyao, T. Ohta, K. Masuda, Y. Tsuruoka, K. Yoshida, T. Ninomiya, and J. Tsujii. Semantic retrieval for the accurate identification of relational concepts in massive textbases. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 1017–1024, 2006.

[122] Y. Miyao, K. Sagae, R. Sætre, T. Matsuzaki, and J. Tsujii. Evaluating contributions of natural language parsers to protein-protein interaction extraction. *Bioinformatics*, 25(3):394–400, 2009.

[123] L. P. Morales, A. D. Esteban, and P. Gervás. Concept-graph based biomedical automatic summarization using ontologies. In *Proceedings of the 3rd Textgraphs Workshop on Graph-Based Algorithms for Natural Language Processing*, pages 53–56, 2008.

[124] A. Morgan, L. Hirschman, A. Yeh, and M. Colosimo. Gene name extraction using FlyBase resources. In *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine - Volume 13*, pages 1–8, 2003.

[125] A. A. Morgan, L. Hirschman, M. Colosimo, A. S. Yeh, and J. B. Colombe. Gene name identification and normalization using a model organism database. *Journal of Biomedical Informatics*, 37(6):396–410, 2004.

[126] A. A. Morgan, Z. Lu, X. Want, A. M. Cohen, J. Fluck, P. Ruch, A. Divoli, K. Fundel, R. Leaman, J. Hakenberg, C. Sun, H.-h. Liu, R. Torres, M. Krauthammer, W. W. Lau, H. Liu, C.-N. Hsu, M. Scheumie, K. B. Cohen, and L. Hirschman. Overview of BioCreAtIvE II: Gene normalization. *Genome Biology*, 9(Suppl 2):S3, 2008.

[127] H. Müller, J. Kalpathy-Cramer, I. Eggel, S. Bedrick, C. E. Charles E. Kahn, Jr., and W. Hersh. Overview of the clef 2010 medical image retrieval track. In *Working Notes of CLEF 2010*, 2010.

[128] M. Narayanaswamy, K. E. Ravikumar, and K. Vijay-Shanker. A biological named entity recognizer. In *Pacific Symposium on Biocomputing*, pages 427–438, 2003.

[129] National center for biomedical ontology. `http://www.bioontology.org/`.

[130] NCBO BioPortal. `http://bioportal.bioontology.org/`.

[131] National Center for Biotechnology Information. *Entrez Programming Utilities Help*, 2010. `http://www.ncbi.nlm.nih.gov/books/NBK25501/`.

[132] National centre for text mining. `http://www.nactem.ac.uk/`.

[133] C. Nédellec. Learning language in logic - genic interaction extraction challenge. In *In Proceedings of the ICML 2005 Workshop on Learning Language in Logic*, pages 31–37, 2005.

[134] Neuroscience information framework. `http://neuinfo.org/`.

[135] Y. Niu and G. Hirst. Analysis and semantic classes in medical text for question answering. In *Proceedings of the ACL 2004 Workshop on Question Answering in Restricted Domains*, 2004.

[136] Y. Niu, G. Hirst, G. McArthur, and R.-G. P. Answering clinical questions with role identification. In *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine*, pages 73–80, 2003.

[137] Y. Niu, X. Zhu, and G. Hirst. Using outcome polarity in sentence extraction for medical question-answering. In *AMIA Anual Symposium Proceedings*, pages 599–603, 2006.

[138] Y. Niu, X. Zhu, J. Li, and G. Hirst. Analysis of polarity information in medical text. In *AMIA Anual Symposium Proceedings*, pages 570–574, 2005.

[139] C. Nobata, N. Collier, and J.-i. Tsujii. Automatic term identification and classification in biology texts. In *Proceedings of the Natural Language Pacific Rim Symposium*, pages 369–374, 1999.

[140] P. V. Ogren. Knowtator: A protégé plug-in for annotated corpus construction. In *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 273–275, 2006.

[141] D. Okanohara, Y. Miyao, Y. Tsuruoka, and J. Tsujii. Improving the scalability of semi-Markov conditional random fields for named

entity recognition. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 465–472, 2006.

[142] F. Olsson, G. Eriksson, K. Franzén, L. Asker, and P. Lidén. Notions of correctness when evaluating protein name taggers. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*, pages 1–7, 2002.

[143] Open biological and biomedical ontologies. `http://www.obofoundry.org/`.

[144] ORBIT project. `http://orbit.nlm.nih.gov/`.

[145] A. Özgür, T. Vu, G. Erkan, and D. R. Radev. Identifying gene-disease associations using centrality on a literature mined gene-interaction network. *Bioinformatics*, 24(13):i277–i285, 2008.

[146] A. Özgür, Z. Xiang, D. R. Radev, and Y. He. Literature-based discovery of IFN-γ and vaccine-mediated gene interaction networks. *Journal of Biomedicine & Biotechnology*, page 426479, 2010.

[147] E. Pafilis, S. O'Donoghue, L. Jensen, H. Horn, M. Kuhn, N. Brown, and R. Schneider. Reflect - augmented browsing for the life scientist. *Nature Biotechnology*, 27:508–510, 2009.

[148] S. Pakhomov. Semi-supervised maximum entropy based approach to acronym and abbreviation normalization in medical texts. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 160–167, 2002.

[149] M. Palakal, J. Bright, T. Sebastian, and S. Hartanto. A comparative study of cells in inflammation, EAE and MS using biomedical literature data mining. *Journal of Biomedical Science*, 14(1):67–85, 2007.

[150] V. Petri, M. Shimoyama, G. Hayman, J. Smith, M. Tutaj, J. de Pons, M. Dwinell, D. Munzenmaier, S. Twigger, and H. Jacob. The rat genome database pathway portal. *Database*, 2011.

[151] I. Petrič, U. Tanja, B. Cestnik, and M. Macedoni-Lukšič. Literature mining method RaJoLink for uncovering relations between biomedical concepts. *Journal of Biomedical Informatics*, 42(2):219–227, 2009.

[152] Pharmacogenomics knowledge base. `http://www.pharmgkb.org/`.

[153] H. Poon and L. Vanderwende. Joint inference for knowledge extraction from biomedical literature. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter*

of the Association for Computational Linguistics, pages 813–821, 2010.

[154] PubMed central open access subset. `http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/`.

[155] S. Pyysalo, A. Airola, J. Heimonen, J. Bjorne, F. Ginter, and T. Salakoski. Comparative analysis of five protein-protein interaction corpora. *BMC Bioinformatics*, 9(Suppl 3):S6, 2008.

[156] S. Pyysalo, F. Ginter, J. Heimonen, J. Bjorne, J. Boberg, J. Jarvinen, and T. Salakoski. BioInfer: A corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8(1):50, 2007.

[157] L. A. Ramshaw and M. P. Marcus. Text chunking using transformation-based learning. In *3rd ACL SIGDAT Workshop on Very Large Corpora*, pages 82–94, 1995.

[158] L. H. Reeve, H. Han, and A. D. Brooks. The use of domain-specific concepts in biomedical text summarization. *Information Processing & Management*, 43(6):1765–1776, 2007.

[159] W. S. Richardson, M. C. Wilson, J. Nishikawa, and R. S. Hayward. The well-built clinical question: A key to evidence-based decisions. *ACP Journal Club*, 123(3):A12–A13, 1995.

[160] S. Riedel, H.-W. Chun, T. Takagi, and J. Tsujii. A Markov logic approach to bio-molecular event extraction. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, pages 41–49, 2009.

[161] S. Riedel and A. McCallum. Fast and robust joint models for biomedical event extraction. In *Proceedings of the 2011 Conference on Emperical Methods in Natural Language Processing*, pages 1–12, 2011.

[162] F. Rinaldi, J. Dowdall, G. Schneider, and A. Persidis. Answering questions in the genomics domain. In *Proceedings of the ACL 2004 Workshop on Question Answering in Restricted Domains*, 2005.

[163] F. Rinaldi, K. Kaljurand, and R. Saetre. Terminological resources for text mining over biomedical scientific literature. *Artificial Intelligence in Medicine*, 52(2):107–114, 2011.

[164] F. Rinaldi, G. Schneider, K. Kaljurand, M. Hess, C. Andronis, O. Konstandi, and A. Persidis. Mining of relations between proteins over biomedical scientific literature using a deep-linguistic approach. *Artificial Intelligence in Medicine*, 39(2):127–136, 2007.

[165] T. C. Rindflesch and M. Fiszman. The interaction of domain knowledge and linguistic structure in natural language processing:

Interpreting hypernymic propositions in biomedical text. *Journal of Biomedical Informatics*, 36(6):462–477, 2003.

[166] T. C. Rindflesch, H. Kilicoglu, M. Fiszman, G. Rosemblat, and D. Shin. Semantic MEDLINE: An advanced information management application for biomedicine. *Information Services & Use*, 31:15–21, 2011.

[167] B. Rink, S. Harabagiu, and K. Roberts. Automatic extraction of relations between medical concepts in clinical texts. *Journal of the American Medical Informatics Association*, 18(5):594–600, 2011.

[168] A. Roberts, R. Gaizauskas, and M. Hepple. Extracting clinical relationships from patient narratives. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 10–18, 2008.

[169] P. Ruch, C. Boyer, C. Chichester, I. Tbahriti, A. Geissbühler, P. Fabry, J. Gobeill, V. Pillet, D. Rebholz-Schuhmann, C. Lovis, and A.-L. Veuthey. Using argumentation to extract key sentences from biomedical abstracts. *International Journal of Medical Informatics*, 76(2-3):195–200, 2007.

[170] D. L. Sackett, W. M. C. Rosenberg, J. A. M. Gray, and R. B. Haynes. Evidence based medicine: What it is and what it isn't. *British Medical Journal*, 312(7023):71–72, 1996.

[171] M. Saeed, M. Villarroel, A. Reisner, G. Clifford, L. Lehman, G. Moody, T. Heldt, T. Kyaw, B. Moody, and R. Mark. Multiparameter intelligent monitoring in intensive care II (MIMIC-II): A public-access intensive care unit database. *Crit Care Med*, 39(5):952–960, 2011.

[172] J. Šarić, L. J. Jensen, R. Ouzounova, I. Rojas, and P. Bork. Extraction of regulatory gene/protein networks from MEDLINE. *Bioinformatics*, 22(6):645–650, 2006.

[173] Y. Sasaki, Y. Tsuruoka, J. McNaught, and S. Ananiadou. How to make the most of NE dictionaries in statistical NER. *BMC Bioinformatics*, 9(Suppl 11):S5, 2008.

[174] J. Seki, K. Mostafa. Discovering implicit associations between genes and hereditary diseases. In *Pacific Symposium on Biocomputing*, pages 316–327, 2007.

[175] B. Settles. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pages 104–107, 2004.

[176] B. Settles. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics*, 21(4):3191–3192, 2005.

[177] H. Shatkay, F. Pan, A. Rzhetsky, and W. Wilbur. Multi-dimensional classification of biomedical text: toward automated, practical provision of high-utility text to diverse users. *Bioinformatics*, 24(18):2086–2093, 2008.

[178] H. Shatkay, J. W. Wilbur, and A. Rzhetsky. Annotation guidelines, 2005. `http://www.ncbi.nlm.nih.gov/CBBresearch/Wilbur/AnnotationGuidelines.pdf`.

[179] D. Shen, J. Zhang, G. Zhou, J. Su, and C.-L. Tan. Effective adaptation of a hidden markov model-based named entity recognizer for biomedical domain. In *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine - Volume 13*, pages 49–56, 2003.

[180] Z. Shi, G. Melli, Y. Wang, Y. Liu, B. Gu, M. Kashani, A. Sarkar, and F. Popowich. Question answering summarization of multiple biomedical documents. In Z. Kobti and D. Wu, editors, *Advances in Artificial Intelligence*, volume 4509 of *Lecture Notes in Computer Science*, pages 284–295. Springer Berlin / Heidelberg, 2007.

[181] M. S. Simpson, D. Demner-Fushman, and G. R. Thoma. Evaluating the importance of image-related text for ad-hoc and case-based biomedical article retrieval. In *AMIA Annual Symposium Proceedings*, pages 752–756, 2010.

[182] N. Smalheiser. The Arrowsmith project: 2005 status report. In A. Hoffmann, H. Motoda, and T. Scheffer, editors, *Discovery Science*, volume 3735 of *Lecture Notes in Computer Science*, pages 26–43. Springer Berlin / Heidelberg, 2005.

[183] N. Smalheiser, V. Torvik, A. Bischoff-Grethe, L. Burhans, M. Gabriel, R. Homayouni, A. Kashef, M. Martone, G. Perkins, D. Price, A. Talk, and R. West. Collaborative development of the arrowsmith two node search interface designed for laboratory investigators. *Journal of Biomedical Discovery and Collaboration*, 1(1):8, 2006.

[184] N. Smalheiser, W. Zhou, and V. Torvik. Anne O'Tate: A tool to support user-driven summarization, drill-down and browsing of PubMed search results. *Journal of Biomedical Discovery and Collaboration*, 3(1):2, 2008.

[185] N. R. Smalheiser and D. R. Swanson. Using Arrowsmith: A computer-assisted approach to formulating and assessing scien-

tific hypotheses. *Computer Methods and Programs in Biomedicine*, 57(3):149–153, 1998.

[186] N. R. Smalheiser, V. I. Torvik, and W. Zhou. Arrowsmith two-node search interface: A tutorial on finding meaningful links between two disparate sets of articles in MEDLINE. *Computer Methods and Programs in Biomedicine*, 94(2):190–197, 2009.

[187] L. Smith, L. Tanabe, R. Johnson nee Ando, C.-J. Kuo, I.-F. Chung, C.-N. Hsu, Y.-S. Lin, R. Klinger, C. Friedrich, K. Ganchev, M. Torii, H. Liu, B. Haddow, C. Struble, R. Povinelli, A. Vlachos, W. Baumgartner, L. Hunter, B. Carpenter, R. Tzong-Han Tsai, H.-J. Dai, F. Liu, Y. Chen, C. Sun, S. Katrenko, P. Adriaans, C. Blaschke, R. Torres, M. Neves, P. Nakov, A. Divoli, M. Mana-Lopez, J. Mata, and W. Wilbur. Overview of BioCreAtIve II: Gene mention recognition. *Genome Biology*, 9(Suppl 2):S2, 2008.

[188] M. Q. Stearns, C. Price, K. A. Spackman, and A. Y. Wang. SNOWMED clinical terms: Overview of the development process and project status. In *Proceedings of the AMIA Symposium*, pages 662–666, 2001.

[189] D. R. Swanson. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, 30(1):7–18, 1986.

[190] D. R. Swanson. Migraine and magnesium: Eleven neglected connections. *Perspectives in Biology and Medicine*, 31(4):526–557, 1988.

[191] D. R. Swanson. Somatomedin C and arginine: Implicit connections between mutually isolated literatures. *Perspectives in Biology and Medicine*, 33(2):157–186, 1990.

[192] D. R. Swanson. Complementary structures in disjoint science literatures. In *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 280–289, 1991.

[193] D. R. Swanson and N. R. Smalheiser. An interactive system for finding complementary literatures: A stimulus to scientific discovery. *Artificial Intelligence*, 91(2):183–203, 1997.

[194] D. R. Swanson, N. R. Smalheiser, and A. Bookstein. Information discovery from complementary literatures: Categorizing viruses as potential weapons. *Journal of the American Society for Information Science and Technology*, 52(10):797–812, 2001.

[195] K. Takahashi, A. Koike, and T. Takagi. Question answering system in biomedical domain. In *Proceedings of the 15th International Conference on Genome Informatics*, pages 161–162, 2004.

[196] K. Takeuchi and N. Collier. Bio-medical entity extraction using support vector machines. *Artificial Intelligence in Medicine*, 33(2):125–137, 2005.

[197] R. M. Terol, P. Martínez-Barco, and M. Palomar. A knowledge based method for the medical question answering problem. *Computers in Biology and Medicine*, 37(10):1511–1521, 2007.

[198] P. Thompson, S. Iqbal, J. McNaught, and S. Ananiadou. Construction of an annotated corpus to support biomedical information extraction. *BMC Bioinformatics*, 10(1):349, 2009.

[199] V. I. Torvik and N. R. Smalheiser. A quantitative model for linking two disparate sets of articles in MEDLINE. *Bioinformatics*, 23(13):1658–1665, 2007.

[200] TREC-9 filtering track collections. `http://trec.nist.gov/data/t9_filtering.html`.

[201] TREC genomics track data. `http://ir.ohsu.edu/genomics/data.html`.

[202] R. Tsai, W.-C. Chou, Y.-S. Su, Y.-C. Lin, C.-L. Sung, H.-J. Dai, I. Yeh, W. Ku, T.-Y. Sung, and W.-L. Hsu. BIOSMILE: A semantic role labeling system for biomedical berbs using a maximum-entropy model with automatically generated template features. *BMC Bioinformatics*, 8(1):325, 2007.

[203] Y. Tsuruoka, M. Miwa, K. Hamamoto, J. Tsujii, and S. Ananiadou. Discovering and visualizing indirect associations between biomedical concepts. *Bioinformatics*, 27(13):i111–i119, 2011.

[204] Y. Tsuruoka and J. Tsujii. Boosting precision and recall of dictionary-based protein name recognition. In *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine - Volume 13*, pages 41–48, 2003.

[205] Y. Tsuruoka and J. Tsujii. Probabilistic term variant generator for biomedical terms. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pages 167–173, 2003.

[206] Y. Tsuruoka, J. Tsujii, and S. Ananiadou. FACTA: A text search engine for finding associated biomedical concepts. *Bioinformatics*, 24(21):2559–2560, 2008.

[207] O. Tuason, L. Chen, L. H., and C. Friedman. Biological nomenclatures: A source of lexical knowledge and ambiguity. In *Pacific Symposium on Biocomputing*, pages 238–249, 2004.

[208] H. Turtle and W. B. Croft. Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9:187–222, 1991.

[209] Orange book: Approved drug products with therapeutic equivalence evaluations. `http://www.accessdata.fda.gov/scripts/cder/ob/default.cfm`.

[210] Databases, resources & APIs. `http://wwwcf2.nlm.nih.gov/nlm_eresources/eresources/search_database.cfm`.

[211] University of Pittsburgh NLP repository. `http://www.dbmi.pitt.edu/nlpfront`.

[212] Y. Usami, H.-C. Cho, N. Okazaki, and J. Tsujii. Automatic acquisition of huge training data for bio-medical named entity recognition. In *Proceedings of BioNLP 2011 Workshop*, pages 65–73, 2011.

[213] O. Uzuner. Recognizing obesity and comorbidities in sparse data. *Journal of the American Medical Informatics Association*, 16(5):561–570, 2009.

[214] O. Uzuner, I. Goldstein, Y. Luo, and I. Kohane. Identifyingn patient smoking status from medical discharge records. *Journal of the American Medical Informatics Association*, 15(1):14–24, 2008.

[215] O. Uzuner, I. Solti, and E. Cadag. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–518, 2010.

[216] O. Uzuner, B. R. South, S. Shen, and S. L. DuVall. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556, 2011.

[217] V. Vincze, G. Szarvas, R. Farkas, G. Mora, and J. Csirik. The BioScope corpus: Biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(Suppl 11):S9, 2008.

[218] A. Vlachos and C. Gasperin. Bootstrapping and evaluating named entity recognition in the biomedical domain. In *Proceedings of the HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology*, pages 138–145, 2006.

[219] T. Wattarujeekrit, P. Shah, and N. Collier. PASBio: Predicate-argument structures for event extraction in molecular biology. *BMC Bioinformatics*, 5(1):155, 2004.

[220] M. Weeber, H. Klein, L. T. W. de Jong-van den Berg, and R. Vos. Using concepts in literature-based discovery: Simulating Swanson's Raynaud-fish oil and migraine-magnesium discoveries. *Journal of the American Society for Information Science and Technology*, 52(7):548–557, 2001.

[221] W. Weiming, D. Hu, M. Feng, and L. Wenyin. Automatic clinical question answering based on UMLS relations. In *Third International Conference on Semantics, Knowledge and Grid*, pages 495–498, 2007.

[222] J. W. Wilbur, A. Rzhetsky, and H. Shatkay. New directions in biomedical text annotation: Definitions, guidelines and corpus construction. *BMC Bioinformatics*, 7:356, 2006.

[223] G. Williams, P. Davis, A. Rogers, T. Bieri, P. Ozersky, and J. Spieth. Methods and strategies for gene structure curation in wormbase. *Database*, 2011.

[224] K. Yamamoto, T. Kudo, A. Konagaya, and Y. Matsumoto. Protein name tagging for biomedical annotation in text. In *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine - Volume 13*, pages 65–72, 2003.

[225] J. Yang, A. M. Cohen, and W. Hersh. Automatic summarization of mouse gene information by clustering and sentence extraction from MEDLINE abstracts. In *AMIA Annual Symposium Proceedings*, pages 831–835, 2007.

[226] A. Yeh, A. Morgan, M. Colosimo, and L. Hirschman. BioCreAtIvE task 1A: Gene mention finding evaluation. *BMC Bioinformatics*, 6(Suppl 1):S2, 2005.

[227] M. Yetisgen-Yildiz and W. Pratt. Using statistical and knowledge-based approaches for literature-based discovery. *Journal of Biomedical Informatics*, 39(6):600–611, 2006.

[228] M. Yetisgen-Yildiz and W. Pratt. A new evaluation methodology for literature-based discovery systems. *Journal of Biomedical Informatics*, 42(4):633–643, 2009.

[229] I. Yoo, X. Hu, and I.-Y. Song. A coherent graph-based semantic clustering and summarization approach for biomedical literature and a new summarization evaluation method. *BMC Bioinformatics*, 8(Suppl 9):S4, 2007.

[230] H. Yu, S. Agarwal, M. Johnston, and A. Cohen. Are figure legends sufficient? Evaluating the contribution of associated text to biomedical figure comprehension. *Journal of Biomedical Discovery and Collaboration*, 4(1):1, 2009.

[231] H. Yu and Y.-G. Cao. Automatically extracting information needs from ad hoc clinical questions. In *AMIA Annual Symposium Proceedings*, pages 96–100, 2008.

[232] H. Yu and M. Lee. Accessing bioscience images from abstract sentences. *Bioinformatics*, 22(14):e547–e556, 2006.

[233] H. Yu, M. Lee, D. Kaufman, J. Ely, J. A. Osheroff, G. Hripcsak, and J. Cimino. Development, implementation, and a cognitive evaluation of a definitional question answering system for physicians. *Journal of Biomedical Informatics*, 40(3):236–251, 2007.

[234] H. Yu and C. Sable. Being Erlang Shen: Identifying answerable questions. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence on Knowledge and Reasonin for Answering Questions*, pages 6–14, 2005.

[235] H. Yu, C. Sable, and H. Zhu. Classifying medical questions based on an evidence taxonomy. In *Proceedings of the AAAI 2005 Workshop on Question Answering in Restricted Domains*, 2005.

[236] G. Zhou, D. Shen, J. Zhang, J. Su, and S. Tan. Recognition of protein/gene names from text using an ensemble of classifiers. *BMC Bioinformatics*, 6(Suppl 1):S7, 2005.

[237] P. Zweigenbaum and D. Demner-Fushman. Advanced literature-mining tools. In D. Edwards, J. Stajich, and D. Hansen, editors, *Bioinformatics: Tools and Applications*, pages 347–380. Springer, 2009.

[238] P. Zweigenbaum, D. Demner-Fushman, H. Yu, and K. B. Cohen. Frontiers of biomedical text mining: Current progress. *Briefings in Bioinformatics*, 8(5):358–375, 2007.