Chapter 21

Application of Biomedical Text Mining

Lejun Gong

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.75924

Abstract

With the enormous volume of biological literature, increasing growth phenomenon due to the high rate of new publications is one of the most common motivations for the biomedical text mining. Aiming at this massive literature to process, it could extract more biological information for mining biomedical knowledge. Using the information will help understand the mechanism of disease generation, promote the development of disease diagnosis technology, and promote the development of new drugs in the field of biomedical research. Based on the background, this chapter introduces the rise of biomedical text mining. Then, it describes the biomedical text-mining technology, namely natural language processing, including the several components. This chapter emphasizes the two aspects in biomedical text mining involving static biomedical information recognization and dynamic biomedical information extraction using instance analysis from our previous works. The aim is to provide a way to quickly understand biomedical text mining for some researchers.

Keywords: bioinformatics, text mining, natural language processing, information extraction

1. Introduction

With the rapid growth of the high-throughput biological technology, the study of biomedical science is entering omics era. It brings several omics data including genomics and transcriptomics; the vast amounts of biological data continue to emerge out of the life science research. The new phenomenon, new discovery, and new experimental data in biomedical research are mostly published in science journals by electronic text form. A large number of biological information is scattered in all kinds of studies. Handling these biomedical literatures could extract more biological information and discover new biomedical knowledge. Manual processing is like looking for a needle in a haystack. Biomedical literature can be seen

as a large unstructured data repository, which makes text mining come into play. Text mining has emerged as a potential solution to achieve knowledge for bridging between the free text and structured representation of biomedical information using artificial intelligence technology including natural language processing (NLP), machine learning (ML), and data mining to process large text collections. Therefore, text-mining technology is a powerful tool for mining valuable information from biomedical literature. There are broader definitions of biomedical text mining. Namely any work that extracts information from text could be considered as text mining, which would include a range of static information recognition, from dynamic information extraction to application of biomedical text mining. In the following sections, we describe the details according to the abovementioned aspects.

2. Natural language processing

Natural language processing is a field of artificial intelligence in computer science with interaction between computers and natural languages. With the mushroom growth of machine learning, much natural language processing research has a great relationship with machine learning. Many machine-learning algorithms have been applied to natural language-processing tasks. Extracting structured data from the complexity of heterogeneous-narrated medical reports is significantly challenged, and the work [1] obtained the results with F1 scores greater than 95% using machine learning, for example, HMM model. Weng et al. [2] proposed the machine learning to classify clinical notes to the medical subdomain. They reported that the classifier of the convolutional recurrent neural network with word embeddings yielded the best performance on iDASH and MGH datasets with F1 scores of 0.845 and 0.870, respectively. Basaldella et al. [3] proposed a hybrid approach using a two-stage pipeline with a machine-learning classifier combining a dictionary approach, and they achieved an overall precision of 86% at a recall of 60% on the named entity recognition task. For helping to obtain biomedical knowledge, a flourishing of ontologies attempted to represent the complexity of the biomedical concepts in text-mining area. The ontologies describe a wide variety of biological concepts spanning from biology to medicine. Moreover, they not only attempt to capture the meaning of a particular domain based on biomedical community but also are key element for knowledge management, and data integration [4]. Ontologies and controlled vocabularies could improve the efficiency and consistency of biomedical data curation, which has a great increasing interest in developing ontologies. For example, gene ontology (GO) [5] is a community-based bioinformatics resource related to gene function used to represent biological knowledge involving three aspects: molecular function, cellular component, and biological process. Molecular function describes the molecular activities of gene products, such as binding or catalysis. Cellular component indicates the place where gene products are active, namely a component of a cell such as an anatomical structure. While biological process represents pathways, larger processes are made up of the activities of multiple gene products. Another is GENIA [6] corpus developed to provide reference materials to allow natural language-processing techniques work for extracting information. GENIA corpus is a semantically annotated corpus of biological literature, which is being compiled and annotated in the scope of GENIA project, aiming at providing high-quality reference materials for bioinformatics. Natale et al. [7] proposed protein ontology (PRO) for enhancing and scaling

up the representation of protein entities represented in OWL and SPARQL format. Ontologies provide machine-readable descriptions for biomedical concepts linking domain-specific vocabulary. Ontology-based mining systems attempt to map a terminology in text to a concept in an ontology. Kim et al. [8] applied syntactic parsing on sentences with annotated GO concepts to exploit similarities of sentential syntactic dependencies for mapping to concepts.

The natural language processing (NLP) components include general tasks including tokenization, morphological analysis, POS tagging, and syntactic parsing.

Tokenization describes the processing that the text is broken into sentences and words. When the text is put into text-mining system, for example, a paper could be viewed as a continuous word stream; they are first broken up into chapters and paragraphs, and then the broken paragraphs continue to be pieced as sentences, words, and phonemes for a more sophisticated processing. For the tokenization task, the tokenizer could extract token features which are types of capitalization, digits, punctuation, special characters, and so on.

POS tagging is aiming at the words to annotate tags based on the context in the text. POS tags divide words into categories based on the role in the sentence. POS tags provide information about the word's semantic content. Nouns usually denote the entities, whereas prepositions express relationships between entities.

Syntactical parsing performs a full syntactical analysis of sentences according to a certain grammar theory including constituency and dependency grammars. Constituency grammars describe the syntactical structure of sentences in terms of phrases, namely element sequences. Most constituency grammars generally contain noun phrases, verb phrases, prepositional phrases, adjective phrases, and so on. Each phrase may consist of smaller phrases or words according to the rules of the grammar. The role of different phrases is contained in the syntactical structure of sentences. For example, a noun phrase may be marked as the subject of sentence, object. Dependency grammars focus on the direct relations between words, not considering the constituents. Dependency analysis uses Direct Acyclic Graph (DAG) to denote the relations between words using nodes and dependencies for edges. For example, a subject depends on the predicate verb, while an adjective depends on the noun and so on.

3. Biomedical text mining

With the enormous volume of biological literature, increasing growth phenomenon due to the high rate of new publications is one of the most common motivations for the biomedical text mining. It is reported that the growth in PubMed/Medline literature is exponential at this rate of publication. Thus, it is very difficult for researchers to keep up with the relevant publications in their own discipline, let alone related other disciplines.

Such a large scale and the rapid growth of biomedical literature data, carrying a lot of biological information, some new phenomena, biomedical discoveries, and new experimental data are often published in recent biomedical literature. Aiming at this massive literature to process, it could extract more biological information for mining hidden biomedical knowledge. These

vast amounts of biomedical literature, even in the field of expert, could not rely on the manual way from fully grasp the status quo and development trend of the research to obtain the information of interest for extracting biomedical knowledge. It is the urgent needs of application of text mining and information extraction from biomedical literature in the field of molecular biology and biomedical knowledge extraction. Biomedical text mining [9] is the frontier research field containing the collection combined computational linguistics, bioinformatics, medical information science, research fields, and so on. The development of biomedical text mining is less than 25 years [10], which belongs to a branch of bioinformatics. Bioinformatics is defined as application information science and technology to understand, organize, and manage biomolecular data. It aims to provide some tools and resources for biological researchers, facilitate them to get biological data, and analyze data, so as to discover new knowledge [11] of the biological world. Biomedical text mining is a sub-field of bioinformatics. It refers to the use of text-mining technology to process biomedical literature object, acquire biological information, organize and manage the acquired bioinformation, and provide it to researchers. Therefore, biomedical text mining can extract various biological information [12], such as gene and protein information, gene expression regulation information, gene polymorphism and epigenetic information, gene, and disease relationship. The biological information can help people to understand life phenomena and understand the rules of life activities. Using the information will help understand the mechanism of disease generation, promote the development of disease diagnosis technology, and promote the development of new drugs in the field of biomedical research. A large number of text-mining methods have been established to assist in the extraction of biological information. These methods could be proposed for extracting information that vary in their degree of reliance on dictionaries, statistical and knowledge-based approaches, automatic rule generation applying part-of-speech (POS) tagger, and some machine-learning algorithms, for example, Hidden Markov Models (HMMs) and decision trees. Cronin et al. [13] classified patient portal messages by a comparison of rule-based and machine-learning approach using a bag of words and natural language-processing (NLP) approaches. The best performance of classifier for individual communication subtypes was random forests for logistical-contact information with 0.963 receiver-operator curve.

In addition, there are some social institutes to focus on the development of biomedical text-mining technology. Based on the rapid development of omics era, the BioCreative called Critical Assessment of Information Extraction system in Biology is a community-wide effort for the evaluation of text mining and information extraction systems applied to the biomedicine domain using natural language processing [14, 15].The researches of biomedical text mining are presented at several conferences including Pacific Symposium on Biocomputing, BioNLP, and Practical Applications of Computational Biology and Bioinformatics [10].

4. Static biomedical information recognition

In the era of system biology, from the system perspective-related information on molecular biology research includes both biomedical entities, some genes, proteins, gene products, drugs, diseases such as basic, static entities to reflect its existence form, called static biomedical information.

There are some biological terminologies that describe domain objects in the medical literature, which is called entity. Such as gene that is the essence of life information, protein information that is the executor of gene function in biomedicine, identification of these entities in the life sciences plays an important role in revealing the phenomenon of life, which is the only way which must be passed to further explore these important biological entities, but also an important task in biomedical text mining. Biological entity representation in biomedical literature is extremely complex. The complexity of performance in both single entity in the form of a word entity, variable word length, and uppercase and lowercase mixed together, for example, urokinase, Cactus, IkapaBalpha, and so on. There are multi-words to form phrases, such as bradykinin B (1) receptor, protein phosphatase 2A, which brought a great deal of difficulty to establish biological entity boundaries; some of the same words or phrases that can be expressed in different categories of biomedical entities, such as c-myc, IL-2 protein can be expressed, can be said to detect gene, through the context, some biomedical entities have different forms of writing, such as protein phosphatase 2A, protein phosphatase, and 2A protein phosphatase 2A and refers to as the same biological entity in biomedical text. The complexity and diversity of the entity recognition of biomedical named entities has become a challenging study. Traditional recognition approaches have three methods containing dictionary-based, rules-based, and statistical machine learning.

Dictionary-based approach needs a detailed term dictionary for entity recognition from the document. Generally, the pre-given term dictionary is edited by the specific biological molecular database. Accordingly, the approach is limited by the term dictionary coverage but it has the relatively higher accuracy. Thus, the approach is widely applied in the actual development in the system. For example, Whatizit [16] and FRENDA [17].

Rule-based approach requires studying the entities' named features and laws for formulating rules to identify the entity, which make the staff to develop biomedical domain background knowledge. For example, biomedical entities are often noun phrases. The first character of a human gene nomenclature is a letter, and the rest is a combination of letters and numbers. Based on the successful rule system, AbGene [18], Tanabe and Wilbur used it to produce a large and high-quality gene protein dictionary from biomedical texts, and it is also used as a component of relationship extraction [19].

Statistical machine-learning approach is the rapid development along with the construction of biomedical corpora. The annotated corpus could be used as the training corpus of statistical machine-learning method. With the help of the biomedical corpora, entities could be identified from the biomedical text. There are some researches in the aspects. ABNER developed by Settles [20] used the Conditional Random Fields (CRF) as the statistical model to identify biomedical entities with the average 72.0% recall, the 69.1% precise, and 70.5% F-score in JNLPBA using morphological and semantic features. Mitsumori et al. [21] proposed an approach to process entity using Support Vector Machine (SVM) as a statistical model with internal and external resource features, which show that the performance of identification is improved by using the external biological dictionary features. Saha et al. [22] used maximum entropy model combined with word-clustering features and feature selection techniques to identify biomedical entities. The approach achieves better performance without using domain knowledge. Li et al. [23] use two-step CRFs to identify biomedical entities, the first CRF model

is used to identify named entity and the second CRF model is used to the types of named entities, which obtained 74.3% F-score in JNLPBA corpus. The mentioned three approaches have their own advantages, respectively. There is also a hybrid approach to be used for identifying biomedical entities. Our research group does some works related to identify biomedical entities involving the abovementioned three methods. We give several examples to illustrate the related static biomedical information recognition.

4.1. Dictionary-based approach to identify biomedical entities

In this section, we introduce our previous work [24–26] which is published in the International Journal of Pattern Recognition and Artificial Intelligence. We first achieve experimental literature and build a concept dictionary based on the authoritative corpus. Using part-of-speech (POS) tagging, phrase block's formulation and designed VWIA algorithm to identify entities for matching biomedical concepts. **Figure 1** [24] describes the pipeline.

4.1.1. Obtain experimental data

The experimental data of the study are from PubMed/Medline using the e-utilities API tool which is also used in the works [27–29] for automatically downloading literatures from the website (http://www.ncbi.nlm.nih.gov/). It looks like the web spider to catch a series of hyperlinks with a Uniform Resource Locator (URL). By obtaining URL using e-utilities related to

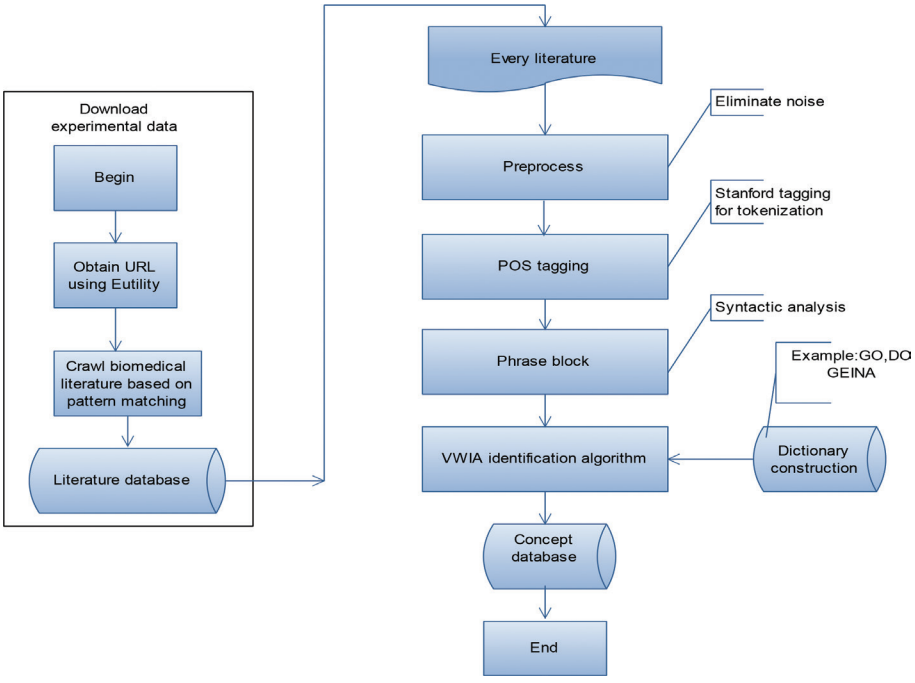

Figure 1. Pipeline of our approach to recognize biomedical concepts.

biomedical literature, the contents are achieved based on pattern mapping, and the crawled contents are stored in the local database to be further processed.

4.1.2. Preprocess, POS tagging, and phrase block

Aiming at the collected literature, to every biomedical literature, there may be several paragraphs, and each paragraph may include one or several sentences. The basic textual units identified by the tokenization as constituent tokens in each sentence are tagged by the Standford POS tagger [30]. It is a tool that appoints parts of speech to every word (for instance, verbs, adjectives, nouns, and so on) by reading text in some language with the implementation of log-linear-part-of-speech taggers. The results are shown in **Figure 2** aiming at an example sentence. Aiming at the features of biomedical concept, the POS-word pairs are important for the biomedical concepts. For example, nouns, adjectives, participles, and so on, could be components of some biomedical concepts, while some words, such as indefinite articles and verbs, would be omitted to the identification of the biomedical concepts. Thus, we extract phrase blocks combined by nouns, adjectives, and participles. These phrase blocks are preprocessed as a unified base form. For example, the word of tagging POS labels "NN" (single noun) and "NNS" (plural noun) transforms the word of single noun for the normalization. **Figure 2** shows an example about the POS tagging [25].

4.1.3. Biomedical dictionaries' construction

Considering the high precision of dictionary-based approach, the approach used it for the identification of the biomedical concepts. Thus, concept dictionary is built by some authorized biomedical ontologies, for example, Disease Ontology (DO), Gene Ontology (GO), and GENIA ontology. GENIA corpus is used to provide reference material for biotext mining. GENIA corpus is also semantically annotated corpus containing 2000 MEDLINE literature with almost 100,000 annotation information and more than 507,325 words including 18,546 sentences for biomedical terms in 3.02 version. Each article is encoded in an XML-based mark-up scheme with ID, title, and abstract in the corpus. Moreover, both abstracts and titles have been marked up for meaningful annotated terms, biologically and semantically. The corpus provides semantically annotated biological terms identifiable with any terminal concepts for GENIA ontology, For instances, "<cons lex = "IL-2_gene" sem = "G#DNA_domain_or_region" > IL-2 gene</cons>", the label "lex" describes the concepts, and the label "sem" represent the type of concept. Due to the structural scheme of GENIA ontology, it could be extracted by the regular expression. These extracted biologically meaningful terms build concept dictionary for the

IL-2/NN gene/NN expression/NN and/CC NF-kappa/NN B/NN activation/NN through/IN CD28/NNP requires/VBZ reactive/JJ oxygen/NN

Figure 2. Sentence with part-of-speech tags generated by Stanford maximum entropy part-of-speech tagger. Tags are NN: normal noun; IN: preposition or conjunction; VBZ: verb in present/past tense; JJ: adjective; CC: conjunction; NNP: proper noun.

biomedical text mining. For example, the concept taxonomy of GENIA 3.02 version contains some biologically relevant nominal categories as shown in **Table 1** [24].

4.1.4. Results

Biomedical texts are divided into a few sentences in biomedical literature, and for every sentence, some phrase blocks are parsed. The algorithm named Variable-step Window Identification Algorithm (VWIA) is developed for identifying biomedical concepts. The approach obtained the overall 95.0% F-measure aiming at the GENIA corpus. The implementation of the approach is shown in **Figure 3** [24].

4.2. Machine-learning approach

In this work [26, 27], we introduce machine learning to help recognize biomedical named entity. The pipeline architecture of our approach is shown in **Figure 4** [26, 27].

The pipeline of our system mainly contains four modules: preprocessing module, training module, tagging module, and testing module.

Categories	Numbers	Categories	Numbers
protein_molecule	21,632	peptide	524
protein_family_or_group	8372	body_part	449
DNA_domain_or_region	8054	atom	341
cell_type	7233	RNA_family_or_group	334
other_organic_compound	4096	polynucleotide	260
cell_line	3974	inorganic	256
protein_complex	2417	nucleotide	239
lipid	2359	mono_cell	222
virus	2126	other_artificial_source	209
multi_cell	1766	protein_substructure	129
DNA_family_or_group	1558	DNA_substructure	107
protein_domain_or_region	1017	protein_N/A	99
protein_subunit	920	carbohydrate	98
amino_acid_monomer	785	DNA_N/A	48
tissue	692	RNA_domain_or_region	39
cell_component	669	RNA_N/A	15
RNA_molecule	602	RNA_substructure	2
DNA_molecule	542	Total	72,185

Table 1. Biomedical concept categories and numbers in GENIA.

Figure 3. Visualization of the dictionary-based approach.

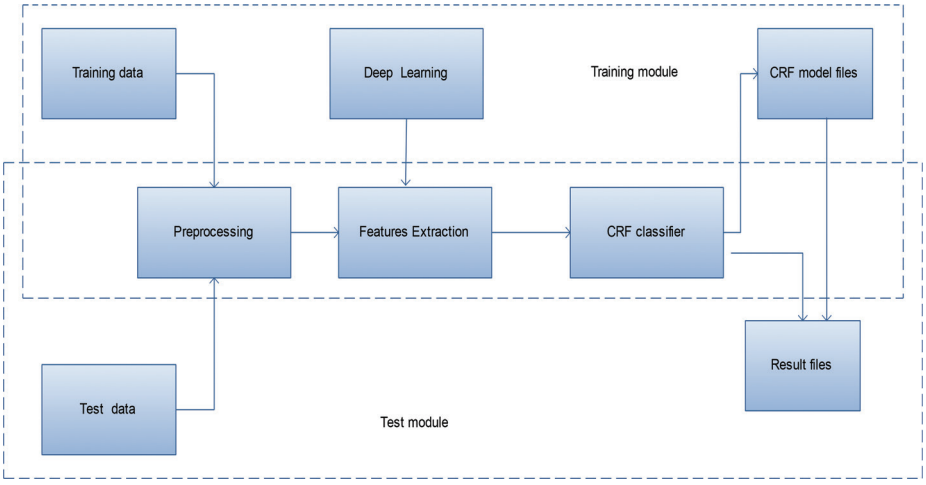

Figure 4. The pipeline architecture of machine-learning approach.

4.2.1. Preprocessing module

Applying machine-learning methods, the original biomedical texts could not directly be processed by the CRF model; they would be preprocessed to form the related formats. The original training, testing, and predicted texts would be processed to the specified format files with feature dataset.

4.2.2. Training module

Trained texts are put into the preprocessing module, aiming at the selected and extracted features to form the specified trained files that are trained with some parameters' set, which could obtain trained model files. The input of training module is the result of the preprocessed training texts, while the output of it is the model files including feature function set and weight parameter.

4.2.3. Tagging module

After the preprocessing module processes the test texts, the formed specified test files are put into a tagging module together with the model files, which could obtain the tagged files. The input of a tagging module is the result of preprocessed testing texts, while the output of it is the tagged result files.

4.2.4. Testing module

The testing module is to measure the performance of our system. After processing the test texts in the preprocessing module, the test standard files are put into a test module together with the tagging result files. The input of a tagging module is the results of preprocessed test files, while the output of it is the identified performance of our system including precision, recall, and F-measure.

4.2.5. Results

The approach considered features including POS features, word surface clue feature (uppercase, lowercase, numbers, specific char, initial) using the Genia corpus 3.02 version to train and test the system's performance with a 10-cross validation. The system's performance is shown in **Table 2** related to the six classes. According to the above method, using Java programming based on Linux OS to develop Biomedical entity recognition Miner system called (BerMiner). **Figure 5** shows the results of the identified biomedical entity.

Project	Precision (%)	Recall (%)	F-measure (%)
DNA	61.31	45.79	52.4
RNA	56.48	44.63	49.69
Protein	76.48	72.64	74.50
Cell-type	71.11	58.51	64.16
Cell-line	71.80	53.58	61.29
Virus	78.56	66.37	71.83

Table 2. Identified biomedical entity's performance based on machine learning [26, 27].

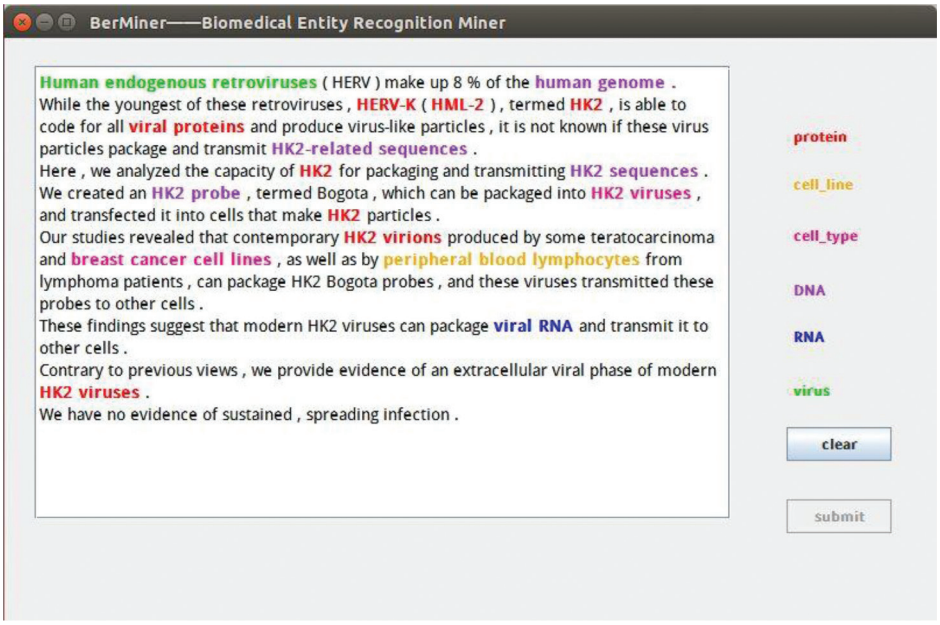

Figure 5. BerMiner extracted results with six categories using machine learning [26, 27].

5. Extraction of dynamic biomedical information

Biomedical entities produce a series of information interaction in the process of genetic information transfer and expression, such as genes and gene interaction relations, the relationship between genes and disease, the relationship between genes and gene product, molecular signal conduction pathways, and so on. The information is represented as the dynamic form. Dynamic biomedical information represents the process of activities of biomedical entities. Dynamic biomedical information is extracted, namely association between biomedical entities which is often extracted based on entity co-occurrence analysis with statistics theory. Glenisson et al. [28] proposed an approach to extract the relationship between genes using vector space method and k-medoids algorithm for gene clustering. Wren [29] explored the method to measure biological entities relationship using mutual information model. Wu et al. [30] researched the interactions between genes and drugs based on text-mining technology, which was divided into three steps. First, the approach identified genes and drugs entities from Medline abstract; the second step is to extract different levels of gene-2-drug pairs. The third step is to rank the gene-2-drug pairs based on mathematical statistical model. Our research group has done the similar works [31–33]. Here, we example the works [31, 33] for describing the process related to the extraction of dynamic biomedical information in detail.

5.1. Relationship extraction based on statistic model

Dynamic information represents the process of activities of biomedical entities. In this study, we focus on the dynamic process of biomedical entities, and extract dynamic biomedical information, namely association between biomedical entities, based on entity co-occurrence analysis which is attached to the statistics theory with more precision. Entity co-occurrence analysis considers that if any two entities occurred in a certain level of paper (e.g., a full text, a paragraph, a sentence, and a phrase), then the two entities could have be related. Different levels have different strengths of entity association. Through syntactic analysis of biomedical texts, the weight of association in phrase level is the highest than in other two levels, and the sentence level is higher than the full text. Due to the multiclass biomedical entities used in this study, extracted associations are also multiple types. We build a data modeling of entity association based on entity co-occurrence analysis with statistics. The data modeling could formally be represented as a three tuples as shown in Eq. (1)

$$D = (E, C, R) \tag{1}$$

Let E be a set of biomedical entities, and C be a set of types of association. Let R be a set of correlation of associations between biomedical entities.

Supposing $\varphi(e)$ represents entity category where $e \in E$; W is a set of weights of association levels where $w_k \in W$. To the two entities ($e_i, e_j \in E$) of the kth level, their co-occurrence frequency is represented as $f_k(e_i, e_j)$, and the correlation between two biomedical entities is defined as Eq. (2)

$$T = \left(e_i, e_j, C\left(\varphi(e_i), \varphi(e_j)\right), R(e_i, e_j)\right) \tag{2}$$

Let $C\left(\varphi(e_i), \varphi(e_j)\right)$ be association category related to the entity category $\varphi(e_i)$ and $\varphi(e_j)$. For instance, $C\left(\varphi(e_i), \varphi(e_j)\right)$ could represent association between gene and disease, or association between gene and microRNA, and so on. Let $R(e_i, e_j)$ be correlated factor between entity $e_i$and e_j as shown in Eq. (3)

$$R(e_i, e_j) = \sum_k w_k f_k(e_i, e_j) \tag{3}$$

After building data modeling of entity association, we further consider extracting these dynamic biomedical information from biomedical literature. Aiming at the entities identified, we design an algorithm of Mining Multiclass Entity Association, named (MMEA), under the data modeling based on co-occurrence statistical analysis as shown in **Figure 6**.

The input of MMEA algorithm is entities identified building on the step of entity recognition. The MMEA algorithm first gets the category of each entity (lines 4–6). Aiming at an entity e_i, the algorithm decides the types of associations between it and other entities (line 7) and computes the correlation factor $R(e_i, e_j)$by Eq. (3) (line 8). The above-achieved results are stored in the four tuples $T = \left(e_i, e_j, C\left(\varphi(e_i), \varphi(e_j)\right), R(e_i, e_j)\right)$for further processing (line 9). The process proceeds with the increment of i entity until all entity associations are obtained (lines 3–11).

1. Input: $E = \{e_1, e_2, ...e_m\}$;
2. Output: T;
3. for each i
4. $e_i \rightarrow \varphi(e_i)$;
5. for j=i+1 to m
6. $e_j \rightarrow \varphi(e_j)$;
7. $(\varphi(e_i), \varphi(e_j)) \rightarrow C(\varphi(e_i), \varphi(e_j))$;
8. Compute $R(e_i, e_j)$ by $\sum_k w_k f_k(e_i, e_j)$;
9. $T = (e_i, e_j, C(\varphi(e_i), \varphi(e_j)), R(e_i, e_j))$;
10. end;
11. end.

Figure 6. MMEA algorithm.

5.2. Relationship visualization

Biomedical text mining focuses on the centrality of user interactivity, and it needs to provide users for interacting with data results. The text-mining visualization facilitates user interactivity with graphical approaches. For example, we developed a circle network graph to allow

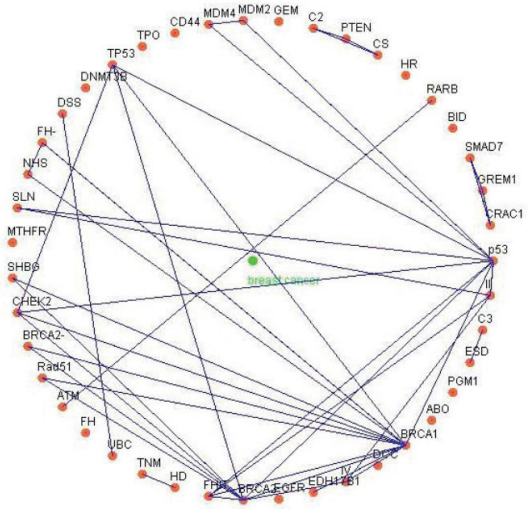

Figure 7. Relationship visualization between genes.

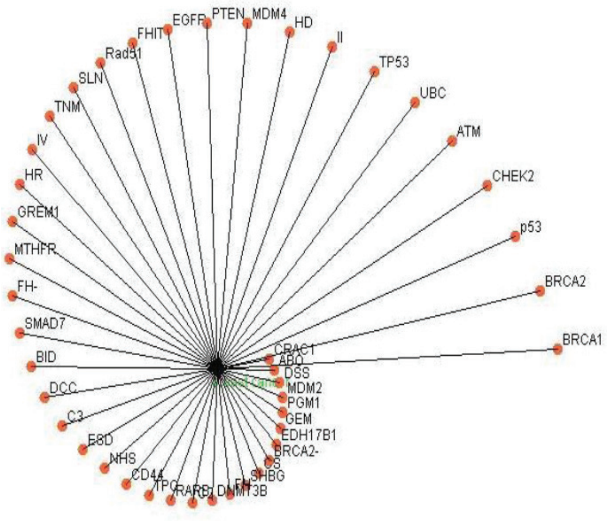

Figure 8. Relationship visualization between disease and gene.

disease researchers to explore the relationship between genes related to breast cancer. To capture the susceptive genes related to breast cancer, a fan-like network visualization is designed by our research group. The node represents the biomedical entities, and the link lines indicate the associations between entities (as shown in **Figures 7** and **8**).

6. Conclusions

This chapter first introduces the rise of biomedical text mining. Then, it describes the biomedical text-mining technology, namely natural language processing, including the several components. In the following sections, it emphasize the two aspects in biomedical text mining involving static biomedical information recognization and dynamic biomedical information extraction using instance analysis which our previous works.

Acknowledgements

This research is supported by the National Natural Science Foundation of China (Grant Nos: 61502243, 61502247, 61272084, 61300240, 61572263, 61502251, and 61503195), Natural Science Foundation of the Jiangsu Province (Grant Nos: BK20130417, BK20150863, BK20140895, and BK20140875), China Postdoctoral Science Foundation (Grant No. 2016M590483), Jiangsu Province postdoctoral Science Foundation (Grant No. 1501072B), Scientific and Technological Support Project (Society) of Jiangsu Province (Grant No. BE2016776), Nanjing University of Posts and Telecommunications' Science Foundation (Grant Nos: NY214068 and NY213088). This work is also supported in part by Zhejiang Engineering Research Center of Intelligent Medicine (2016E10011).

Conflict of interest

In this chapter, the instances are from our previous works [24–27, 31–33].

Author details

Lejun Gong

Address all correspondence to: glj98226@163.com

Jiangsu Key Laboratory of Big Data Security and Intelligent Processing, School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing, China

References

[1] Zheng S, Lu JJ, Ghasemzadeh N, Hayek SS, Quyyumi AA, Wang F. Effective information extraction framework for heterogeneous clinical reports using online machine learning and controlled vocabularies. JMIR Medical Informatics. 2017;**5**(2):e12. DOI: 10.2196/medinform.7235

[2] Weng WH, Wagholikar KB, McCray AT, Szolovits P, Chueh HC. Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. BMC Medical Informatics and Decision Making. 2017;**17**(1):155. DOI: 10.1186/s12911-017-0556-8

[3] Basaldella M, Furrer L, Tasso C, Rinaldi F. Entity recognition in the biomedical domain using a hybrid approach. Journal of Biomedical Semantics. 2017;**8**(1):51. DOI: 10.1186/s13326-017-0157-6

[4] Krallinger M, Leitner F, Vazquez M, Salgado D, Marcelle C, Tyers M, Valencia A, Chatr-aryamontri A. How to link ontologies and protein-protein interactions to literature: Text-mining approaches and the BioCreative experience. Database: The Journal of Biological Databases and Curation. 2012;**2012**:bas017

[5] Gene Ontology Consortium. Gene ontology consortium: Going forward. Nucleic Acids Research. 2015;**43**(Database issue):D1049-D1056

[6] Kim JD, Ohta T, Tateisi Y, Tsujii J. GENIA corpus – semantically annotated corpus for bio-textmining. Bioinformatics. 2003;**19**(Suppl 1):i180-i182

[7] Natale DA, Arighi CN, Blake JA, et al. Protein ontology (PRO): Enhancing and scaling up the representation of protein entities. Nucleic Acids Research. 2017;**45**(D1):D339-D346

[8] Kim JJ, Park JC. Bioie: Retargetable information extraction and ontological annotation of biological interactions from the literature. Journal of Bioinformatics and Computational Biology. 2004;**2**(3):551-568

[9] Cohen AM, Hersh WR. A survey of current work in biomedical text mining. Briefings in Bioinformatics. 2005;**6**(1):57-71. Review

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

[11] Luscombe NM, Greenbaum D, Gerstein M. What is bioinformatics? A proposed definition and overview of the field. Methods of Information in Medicine. 2001;**40**(4):346-358

[12] Krallinger M, Erhardt RA, Valencia A. Text-mining approaches in molecular biology and biomedicine. Drug Discovery Today. 2005;**10**(6):439-445 Review

[13] Cronin RM, Fabbri D, Denny JC, Rosenbloom ST, Jackson GP. A comparison of rule-based and machine learning approaches for classifying patient portal messages. International Journal of Medical Informatics. 2017 Sep;**105**:110-120

[14] Leitner F, Mardis SA, Krallinger M, Cesareni G, Hirschman LA, Valencia A. An overview of BioCreative II.5. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2010;**7**(3):385-399

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

[16] Rebholz-Schuhmann D, Kirsch H, Couto F. Facts from text–is text mining ready to deliver? PLoS Biology. 2005;**3**(2):e65

[17] Barthelmes J, Ebeling C, Chang A, et al. BRENDA, AMENDA and FRENDA: The enzyme information system in 2007. Nucleic Acids Research. 2007;**35**(Database issue):D511-D514

[18] Tanabe L, Wilbur WJ. Tagging gene and protein names in biomedical text. Bioinformatics. 2002;**18**(8):1124-1132

[19] Tanabe L, Wilbur WJ. Generation of a large gene/protein lexicon by morphological pattern analysis. Journal of Bioinformatics and Computational Biology. 2004;**1**(4):611-626

[20] Settles B. ABNER: An open source tool for automatically tagging genes, proteins and other entity names in text. Bioinformatics. 2005;**21**(14):3191-3192

[21] Mitsumori T, Fation S, Murata M, et al. Gene/protein name recognition based on support vector machine using dictionary as features. BMC Bioinformatics. 2005;**6**(Suppl 1):S8

[22] Saha SK, Sarkar S, Mitra P. Feature selection techniques for maximum entropy based biomedical named entity recognition. Journal of Biomedical Informatics. 2009;**42**(5):905-911

[23] Li L, Zhou R, Huang D. Two-phase biomedical named entity recognition using CRFs. Computational Biology and Chemistry. 2009;**33**(4):334-338

[24] Gong L, Yang R, Liu Q, Dong Z, Chen H, Yang G. A dictionary-based approach for identifying biomedical concepts. International Journal of Pattern Recognition and Artificial Intelligence. 2017;**31**(9):1757004. http://dx.doi.org/10.1142/S021800141757004X

[25] Gong L, Sun X. ATRMiner: A System for Automatic Biomedical Named Entities Recognition. Yantai: ICNC 2010; 2010. pp. 3842-3845

[26] Gong LJ, Yang RG, Yang HY, Jiang KY, Yang G. BerMiner: A machine learning system for identifying bio-entity. 2015 International Conference on Software Engineering and Information System (SEIS 2015); 2015:447-450

[27] Yang RG, Wu ZX, Yang Z, Yang G, Gong LJ. Identifying biomedical entity based on deep learning. 2015 International Conference on Software Engineering and Information System (SEIS 2015); 2015:713-718

[28] Glenisson P, Coessens B, Van Vooren S, et al. TXTGate: Profiling gene groups with text-based information. Genome Biology. 2004;**5**(6):R43

[29] Wren JD. Extending the mutual information measure to rank inferred literature relationships. BMC Bioinformatics. 2004;**5**:145

[30] Wu Y, Liu M, Zheng WJ, et al. Ranking gene-drug relationships in biomedical literature using latent dirichlet allocation. Pacific Symposium on Biocomputing. 2012:422-433

[31] Gong LJ, Yang RG, Dong ZJ, Chen H, Yang G. Extraction of disease-centred dynamic biomedical information from literature. Journal of Computational and Theoretical Nanoscience. 2015;**13**(1):722-727 12

[32] Gong LJ, Yang RG, Sun X. Prioritization of disease susceptibility genes using LSM/SVD. IEEE Transactions on Biomedical Engineering. 2013;**60**(12):3410-3417

[33] Gong LJ, Wei YB, Xie JM, Yuan ZD, Sun X. Text mining approach for relationships between genes and diseases. Dongnan Daxue Xuebao (Ziran Kexue Ban)/Journal of Southeast University (Natural Science Edition). 2010;**40**(3):486-490