A Survey on Biomedical Named Entity Recognition and Normalization

Ruoyao Ding* and Huaxing Chen

1 School of Information Science and Technology, Guangdong University of Foreign Studies, Guangdong, P.R. China

Abstract

With a rapidly growing amount of biomedical information available only in textual form, there is considerable interest in applying NLP techniques to extract such information from the biomedical literature. Much of the research has paid special attention to extracting information about biomedical named entities. In this paper, we conducted a survey on biomedical named entity recognition and normalization, focusing on gene mention recognition and normalization. We believe this can help researchers to find work of their interest and interpret their own research.

Keywords

Recognition; Normalization; Text Mining

1 Introduction

Biomedical researchers usually describe their experimental results in research publications. With the rapid growth of biomedical publications, the information of interest needs to be extracted automatically to avoid the time consuming and labor intensive process.

Named entity recognition and normalization are two common tasks in the biomedical text mining field. Together they provide a means to extract the unstructured information buried in the literature and put the extracted information to structured form. There already has been some work on the survey of biomedical named entity recognition and normalization. However, a more comprehensive and most updated version is still needed. In this paper, we conduct a survey to present current work on biomedical named entity recognition and normalization. Given the primary importance genes and their products play in biological and medical studies, this survey will focus on gene mention recognition and normalization. We hope this can help researchers in biomedical text mining field to find the information of their interest and interpret their own research.

2. Gene Mention Recognition

The task of gene mention (GM) recognition is to automatically recognize gene/protein names mentioned in text. This task has received wide attention, and has been used in several challenge evaluations such as BioCreative I [1] and BioCreative II [2]. Other annotated corpora have also been constructed for system development and evaluation purpose.

There are several challenges of the gene mention recognition task:

(1) **No. of genes:** The number of gene names is in the millions and new names are created continuously.

(2) **Name variations:** Authors usually do not use proposed standardized gene names.

(3) **Polysemy:** Gene names often also refer to other entities such as disease names.

2.1 Gene Mention Recognition systems

Approaches to gene mention recognition can be categorized into two major classes: rule-based approaches and machine learning-based approaches.

While rule-based gene mention recognition approaches do not require annotated data to train a system, they do require domain experts to be closely involved in developing the rules. The following three systems are examples of gene mention detectors that rely on manually developed rules.

Hanisch et al. [3] presented a dictionary matching based system that detects fly, mouse and yeast gene names from biomedical text. Fukuda et al. [4] proposed a method which incorporates two new concepts called c-term (a concept based on orthography) and f-term (a concept that is based on terms that correspond to types of biological entities) (details about those two terms will be introduced later in the Gene Normalization chapter). Narayanaswamy et al. [5] developed a system which extracts multiple types of named entities including gene names. Their system is based on a manually developed set of rules that
rly upon some crucial lexical information, linguistic constraints of English, and contextual information and develop the notion of c-term and f-term in named entity recognition.

The machine learning-based gene mention recognition approaches require annotated data to train a system. Thus, domain expertise is now required in the development of the data annotation and less during the system training.

In the machine learning-based gene mention recognition approaches, the gene mention recognition task is often treated as a sequence labelling problem (label the tokens in the text using the tags). BIO (or IO) tags for the text are commonly used to represent the boundaries of gene mentions where B represents the beginning of the gene name in text, I is assigned to a token inside the gene mention and O is assigned to tokens that are outside the gene mentions.

Among the machine-learning based systems, Banner [6] is widely used for recognizing biomedical named entities including gene mentions. It is based on conditional random fields and applied orthographic, morphological and shallow syntax features. Liu et al. [7] trained a classification system using Conditional random field (CRF) [8] to classify each word in the literature to the BIO tags. They applied BioThesaurus [9], a comprehensive collection of gene names to entries in the UniProt Knowledgebase, for dictionary lookup and used the matching information as a feature. Huang et al. [10] considered the gene mention task as a classification problem and applied support vector machine (SVM) [11] to solve it. Chen et al. [12] proposed a gene mention recognition system for biomedical literature using a dictionary and Support Vector Machine. Zhou et al. [13] proposed an ensemble of classifiers for gene mention recognition. They combined three classifiers, one Support Vector Machine and two discriminative Hidden Markov Models using a simple majority voting strategy. Other machine learning based gene mention systems can be found in [14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24].

2.2 Gene Mention Recognition Corpora

High quality gene mention corpora are important for the development of any type of gene mention recognition system. Even for the rule-based system, more accurate rules can be made by analyzing the instances in the corpora.

The GENIA corpus [25] is a collection of 2000 abstracts extracted from Medline database. Multiple biomedical named entities, including gene names, are annotated. It is focused on a subset of human hematology. The PennBioIE corpus [26] consists of 1414 Medline abstracts on cancer. 24 types of biomedical named entities, including gene names, are annotated. The BioCreative 1 GM corpus [1] contains 15,000 sentences from Medline abstracts. Genes and related entities mentions are annotated. The BioCreative 2 GM corpus [2] contains 20,000 sentences from Medline abstracts (15,000 of which were used previously in BioCreative 1).

3. Gene Normalization

The task of gene normalization (GN) is to automatically link a gene mention to a database entry for the gene (product). Other than the challenges stated in the gene mention recognition task, the challenges for the gene normalization task also include:

1. identifying the species for the gene mentions since most gene (product) knowledge bases contain species-specific entries, and
2. disambiguation since multiple gene entries may share the same short name (symbol).

3.1 Gene Normalization Systems

The following were the top performing systems in the BioCreative I [27] and BioCreative II [28] Challenge GN Tasks. ProMiner [29] is a dictionary-based GN system which is characterized by the inclusion of different biomedical dictionaries and manual clean-up of a dictionary. BioTagger [7] tackles the GN problem with the steps: (1) dictionary lookup to obtain a list of mapping pairs of gene mention and database identifier, (2) machine learning that considers features such as the gene mention recognition, name ambiguity, and token shape information, and (3) a similarity based method to associate Entrez gene records with phrases detected by the gene mention tagger. GNAT [30] is a GN system encompassing four steps: named entity recognition for genes and species, validation of gene mentions, correlating gene mentions with species, and finally gene mention disambiguation. GeNo [31] tackles the GN problem by employing a carefully crafted suite of symbolic and statistical methods.

In BioCreative III [32], the GN task was further extended to cover genes of all relevant species in the literature corpora. Among the systems, Bhattacharya et al. [33] tried to associate a species name with a gene name by considering their proximity to the gene mention. Dai et al. [34] employed a multistage GN procedure and selected dictionary entries from only the top 22 most common species in NCBI (from 7283 species) to speed up the GN process. A document-level gene normalization system, called GeneTUKit [35], employed features from the local context as well as the global context of the whole full-text article. GenNorm [36] follows three steps: gene name recognition, species assignment, and species-specific gene normalization, and uses SR4GN [37] for assigning species to gene mentions. GenNorm has been widely used in text mining systems that require GN, such as in PubTator [38] and in an event extraction pipeline [39]. GNNormPlus [40], as an updated version of GenNorm, refined the gene mention process by training the mention recognizer on a new corpus with gene, gene family and protein domain annotations. It also integrates several advanced text mining techniques, including SimConcept for resolving composite gene names.

3.2 Gene Normalization Corpora

High quality gene normalization corpora are important for the development of any type of gene normalization system.

EJBI – Volume 15 (2019), Issue 2
The BioCreative I gene normalization corpus [27] and the BioCreative II gene normalization corpus [28] focused on the GN task for yeast, fly, and mouse genes and human genes respectively. Both of these corpora annotate gene mentions found in abstracts. In contrast, the BioCreative III gene normalization corpus [33] annotates full length articles and is not limited to specific species.

The BioCreative I gene normalization corpus consists 15,000 abstracts for training, 468 abstracts for developing, and 750 abstracts for testing. All these abstracts are annotated in abstract level, not mention level (only a list of database identifiers is given for each abstract, without any location information). No corresponding gene name in the abstracts for the database identifier is provided in this corpus. The BioCreative II gene normalization corpus consists 281 abstracts for training, and 262 abstracts for testing. All these abstracts are also annotated in abstract level, but the corresponding gene names in the abstracts are given for each database identifier. The BioCreative III gene normalization corpus consists 32 fully annotated articles and 500 partially annotated articles for training. For testing, it provides 50 articles as gold standard and 507 articles as silver standard.

4. Other Biomedical Named Entity Recognition and Normalization

There has been considerable interest in the detection and normalization of other types of biomedical entities such as diseases, chemical compounds and drugs.

4.1 Other Biomedical Named Entity Recognition

tmChem [41] is a chemical named entity recognition system created by combining two Conditional random field (CRF) models in an ensemble. The two models in the system used different tokenization methods, feature sets, CRF implementations, CRF parameters. Lu et al. [42] developed a chemical named entity recognition system based on mixed CRFs with word clustering. Lowe et al. [43] proposed a system for chemical entity recognition based on grammar and dictionary. Their system uses a mixture of expertly curated grammars and dictionaries, as well as dictionaries automatically derived from public resources.

Chowdhury et al. [44] presented a CRF based approach for disease mention recognition. The features they used include disease specific contextual features, orthographic features, general linguistic features, syntactic dependency features and dictionary lookup features. Kaewphan et al. [45] developed a system for disease mention recognition. Their system was based on an existing named entity system, NERsuite, supplemented with UMLS dictionary features.

Other biomedical named entity recognition systems can be found in [46, 47, 48, 49, 50].

4.2 Other Biomedical Named Entity Normalization

Leaman et al. [41] paired their chemical named entity recognition system with a dictionary approach for normalization. They used a dictionary of chemical entities and their names that was collected from MeSH and ChEBI. DNORM [51] is a disease normalization system, which uses a linear model to score the similarity between mentions and concept names. DNORM has an interesting approach of learning term variation directly from training data. Kaewphan et al. [45] developed a disease normalization system, which was based on their disease mention system. They combined compositional word vector representations with CRF to map the recognized mentions to the UMLS concepts. Other biomedical named entity normalization works can be found in [52, 53, 54, 55].

5. Discussion

1.1 Biomedical Named Entity Recognition based on Deep Learning

In recent years, deep learning has drawn much attention in biomedical named entity recognition. Hence, we will next describe some novel work of biomedical named entity recognition in the last three years based on deep learning. Hemati et al. [56] combined Long Short Term Memory (LSTM) neural networks and CRF to detect drug named entity, and achieved state-of-the-art performance. Korvigo et al. [57] first used Convolutional Neural Network (CNN) to encode the text, then applied Recurrent Neural Networks (RNN) to recognize drug named entity. Xu et al. [58] constructed a LSTM+CRF network to tackle the task of disease named entity recognition. Zhao et al. [59] developed a neural network based on CNN to recognize disease mention. Zhang et al. [60] constructed a network using LSTM+CRF structure to recognize the named entities in electronic health records. Zhu et al. [61] used n-gram and context as input of CNN to detect named entities in biomedical text. Luo et al. [62] developed an attention based bidirectional-LSTM+CRF model to recognize drug named entity. Similarly, Habibi et al. [63] compared LSTM+CRF with pure CRF model, and shown that deep learnin model outperformed traditional machine learning model in the tasks of recognizing gene mention, chemical mention, species mention, and disease mention. Lyv et al. [64] constructed three models based on RNN, RNN+CRF, and BiLSTM+RNN. The three models were compared in the task of gene mention recognition. Experimental results shown BiLSTM+RNN model outperformed the other two.

5. Conclusion

Named entity recognition and normalization are tasks to recognize entities mentioned in natural language text and link them to database IDs. We have conducted a survey of works related to biomedical named entity recognition and normalization, focusing on gene mention recognition and normalization. We believe this work will assist researchers to find the information of their interest and interpret their own research. In the further, we plan to conduct another study on biomedical named entity relation extraction.
References


24. Tsai RT, Sung CL, Dai HJ, Hung HC, Sung TY, Hsu WL. NERBio: using selected word conjunctions, term normalization, and global patterns to improve biomedical


