A Short Survey of Biomedical Relation Extraction Techniques

Elham Shahab
Department of Computer Engineering
Islamic Azad University, Yazd Branch
ma.shahab@iauyazd.ac.ir

ABSTRACT
Biomedical information is growing rapidly in the recent years and retrieving useful data through information extraction system is getting more attention. In the current research, we focus on different aspects of relation extraction techniques in biomedical domain and briefly describe the state-of-the-art for relation extraction between a variety of biological elements.

KEYWORDS
Information Extraction, Biomedical text mining, Relation Extraction.

ACM Reference format:
Elham Shahab. 2017. A Short Survey of Biomedical Relation Extraction Techniques.

1 INTRODUCTION
Biomedical literature is growing rapidly, Cohen and Hunter in [17] explain how the growth in PubMed/MEDLINE publications is phenomenal, which makes it a potential area of research with respect to information and data mining techniques. In fact, it is quite difficult for biomedical scientists to adjust new publications and come up with relevant publications in their own research area. To address this, text mining and knowledge discovery is getting more attention these days in biomedical sciences. In fact, automated text processing methods try to overcome text information overload and transform those data into machine understandable format. Text mining and knowledge extraction techniques along with statistical machine learning algorithms are widely used in medical and biomedical domain such as [45, 64]. In particular, text mining methods have been applied in a variety of biomedical branches and domains such as gene clustering, protein structure prediction, clinical diagnosis, biomedical hypothesis and etc. In this section, we briefly describe some of the relevant research in biomedical domain and explain some of the state-of-the-art relation extraction techniques with respect to data mining approaches in biomedical discipline.

2 RELATION EXTRACTION
Determining the relationships among biomedical entities is the key point in relation extraction in Biomedical domain. The ultimate goal is to locate the occurrence of a specific relationship type between given two entities. There are lots of extraction format available in biomedical domain such as RDF [27, 41] and XML format [9, 19, 44] which is widely used. For instance, in the genomic area, extracting interactions between genes and proteins such as gene-diseases or protein-protein relationships is very important and getting more attention these days. Relation extraction is usually integrated with the similar challenges as NER, such as creation of high quality annotated data for training and assessing the performance of relation extraction systems. There are different text mining techniques [4] such as topic modeling [2, 3], information extraction [59, 62], text summarization [5], and clustering [4, 23] for relation extraction between some of the different types of biological elements such as genes, proteins and diseases that will be discussed in the following sections.

3 BIOMEDICAL RELATION EXTRACTION TECHNIQUES
Knowledge and Information extraction and in particular relation extraction tasks have widely studied various biomedical relations. There are lots of ongoing research in biomedical relation extraction due to critical roles of genes and proteins interactions in different biological processes. Many different approaches for biomedical relation extraction have been proposed which can be a simple systems that only rely on co-occurrence statistics to complex ones which use syntactic analysis and dependency parse trees. The entities co-occur based technique is considered as a the most straightforward technique which is based on this fact that if they mentioned together more frequently, there is a chance that they might be related together in some way. For example, Chen et al. [15] introduce a co-occurrence statistics method to calculate and evaluate the degree of association between disease and relevant drugs from clinical narratives and biomedical literature. An other approach in this area is Rule-based approaches. In this technique a set of methods used for biomedical relation extraction. Usually, rules are defined manually by domain experts [54] or automatically generated by using machine learning methods [28] from an annotated corpus. Hakenberg et al. [28] define and extract syntactical patterns learned from labeled examples and match them against arbitrary text to detect protein-protein interactions. Classification-based techniques are also widely used methods for relation extractions in biomedical domain [59]. For example, Rink et al. [52] identify a set of features from multiple knowledge sources such WordNet and Wikipedia. In the next phase train and then apply a supervised machine learning technique, Support Vector Machine (SVM), to extract the relations between medical records and treatments. In addition, Bundschus et al. [12] have applied a supervised
machine learning method that detects and classifies relations between diseases and treatments extracted from PubMed abstracts and between genes and diseases in human GenerIF database.

Relation extraction methods improved fundamentally by considering the syntactic and semantic structures. Specifically, syntactic parsing methods, including dependency trees (or graphs) are able to produce syntactic information about the biomedical text which reveals grammatical relations between words or phrases. For example, Miyao et al. [40] conducted a comparative of several state of the art syntactic parsing methods, including dependency parsing, phrase structure parsing and deep parsing to extract protein-protein interactions (PPI) from MEDLINE abstracts.

Having faced the increasing growth of biomedical data, many approaches utilized machine learning techniques to extract useful information from syntactic structures rather than applying manually derived patterns [55]. Airola et al. [1] propose an all-path graph kernel to calculate the similarity between dependency graphs, and then use the kernel function to train a support vector machine to detect protein-protein interactions. Miwa et al. [39] describe a method to combine kernels and syntactic parsers for PPI extraction. Furthermore, Kim et al. [32] introduce four genic relation extraction kernels defined on the shortest dependency path between two named entities.

Semantic role labeling (SRL), a natural language processing technique that identifies the semantic roles of these words or phrases in sentences and expresses them as predicate-argument structures, is also useful when it is complemented with syntactic analysis. [57, 60] are examples which have used SRL.

In the following, we describe some of works done for relation extraction between a variety of biological elements.

### 3.1 Gene-Disease

Chun et al. [16] describe a classification-based approach for relation extraction. First they use a dictionary-based longest matching technique which extracts all the sentences that include at least one pair of gene and disease names. Then, they apply a Maximum Entropy-based NER to filter out false positives produced in previous stages. They reach the precision of 79% and recall of 87% which significantly outperforms previous methods. Bundschus et al. [12] also propose a classification-based method, Conditional Random Field (CRF), to identify and classify relations between diseases and treatments and relations between genes and diseases. Their system utilizes supervised machine learning, syntactic and semantic features of context. For more information, see [10, 22, 51, 61].

### 3.2 Gene-Protein

Fundel et al. [25] use Stanford Lexicalized Parser to create dependency parse trees from MEDLINE abstracts and complement this information with gene and protein names obtained from ProMiner NER system [29]. Then the system applies a few different relation extraction rules to identify gene-protein and protein-protein interactions. They achieved better precision and F-measure and significantly outperformed previous approaches. Saric et al. [54] present a rule-based method to extract gene-protein relations. They integrate NLP techniques to preprocess and recognize named entities (e.g. genes and proteins), then apply a separate grammar module, combining syntactic properties and semantic properties of the relevant verbs, to extract relations. Some other works include [18, 34].

### 3.3 Protein-Protein

Raja et al. [47] introduce a system called PPInterFinder to extract human protein-protein interactions (PPI) from MEDLINE abstracts. PPInterFinder integrates NLP techniques (Tregex for relation keyword matching) and a set of rules to identify PPI pair candidates and then apply a pattern matching algorithm for PPI relation extraction. [38] presents a statistical unsupervised method, called BioNoculars. BioNoculars uses a graph-based method to construct extraction patterns for extracting protein-protein interactions. [58] performs a comprehensive benchmarking of nine different methods for PPI extraction that utilizes convolution kernels and confirms that kernels using dependency trees generally outperform kernels based on syntact trees. Similarly, [65] describes various methods for PPI extractions. For more approaches, see [1, 11, 33, 53].

### 3.4 Protein-Point mutation

The problem of point mutation extraction is to link the point mutation with its related protein and organisms of origin. Lee et al. [35] introduce Mutation GraB (Graph Bigram), that detects, extracts and verifies point mutation from biomedical literature. They test their method on 589 articles explaining point mutations from the G protein-coupled receptor (GPCR), tyrosine kinase, and ion channel protein families, and achieve the F-score of 79%, 72% and 76% for the GPCRs, protein tyrosine kinases and ion channel transporters respectively.

A few other algorithms have been developed for point mutation extraction. Rebholz-Schuhmann et al. [50] introduce a method called MEMA that scans MEDLINE abstracts for mutations. Baker and Witte [7, 8, 63] describe a method called Mutation Miner that integrates point mutation extraction into a protein structure visualization application. [30] presented MuteXt, a point mutation extraction method applied to G protein-coupled receptor (GPCR) and nuclear hormone receptor literature. [20] describes a automatic method for cancer and other disease-related point mutations from biomedical text.

### 3.5 Protein-Binding site

Ravikumar et al. [49] propose a rule-based method for automatic extraction of protein-specific residue from the biomedical literature. They use linguistic patterns for identifying residues in text and then apply a graph-based method (sub-graph matching [37]) to learn syntactic patterns corresponding to protein-residue pairs. They achieved a F-score of 84% on an automatically created dataset and 79% on a manually annotated corpus and outperforms previous methods. Chang et al. [14] describe an automatic mechanism to extract structural templates of protein binding sites from the Protein Data Bank (PDB). For more information about binding of other ligands to proteins, see [36].
3.6 Other Types of Interactions

Recently, there has been an increasing attention to the more complex task of identifying of nested chain of interactions (i.e. event extractions) rather than identifying binary relations. Because biomedical events are usually complex, effective event extraction normally needs extensive analysis sentence structure. Deep parsing methods and semantic processing are specially very helpful due to the capability of examining both syntactic as well as semantic structures of the biomedical text. For an overview of the currently available methods, see [6].

Event extraction has started to be widely used for annotation of biomedical pathways, Gene Ontology annotation and the enhancement of biomedical databases [55]. For example, [24] presents a NLP-based system, GENIES, to extract molecular pathways from biomedical literature.

There are several corpora in the biomedical domain that have integrated event annotations such as BioInfer corpus [46]. GENIA Event Corpus [31] and the Gene Regulation Event Corpus [57] are other annotated event corpora which are widely employed in biomedical text mining. For a comprehensive overview of the biomedical event extraction and evaluation, see [6, 55].

In addition, there are studies for identifying drug-drug interaction (DDI) in biomedical text. DDI can occur when two drugs interact with the same gene. Percha et al. [42] use a NLP technique [18] to identify and extract gene-drug interactions and propose a machine learning technique to predict DDIs. Some other works for DDI are [26, 43, 56].

4 DISCUSSION

Although relation extraction between various biological elements (e.g. genes, proteins and diseases) from biomedical literature has attained extensive attention recently, yet these text mining techniques have not been applied to extract relations between other types of molecules, particularly complex macromolecules to these important biological processes (e.g. glycan-protein interactions). The potential reasons of why extracting carbohydrate-binding proteins relationship from biomedical text have almost remained untouched, are as follows:

(1) Raman et al. [48] explains that the progress of glycomics has coped with distinctive challenges for developing analytical and biochemical tools to investigate glycan structure-function relations compared to genomics or proteomics. Glycans are more varied in terms of chemical structure and information density than DNA and proteins. In other words, carbohydrate-binding proteins are greatly heterogeneous in terms of their sequences, structures, binding sites and evolutionary histories [21]. This complicates the development of analytical techniques to accurately define the structure of glycans which accordingly makes the investigation and understanding of glycan-protein relations difficult. Therefore, the amount of knowledge in this domain is not comparable to genomic area where, it has led to less concentration on this field.

(2) In comparison with genomic area, the glycan-related knowledge bases (e.g. ontologies, databases, etc) which can be used as background knowledge to analyze the biomedical literature for information extraction is very restricted in terms of quantities and qualities. As we explained before (section 1 and 3), there exist many different ontologies and corpora about genes, diseases and proteins which are widely used in text mining, but there are barely a few ones for glycomics research. For example, UniCarbKB\(^1\) is a knowledge base and a framework that includes structural, experimental and functional data about glycomic experiments. Consortium for Functional Glycomics (CFG)\(^2\), funded by US National Institute of General Medical Sciences, is another collaborative effort which facilitates access to databases and services about glycomic research. However, according to previous reason, the algorithms used to automatically produce glycan structures are laborious and not quite accurate which may result in lower quality information whereas the databases and ontologies in genomic area contain curated data. Also, the amount of knowledge in glycomics research area is extremely small in terms of number of concepts and relations and instances in ontologies and/or the volume of data in databases as opposed to the fairly rich ontologies about genes, proteins and diseases [13].

5 CONCLUSION

Nonetheless, glycoproteomics is an emerging research area and there are many interesting future directions regarding information extraction and knowledge discovery in this domain. Glycoproteomics literature is barely touched by text mining community (due to aforementioned reasons), thus, there is a great demand for creating curated and high quality ontologies for glycoscience information. As we mentioned, UniCarbKB is an example of such systems. Even though, UniCarbKB provides critical information, it really is a database, not an ontology. Additionally, it does not contain a large amount of information. However, UniCarbKB research group has recently started to represent the data in RDF to unify the content and also begun to extend it to encompass more knowledge [13].

Another interesting direction is not only to create ontologies, but also to integrate them to invaluable existing ontologies in genomic area and linked open data which is very beneficial, because: 1) Although different ontologies contain different set of concepts, they have inter-relations to each other. This makes new interesting discoveries of hypotheses as well as relation extractions where it would not be possible using ontologies individually. 2) It facilitates the development of various applications for knowledge discovery (e.g. faceted browsing, data visualization, etc) in this domain. There are other interesting research directions in the area of knowledge discovery from glycoproteomics literature, and our propositions are barely scratching the surface.

REFERENCES

[1] Antti Airola, Sampo Pyysalo, Jari Björne, Tapio Pahikkala, Filip Ginter, and Tapio Salakoski. 2008. A graph kernel for protein-protein interaction extraction. In Proceedings of the workshop on current trends in biomedical natural language processing. Association for Computational Linguistics, 1–9.

\(^1\)http://www unicarckb org/

\(^2\)http://www functionalglycomics org/
A Short Survey of Biomedical Relation Extraction Techniques

[50] Dietrich Rebholz-Schuhmann, Stephane Marcel, Sylvie Albert, Ralf Tolle, Georg Casari, and Harald Kirsch. 2004. Automatic extraction of mutations from Medline and cross-validation with OMIM. Nucleic Acids Research 32, 1 (2004), 135–142.

[51] Thomas C. Rindflesch, Bisharah Libbus, Dimitar Hristovski, Alan R Aronson, and Halil Kilicoglu. 2003. Semantic relations asserting the etiology of genetic diseases. In AMIA Annual Symposium Proceedings, Vol. 2003. American Medical Informatics Association, 554.

[52] Bryan Rink, Sanda Harabagiu, and Kirk Roberts. 2011. Automatic extraction of relations between medical concepts in clinical texts. Journal of the American Medical Informatics Association 18, 5 (2011), 594–600.

[53] Barbara Rosario and Marti A Hearst. 2005. Multi-way relation classification: application to protein-protein interactions. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 732–739.

[54] Jasmin Šarić, Lars Juhl Jensen, Rossitza Ouzounova, Isabel Rojas, and Peer Bork. 2006. Extraction of regulatory gene/protein networks from Medline. Bioinformatics 22, 6 (2006), 645–650.

[55] Matthew S Simpson and Dina Demner-Fushman. 2012. Biomedical text mining: A survey of recent progress. In Mining Text Data. Springer, 465–517.

[56] Luis Tari, Saadat Anwar, Shanshan Liang, James Cai, and Chitta Baral. 2010. Discovering drug–drug interactions: a text-mining and reasoning approach based on properties of drug metabolism. Bioinformatics 26, 18 (2010), i547–i553.

[57] Paul Thompson, Syed A Iqbal, John McNaught, and Sophia Ananiadou. 2009. Construction of an annotated corpus to support biomedical information extraction. BMC bioinformatics 10, 1 (2009), 349.

[58] Domonkos Tikk, Philippe Thomas, Peter Palaga, Jörg Hakenberg, and Ulf Leser. 2010. A comprehensive benchmark of kernel methods to extract protein–protein interactions from literature. PLoS computational biology 6, 7 (2010), e1000837.

[59] E. D. Trippe, J. B. Aguilar, Y. H. Yan, M. V. Nural, J. A. Brady, M. Assefi, S. Safaei, M. Allahyari, S. Pouziyeh, M. R. Galinski, J. C. Kissinger, and J. B. Gutierrez. 2017. A Vision for Health Informatics: Introducing the SKED Framework An Extensible Architecture for Scientific Knowledge Extraction from Data. ArXiv e-prints (2017). arXiv:1706.07992

[60] Richard TH Tsai, Wen-Chi Chou, Ying-Shan Su, Yu-Chun Lin, Cheng-Lung Sung, Hong-Jie Dai, Irene TH Yeh,Wei Ku, Ting-Yi Sung, and Wen-Lian Hsu. 2007. BIOSMILE: A semantic role labeling system for biomedical verbs using a maximum-entropy model with automatically generated template features. BMC bioinformatics 8, 1 (2007), 325.

[61] Richard TH Tsai, Po-Ting Lai, Hong-Jie Dai, Chi-Hsin Huang, Yue-Yang Bow, Yen-Ching Chang, Wen-Harn Pan, and Wen-Lian Hsu. 2009. HypertenGene: Extracting key hypertension genes from biomedical literature with position and automatically-generated template features. BMC bioinformatics 10, Suppl 15 (2009), 59.

[62] Daya C Wimalasuriya and Dejing Dou. 2010. Ontology-based information extraction: An introduction and a survey of current approaches. (2010).

[63] René Witte and Christopher JO Baker. 2005. Combining biological databases and text mining to support new bioinformatics applications. In Natural Language Processing and Information Systems. Springer, 310–321.

[64] Illhoi Yoo, Patricia Alafaireet, Miroslav Marinov, Keila pena-Hernandez, Rajitha Gopidi, Jia-Fu Chang, and Lei Hua. 2012. Data mining in healthcare and biomedicine: a survey of the literature. Journal of medical systems 36, 4 (2012), 2431–2448.

[65] Deyu Zhou and Yulan He. 2008. Extracting interactions between proteins from the literature. Journal of biomedical informatics 41, 2 (2008), 393–407.