Information Extraction for Biomedical and Biological Literature: A Review

Purabi Kalita
Department of Information Technology
Gauhati University

Rashmi Choudhury
Department of Information Technology
Gauhati University

ABSTRACT

Information Extraction is a technique whose importance in this era of information and technology is growing rapidly. It is nothing but a process of scanning text for information relevant to some interests like extracting entities, relations and events. Likewise all other fields, Bioinformatics also depends on information processing. The information available in Biomedical and other Biological researches need to be automated to make them reachable easily. For this, the textual information available is extracted to relational form. There are various researchers working on this field. Most of them are dealing with information extraction for protein interaction, DNA expression arrays etc. The main aim of this review paper is to study different research papers on IE for Biology and about their method of extracting biomedical or biological literature.

General Terms

Information Extraction, Natural Language Processing, Biomedical literature, Information.

Keyword

Information Extraction, Natural Language Processing, Biomedical literature, Information.

1. INTRODUCTION

NLP is the processing of natural language i.e. human language by computer. NLP contains different processes like parsing, tagging, information retrieval, information extraction etc. Here we discuss basically on information extraction in the context of biology.

Information Extraction (IE) is the name given to any process which selectively structures and combines data which is found, explicitly stated or implied, in one or more texts. IE aims at extracting structured information from natural language texts. IE is different from other NLP technologies due to its evaluating and comparing role. IE technology has not yet reached the market but it could be of great significance to information end-user industries of all kinds, especially finance companies, banks, publishers and governments. Computational linguistic techniques and theories are playing a strong role in this emerging technology, which should not be confused with the more mature technology of Information Retrieval (IR), which selects a relevant subset of documents from a larger set. IE extracts information from the actual text of documents. IE is interested in the structure of the texts, whereas one could say that, from an IR point of view, texts are just bags of words. [5, 6]

IE programs generally consist of smaller IE tasks such as document retrieval, segmentation, entity extraction, normalization, co-reference resolution and relationship extraction.

The application of information extraction of text can create a structured view of the information present in the free text. The overall goal of IE is to convert the unstructured documents into relational form so that it can be more easily machine readable.

Information Extraction is an outgrowth of work in automated NLP which began in 1950s with work on transformational grammar by Zellig Harris and later Noam Chomsky. IE starts rapidly from 1980s with a series of conferences known as Message Understanding Conferences (MUCs).

2. INFORMATION EXTRACTION AND BIOLOGY

Despite of the widespread use of computer in biological field, the end result in almost all the researches are available in texts and figures. And also a large part of the information required for biology research can only be found in free-text form, as in MEDLINE abstracts, or in comment fields of relevant reports, as in GenBank feature table annotations. Such information is important for many types of analysis, such as classification of proteins into functional groups, extraction of protein-protein interaction facts, discovery of new functional relationships, maintaining information of material and methods. However the information available in free text forms or in comment fields are not suitable to be used in computerized or automated systems. The role of Information Extraction comes here. IE can extract the information from MEDLINE abstracts or other text sources and convert it to computer readable form can help biologists in their complex biological analysis. [5, 6, 7,8]

There are several algorithms and methods for implementing Information Extraction. Existing methods can be divided into Supervised and Rule based. In the first case, the system learns a statistical/probabilistic model that explains how the relevant entities or relationships are created. Rule based methods are based on predefined rules that capture syntactical, semantic and lexical knowledge required for identifying entities and relationships. For the extraction of independent entities both the methods are useful while for the extraction of relationships supervised methods become more helpful.

Information Extraction has become a very active field in biology. We can find a large number of research papers in this area. A growing number of workshops and conferences are arranged on information extraction and its role in biology.

Generally it becomes a problem to have an access to a full length text for Information Extraction. But in biology, abstracts are collected and indexed in MEDLINE host at the National Library of Medicine (NLM). The system in NLM is called PUBMED and it indexes 9,741 different journals in Medicine and Molecular biology. [4]

To apply Information Extraction in Biological field, different approaches are available. One is statistics of term occurrence. This deals with documents with similar theme. One of the earliest applications of these methods in biology was a general text clustering algorithm developed by Wilbur and coffee based on word frequency vectors to find related MEDLINE documents. These approaches are limited as words are often ambiguous.

Another approach is deeper syntactic analysis applied mostly in detection of protein-protein and protein-drug interactions. This method is limited to small corpora and is not surely scale up to large sized MEDLINE abstracts database.

The mixed approach i.e. the combination of the above two methods seems to be highly appropriate for biological applications. Linguistics tools are good at determining individual terms while statistical methods are useful for finding relationships between terms in a probabilistic way. [4,11]

3. STUDY OF WORKS IN INFORMATION EXTRACTION FOR BIOMEDICAL AND BIOLOGICAL LITERATURE

During our survey, we have come across a number of research papers dealing with the aforementioned topic: the role of Information Extraction in Biological and Biomedical literature.

3.1 Review Paper 1 [1]

A pilot project has been performed by Bisharah Libbus, Thomas C. Rindflesch of University of North Carolina with an aim to construct a tool for identifying and extracting biomedical information from texts. With this tool, they are aiming to provide as much information as possible to the molecular biologists. Generally such a tool exploits terms and relations identified automatically in text by both statistical and symbolic methods in addition to information supplied by National Library of Medicine indexes. Output would consist of structured information based on terms and relations found in the text. They have selected Diabetes as representative multigenic, human traits and have decided to limit their concentration to genomic information and clinical observations with a limited number of specific variables as Disease, Findings, Genes, Alleles, Mutations, Polymorphism, Chromosomes and Genotypes.

In order to identify the specific entities, they have modified an existing Prolog problem which is able to identify Gene, Cell and Drug in texts. For identification of Disease and Findings, they have used Metamap, a program that maps texts to concepts in the Unified Medical Language System (UMLS) Metathesaurus. To identify values for the variables, two paths are pursued parallel. In one path they have produced a syntactic parse structure which identifies simple noun structure in the texts. Based on this result, Metamap maps the parsed sentences to concepts in UMLS Metathesaurus. Each concept in Metathesaurus is categorized with one or more semantic types such as ‘disease or syndrome’ or ‘peptide or amino acid’. This allows the program to identify disease and findings and also some genomic phenomena. For example, Metamap determines that the Noun Phrase Diabetes corresponds to the Metathesaurus concept Diabetes Mellitus with semantic type ‘disease or syndrome’. [9]

Now the other path uses several statistical and empirical methods to identify Genes and protein names. An example is given below.

DNA/NN sequencing/VBG revealed/VBN no/DT

Additional/JJ mutations/NNS in/IN the/DT coding/JJ

Region/NN of/IN the/DT PPAR/MULTIGENE

Gamma/MULTIGENE gene/MULTIGENE in/IN

Genotypes/NNS A12A/c1431c/GENE or/CC

A12A/t1431t/GENE/.

In the above example, PPAR, gamma, A12A/c1413c and CCA12A/t1431t are identified as Genes.

They had produced the preliminary results by running the program on a sample set of 1075 MEDLINE abstracts dealing with the molecular genetics of Diabetes. Output was provided in two formats, one suitable for human and other is machine readable.

The system contains some errors due to limitation of NLP technique though it provides nonetheless a useful basis for genetic information investigation. They had used a database as well as UNIX commands to generate distributed and co occurrence information. In addition to finding parameters related to Diabetes they were also able to extract co occurrence relationship between any two or more terms. It is also stated by the authors that, the databases provided by national Centre for Biotechnology Information at National Library of Medicine can be used in association with their output to provide extensive molecular information on genes and loci of interest. The particular advantage of this research paper is that it has the potentiality to uncover new relationships.

3.2 Review Paper 2 [2]

In the work done by Ronen Feldman, yizhar Regev, Michal Finkelstein-Landau, Eyal Hurvitz & Boris Kogan from ClearForest Corp, USA and Israel, for text mining of biomedical literature, a rule based Information Extraction method has been used.

The paper first describes a structure driven rule based strategy where the predefined semantic relationships are extracted using deep syntactic and semantic analysis of sentences. This is based on a generic multilevel NLP system where the addition of new entities or relations or domains is easy.

Secondly, the paper describes another IE approach using generic syntax based templates. As writing patterns for all possible lexical and semantic combinations for certain relationship is very time consuming, so they have taken a verbRelationTemplate by considering a collection of MEDLINE abstracts where all the relevant entities (such as Gene, Disease, Tissue etc) are available. Here, the generic template verbRealtion extracts two noun phrases NP1 and NP2 by a verb. For example, in the sentence, ‘MC3-R is potently activated by gamma-MSH peptides’, ‘MC3-R’ will be extracted as NP1 and ‘gamma-MSH peptides’ will be extracted as NP2. The verb is ‘activated’.

The framework used for the process is known as DIAL (Declarative Information Analysis Language). The building block of DIAL is rules. Rules are sequences of pattern matching elements. The pattern matching elements must obey the assignment of the rule's parameters and actions concerning external variables. DIAL enables the user to implement separately the different operations required for performing Information Extraction: tokenization, sectioning (recognizing paragraph and sentence boundaries), and morphological and lexical processing, parsing and domain semantics. DIAL has built-in modules that perform the general tasks of tokenization and part-of-speech tagging. In addition, they have developed a general library of rules that perform noun phrase and verb phrase grouping and separate libraries for recognizing common entities, such as companies or persons. So, basically the ClearForest Corp has used an IE module that can incorporate specific customized rules and infrastructure libraries for some specific domain and task. Such module can be called a Rulebook.

Basically, the authors of this paper have aimed to use information extraction as a sub area of text mining approach. The extracted information such as terms, relationships, categories is used to support a range of data mining operation on available documents. They also had focused on the visualization tool. The paper has enabled one to visualize relationships between entities extracted from documents. One can view semantic map or collocations of relationships.

This paper has proposed a machine assisted indexing which helps human experts to review the results extracted by the IE system. He/she doesn't have to review the whole paper but only the suggested portion of texts extracted by the system which reduces the time and other overhead to check a large amount of biomedical literature.

3.3 Review Paper 3 [3]

"Information Extraction from Biomedical Literature Using Text Mining Framework", a paper by Latha. K, Kalimuthu.S, Dr.Rajaram.R demonstrates an optimized method of three distinct stages such as text gathering, text pre-processing and data analysis using Support Vector Machine clustering algorithm for information Extraction from some Biomedical literature.

In the first stage, they have gathered 1000 sample sets of biomedical documents from PUBMED, MEDLINE and NLM. In the second stage, documents are processed in various stages such as tokenization, data cleaning, stop word removal, stemming and identification of interesting terms. Here stemming is done using Paice/Husk algorithm as it produces more mean modified hamming distance compared to other algorithms. Also this algorithm removes extreme outliers. Now the processed documents are analyzed using Support Vector Machine algorithm. In this stage several data mining methods are applied as this is the main place of information extraction. In this paper biomedical literatures are analyzed using support vector machine algorithm as it is a non parametric algorithm. The algorithm uses two parameters p and q. Parameter p controls the number of outliers. Parameter q is the Gaussian kernel parameter which determines the scale at which data is probed and as it is increased, clusters begin to split. Thus the framework tries to extract information from the biomedical text in a more precise way.

They have used java as font end and oracle as back end for implementing the framework.

The main advantage of this project is that along with applying data mining algorithms to texts it tries to reduce the number of documents to be checked. The aim of this thesis is to provide an optimized way of information extraction.

Review Paper 4 [10]

The paper "Protein Structures and Information Extraction from Biological Texts: The PASTA" System describes the Protein Active Site Template Acquisition System which addresses the problem of maintaining growing numbers of protein structure literatures by performing automatic extraction of information relating to the roles of specific amino acid residues in protein molecules from online articles. PASTA is the first information extraction system developed for protein structure domain. Based on the PASTA system a structured database has been built from the extracted information and is made available through network. PASTA extraction task involves two steps: Terminological Tagging, where primary entities in the domain are identified and classified (example. Protein, Species, residues etc.) and PASTA template design, where PASTA templates can be some entities or some relations between entities or a stereotypical event or scenario. The corpus for PASTA is 1513 MEDLINE abstracts.

The evaluation results of the system shows that the system is capable of performing close to but not quite at the state of the art for IE system. PASTA results compare favourably with most other IE systems in biomedical domain. The most important point here is that the pleasing factor of PASTA system is the negligible discrepancy between precision and recall indicating system maturity and tolerance to unseen data. [10]

4. DISCUSSION AND CONCLUSION

Information Extraction for Biological and Biomedical literature is an emerging field of Bioinformatics and also for NLP. There are various ongoing researches in this field and a large number of papers are available. As this is the era of fast information, fast communication, automation is very much necessary in all research fields. All research findings need to be made computer readable so that those can be gained online and fast communication of those findings can be made available to people all over the world. But the problem here is that most of the research findings, basically biological and biomedical results are texts and figures which are unstructured and free. These texts need to be made relational and structured so that they can be computerized. But again Natural language always faces the problem of ambiguity, lack of reasoning capacity of computer. So there are a number of techniques available to make natural language understandable to computer. Information Extraction is such a process. IE extracts the most important terms and entity relationships from the available texts/documents so that we need not have to work on a large volume of data unnecessarily. In Biomedical field IE has a very important role. Most of Biomedical findings deal with terms like Disease, Gene, Protein, syndrome, various Gene names, amino acid, Diabetes etc. During our survey with IE in Biological and Biomedical field, a large number of research papers have been come across which deal with protein-protein interaction, analysis of DNA expression arrays, protein functional description, detection of disease related genes etc. Despite of large number of researches in this area, the developments are not up to the mark. This is due to the limitations in NLP techniques. Developments have not yet crystallized into general agreement on a set of standard evaluation criteria. [5, 6, 7, 8, 12]

No Information Extraction system is 100% correct. There are always some instances that are missed by the system. This is due to the limitation of computer to trace all possible phrasing and contexts used by humans.

5. ACKNOWLEDGMENTS

We are highly indebted to our faculty Mr. Mirzanur Rahman and Mr. Abdul Hannan for their guidance and constant supervision as well as for providing necessary information regarding our study. We would like to express our special thanks to Mr. Samarjeet Borah, convener of NCWBCB 2014 for his guidance and encouragement.

6. REFERENCES

[1] Libbus, Thomas C. Rindflesch, "NLP-Based Information Extraction for Managing the Molecular Biology Literature", AMIA 2002 Annual Symposium Proceedings

[2] Ronen Feldman, Yizhar Regev, Michal Finkelstein Landau, Eyal Hurvitz, Boris Kogan, "Mining Biomedical Literature using Information Extraction", ClearForest Corp, USA, Israel, October 2002.

[3] Latha.K, Kalimuthu.S, Dr.Rajaram.R, "Information Extraction from Biomedical Literature using Text Mining Framework", International Journal of Imaging Science and Engineering, GA, USA, ISSN: 1934-9955, VOL.1, NO.1, January 2007.

[4] Christian Blaschke, Lynette Hirschman, Alfonso Valencia, "Information Extraction in Molecular Biology", Henry Steward Publication, 1467-5463, Briefings In Bioinformatics, VOL.3,NO.2, 1544-165, JUNE 2002.

[5] Tatsuhiko Tsunoda, Limsoon Wong, "Natural Language Processing for Biology", Pacific Symposium on Biocomputing 5:488-489 (2000).

[6] K. Bretonnel Cohen and Lawrence Hunter, "Natural Language Processing and Systems Biology", http://compbio.ucdenver.edu/.

[7] S. Mukherjea, L. V. Subramanian, G. Chanda, S. Sankararaman, R. Kothari, V. Batra, D. Bhardwaj, B. Srivastava," Enhancing a biomedical information extraction system with dictionary mining and context disambiguation", IBM J. RES. & DEV. VOL. 48 NO. 5/6 September/November 2004.

[8] L. Hirshman, J. C. Park, J. Tsujii, L. Wong, and C. H.Wu, "Accomplishments and Challenges in Literature Data Mining for Biology," BioInform. Rev. 18**,** No. 12, 1553–1561 (2002).

[9] UMLS; see http://umlsks.nlm.nih.gov/.

[10] R. Gaizauskas, G. Demetriou, P. J. Artymiuk and P. Willett, "Protein Structures and Information Extraction from Biological Texts: The PASTA System", Oxford University Press, vol.19 no.1, pages 135-143, February 15 2014.

[11] Jerry R. Hobbs, "Information Extraction from biomedical text", Journal of Biomedical Informatics-Special Issue, vol. 35, Issue 4, pages 260-264, August 2002.

[12] Aaron M. Cohen, William R. Hersh, "A survey of current work in biomedical text mining", HENRY STEWART PUBLICATION, vol. 6, no. 1, pages 57-71, March 2005.