

Extraction of Protein Interaction Information from Unstructured Text Using a link Grammar Parser

Rania A. Abul Seoud¹, Nahed H. Solouma², Abou-Bakr M. Youssef³, Yasser M. Kadah³

^{1,3}Misr University for Science and Technology, 6 October City, Egypt

²Laser Institute, Cairo University, Giza, Egypt

³Biomedical Engineering Department, Faculty of Engineering, Cairo University, Giza, Egypt

E-mail: r-abulseoud@k-space.org

Abstract- As research into disease pathology and cellular function continues to generate vast amounts of data, pertaining to protein and gene interactions, there is a critical need to capture these results in structured formats allowing for computational analysis. Although many efforts have been made to create databases that store this information in computer readable form, populating these sources largely requires a manual process of interpreting and extracting interaction relationships from the biological research literature. Being able to efficiently and accurately automate the extraction of interactions from unstructured text, would improve the content of these databases, and provide a method for managing the continued growth of new literature being published. Hence, it is important to have a fully automated extraction system for the biomedical domain for extracting protein interactions from scientific publications. In this paper, we present a completely automated NLP-based information extraction system, to identify protein interactions in biomedical text. Our approach is based on first, tagging biological entities with the help of biomedical and linguistic protein names databases. The system extracts complete interactions by analyzing the matching contents of syntactic roles and their linguistically significant combinations.

Keywords - Bioinformatics; Natural Language Processing; Information Extraction; Protein-Protein Interactions; Link Grammar.

I. INTRODUCTION

Genomic research in the last decade has resulted in the production of a large amount of Information about protein function. That generated data is highly connected; hence, should such data be made easily available. In addition, scientists in the field are aided by many online databases covering different aspects of protein function, such as protein-protein interaction DIP¹ and BIND², CSNDB³ and SPAD⁴. However, since they are dependent on human experts, they rarely store more than a few thousand of the best-known protein relationships and do not contain the most recently discovered facts and experimental details. There is an urgent need for an automatic system capable of accurate extracting protein function information from literature.

Many approaches have been proposed for information extraction (IE) from scientific publications, ranging from simple statistical methods to advanced natural language processing (NLP) systems. In the biomedical context, the first step towards information extraction is to recognize the names of proteins, genes, drugs and other molecules. The next step is to recognize interaction events between such entities. These basic information extraction approaches rely on the matching of pre-specified templates (patterns) or rules. A

number of groups reported application of pattern-matching-based systems for protein-function information extraction. The shortcoming of such systems is their inability to process correctly anything other than short, straightforward statements, which are quite rare in information-saturated MEDLINE and PubMed abstracts. Several attempts have been made to utilize *shallow-parsing techniques* for the task of biological information. Shallow parsers perform partial decomposition of a sentence structure. In some cases, shallow-parsers are used in combination with various heuristic and statistical methods. Information extraction systems based on the *full-sentence parsing approach* deal with the structure of an entire sentence. However, full parsers are significantly slower and require more memory. A problem of parsing ambiguity can be reduced by employment of domain-specific *context-sensitive grammars*. This approach has been implemented in a system called MedLee⁵. Another system is called GENIES [1] which utilizes a grammar based NLP engine for information extraction. *Context-free parsing systems*, on the other hand, are general enough to be applicable to any domain, but completely generic systems seem to be impractical and inefficient. The PathwayAssist system uses an NLP system, MedScan⁶, for the bio-medical domain that tags the entities in text and produces a semantic tree. Slot filler type rules are engineered based on the semantic tree representation to extract relationships from text. Recently, it has been extended as GeneWays⁷, which also provides a Web interface that allows users to search and submit papers of interest for analysis. The BioRAT⁸ system uses manually engineered templates that combine lexical and semantic information to identify protein interactions. *Grammar engineering approaches*, on the other hand use manually generated specialized grammar rules that perform a deep parse of the sentences. *Machine learning approaches* have also been used to learn extraction rules from user tagged training data. These approaches represent the rules learnt in various formats such as decision trees or grammar rules. Recently, extraction systems have also used *Link Grammar* to identify interactions between proteins. Their approach relies on various linkage paths between named entities such as gene and protein names. The IntEx (A Syntactic Role Driven Protein-Protein Interaction Extractor for Bio-Medical Text) system, [2] has used a dependency based English grammar parser, the LG (Sleator and Temperley 1993), to identify syntactic roles for information extraction.

This paper, divided the information extraction system into

¹ <http://dip.doe-mbi.ucla.edu/>

² <http://www.bind.ca/>

³ <http://geo.nih.gov/jp/csndb/>

⁴ <http://www.grt.kyushu-u.ac.jp/eny-doc/>

⁵ <http://lucid.cpmc.columbia.edu/medlee/>

⁶ <http://www.ariadnegenomics.com/products/medscan/>

⁷ <http://geneways.genomecenter.columbia.edu/>

⁸ <http://bioinf.cs.ucl.ac.uk/biorat/>

two components for flexibility and efficiency. One is to apply the natural language processing techniques on the abstracts and the other is to apply the information extraction techniques to interpret results produced by the first component. Many natural language processing approaches at various complexity levels have been reported for extracting biochemical interactions. While some algorithms using simple template matching are unable to deal with the complex syntactic structures, others exploiting sophisticated parsing techniques are hindered by greater computational cost. This paper investigates link grammar parsing for extracting protein - protein interactions. The information extraction system efficiently processes sentences from PubMed abstracts using a dependency based English grammar parser to produce set of simple sentences with various syntactically links. The *Link Grammar* [3] used to identify interactions between proteins. This approach relies on various linkage paths between named entities such as protein names. The most of the system was coded in Perl language and was compiled into a Windows application.

II. METHODOLOGY

A. System Overview:-

The proposed information extraction system consists of four modules: - (as shown in Figure 1)

1) *Information Retrieval (IR) module*: the user first, provides an initial search specification, which he/she thinks it best represents and characterizes the required protein. Then the information retrieval module starts retrieving all PubMed's abstracts satisfying user's specification.

2) A *preprocessor module*: aimed to identify and tag protein names and it consists of:-

(i) **Entity Tagging**: Reads the XML-based format of a PubMed record. Analyses each abstract to identify sentences that mention interaction of wanted proteins. Regular expressions are used to mark the name of proteins in each sentence. Each sentence is considered an "evidence" for an interaction.

(ii) **Preprocessor**: Removes some constructs that cause the Link Grammar Parser to produce an incorrect output such as parentheses in the sentences. Forcing the LG parser to recognize the biological names as noun forms since the parser recognizes words that start with an uppercase letter as a noun. Therefore, the pre-processor converts each protein name to a word starting with an upper case letter. The pre-processor performs minor punctuation corrections on the spacing of commas and semi-colons in the text. It is also reduces the processing time for the abstracts by filtering out sentences that do not contain interactions. It selects the sentences containing at least one protein name. In this paper, it is interested in protein-function information extraction; so the pre-processor outputs only those sentences that contain at least one identified protein.

(3) *Link Grammar Parse Module*: constructing the set of Link Grammar representation of a sentence. The information extraction system uses the Link Grammar Parser (LGP) by [4] as the Natural Language Processor to produce a set of a set Link- Grammar representation representing each sentence.

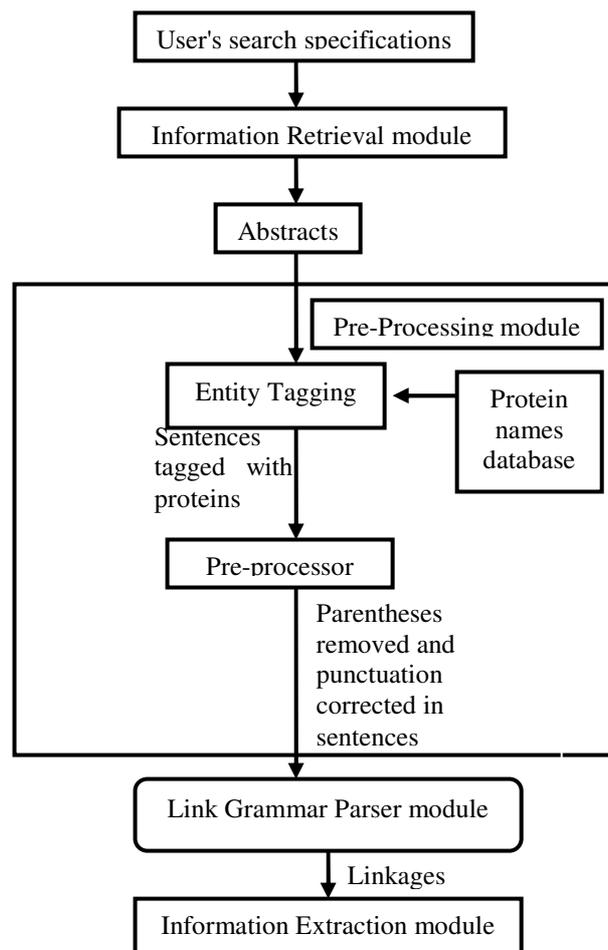


Fig. 1. System Architecture.

The LGP based on a syntactic dependency grammar of the English language producing links between the words in a sentence that correspond to the syntactic structure of the sentence via subject, object, determiner etc.

Link Grammar:-

Link grammar was first introduced by Sleator and Temperley to simplify English grammar with a context-free grammar [3]. The basic idea of link grammar is to connect pairs of words in a sentence with various links. Each word is viewed as a block with connectors coming out. There are various types of connectors, and connectors may point to the right or to the left. A link consists of a left-pointing connector connected with a right-pointing connector of the same type on another word. A valid sentence is one in which all the words are connected in some way (a complete linkage). Rather than examine the basic context of a word within a sentence, the link grammar is based on words within a text form "links" with one another. These links are used not only to identify parts of speech (nouns, verbs, and so on), but also to describe in detail the *function* of that word within the sentence. The Link Grammar is based on a characteristic that if one draws arcs between related words in a sentence [5], the sentence is ungrammatical if arcs cross one another and grammatical if they do not. In Link Grammar, a *linkage* is a single successful parse of a sentence: a set of links in which none of the connecting arcs crosses. A sample parse of the

sentence, "The dog chased a cat." Is shown in figure 2 [2].

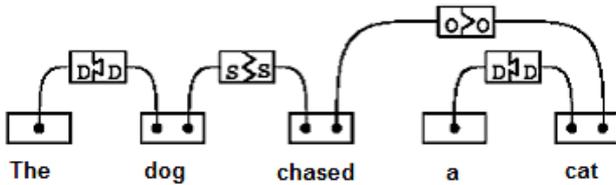


Fig. 2. Link Grammar Representation of a Sentence [2].

In the above example the link between 'dog' and 'chased' is 'S' ('S' links Subject-noun to verbs), the link between 'chased' and 'cat' is 'O' ('O' links verbs to direct or indirect Objects) and the link between 'the' and 'dog' is 'D' ('D' links determiners to nouns). A sample parse output of the LG parser in Bioinformatics domain for the sentences "The SAC6 gene was found by suppression of a yeast action mutation." and for the sentence "HMBA could inhibit the MEC-1 cell proliferation by down-regulation of PCNA expression." Shown in figure 3

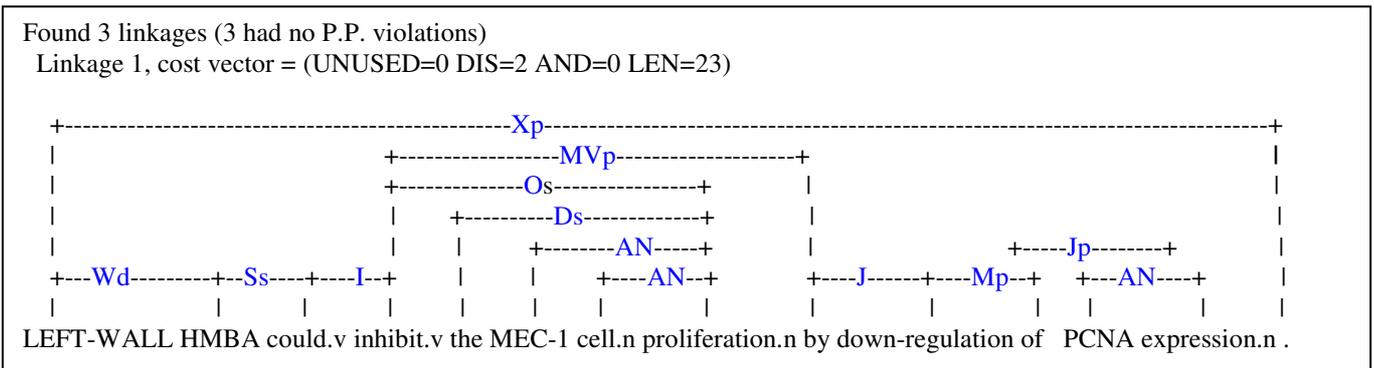
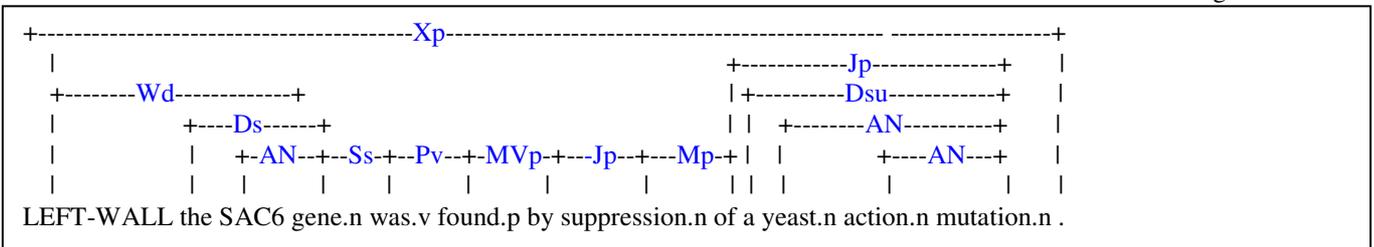


Fig. 4. Link Grammar Parser's output.

The primary parts of speech are labeled with .n and .v to indicate that these words are nouns and verbs, respectively. The labels of the links between words indicate the type of link. The Link Grammar parser Used is Lingua::LinkParser 1.08⁹ (Link Grammar Parser program runs as an EXE file) at Carnegie Mellon University, which is available in CPAN¹⁰. The Lingua::LinkParser provides 107 primary types of links (indicated by the uppercase letters); with many, additional subtypes further detailing the relationship of words (indicated by the lowercase characters). The parser also uses a dictionary that contains the linking requirements of each word and the possible part of speech assignments for the entries. The Link Grammar (LG) parser has around seven hundred definitions that capture many phenomena of English grammar. It can handle: noun-verb agreement, questions,

imperatives, complex and irregular verbs(wanted, go , denied, etc.), different types of nouns and many other things. The dictionary of LG parser has about 60000 word forms, with wide coverage of syntactic constructions [2]. For more details, please visit the web sites¹¹ and¹².

As we were looking for a dependency-tree based parser, the LG parser is selected to provide the syntactic constructs that relate to the linguistic rules for a sentence. Hence, by using the parser, the proposed approach is linguistically oriented. Any certain link types allow us to extract the constituents of sentence irrespective of the tense [4]. The LG parser's ability to detect multiple verbs and their constituent linkage in a complex sentence makes it better suited for the proposed approach. The LG parsers' dictionary can also be easily enhanced to produce better parses for biomedical text [6]. The highly technical terminology and the complex grammatical constructs that are present in the biomedical abstracts make the extraction task difficult. That's why our IE system is based on a deep parse tree structure presented by the Link Grammar and it considers a thorough case based

analysis of contents of various syntactic roles of the sentences [6] like their subjects(S), verbs (V), objects (O) and modifying phrases (M) as well as their linguistically significant and meaningful combinations like SV-O, S-O, S-V-M or S-M.

(4) Information Extraction (IE) Module:

The main component of this module is a set of rules, which can be applied to first identify all the main verbs, i.e., the verbs that truly represent the action in the verb phrase, in the text and then predict the subject for each of these. The module also helps to find out the object of the verb, when present, as well as the modifiers of all verbs and nouns. This set of rules predicts the subject and object of a key verb

⁹ <http://search.cpan.org/~dbrian/Lingua-LinkParser1.08/>
¹⁰ <http://search.cpan.org/>

¹¹ http://www.foo.be/docs/tpj/issues/vol5_3/tpj0503-0010.html
¹² <http://www.link.cs.cmu.edu/link/submit-sentence-4.html>

(interaction word) as well as modifiers of all verbs and nouns. The subject/object prediction scheme begins once the sentence has been passed through the link parser and the linkage for that sentence has been obtained.

When we search for some event in a document, we usually first think of some key words, the presence of which we think will most probably indicate an instance of the required event. In our case this key words is called "*interaction words*" the words that convey a biologically significant action between two gene/protein names. For example in sentence "*HMBA could inhibit the MEC-1 cell proliferation by down-regulation of PCNA expression.*", here 'inhibits' is the only verb, but the sentence has two interactions HMBA, inhibits, MEC-1 cell proliferation and HMBA, down-regulation, PCNA expression. That's why our IE system is based on a deep parse tree structure presented by the LG and it considers a thorough case based analysis of contents of various syntactic roles of the sentences. The main verb "inhibit," describes the action performed by "HMBA" on "MEC-1," is an example of interaction word. Some other example of interaction words are "bind", "down-regulation", "phosphorylation", etc. this interaction words could be collected from UMLS specialist Lexicon¹³ and WordNet¹⁴. Example of Interaction Extractor Algorithm for the same sentence is as follow:-

1. The Algorithm uses the links given by the LG parser to obtain this syntactic constituents: *Subject (S): "HMBA", Object (O): "the MEC-1 cell proliferation"* and Modifying Phrase (MV): "*by down-regulation of PCNA expression*"

2. The system identifies the roles based on the information they contain. For this sentence the subject "HMBA" contains one protein name, Object "the MEC-1 cell proliferation" contains one protein name, and modifying phrase "by down-regulation of PCNA expression" contains one interaction word and one protein name. For each syntactic role of the sentence, the role type matcher identifies the type of each role as either "elementary", "partial" based on its matching content. Here the subject is *Elementary*, object is *Elementary* and modifying phrase is *Partial*.

3. The main verb "inhibit" is identified and we try to extract interaction between subject and object. As main verb is an interaction word, we obtain the interaction:

("HMBA", "inhibit", "the MEC-1 cell proliferation")

4. Now we go even further and extract interaction between subject and modifying phrase. Thus we obtain interaction:

("HMBA", "down-regulation", "PCNA expression")

Each occurrence of the key verb (interaction word), as a main verb is considered to be one occurrence of the required event. So, by finding the subject, object, as well as all available modifiers, almost all information about that instance of the event can be extracted from the document. The development of the extraction module is still an ongoing research project and formal results will be published in details elsewhere.

III. RESULTS

The parsing took 18 ms per sentence on a 600MHz Pentium

¹³ <http://www.nlm.nih.gov/pubs/factsheets/umls.html>

¹⁴ <http://wordnet.princeton.edu/>

III processor with 128MB of RAM. This means that system is faster than similar systems, and preliminary evaluation indicates that performance can be further increased by a factor of 3–5 using better implementations of programming components such as more efficient memory management. The scope of the proposed system was limited to sentences describing human protein function. It is important to note, that using the Link Grammar in the proposed information extraction system makes it applicable to a large number of areas ranging from pathway analysis to clinical information and protein structure-function relationships.

IV. DISCUSSION

The link grammar parser is a robust system, which handles almost all aspects of English grammar. Although it is a dictionary-based system, it can handle sentences admirably well even if they have two words or more, which are not in the dictionary and predict the pan-of-speech for these words with a fair degree of accuracy. Link grammar parsing can handle many syntactic structures and is computationally relatively efficient.

V. CONCLUSION

This paper, presents an information extraction system based on NLP for the purpose of analysis biomedical literature. It is concluded that its performance is satisfactory for the real-time PubMed processing. Also, the syntactic role-based approach compounded with linguistically sound interpretation rules applied on the full sentence's parse can achieve better performance than existing systems which are based on manually engineered pattern and are both costly to develop and not as scalable as the automated mechanisms. Currently an Algorithmic approach is tested, which utilizes the linkage representation generated by the Link Grammar parser to extract protein function information with high precision. The details of this design and implementation will be described elsewhere.

REFERENCES

- [1] C. Friedman, P. Kra, H. Yu, M. Krauthammer, A. Rzhetsky, *GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles*, *Bioinformatics* 17 (2001) S74#/82.
- [2] Sayed T. Ahmed, D. Chidambaram, H. Davulcu, C. Baral. "IntEx: A Syntactic Role Driven Protein-Protein Interaction Extractor for Bio-Medical Text", 2005.
- [3] Sleator, D. and D. Temperley, "*Parsing English with a Link Grammar*" *Third International Workshop on Parsing Technologies*, 1993.
- [4] D. Grinberg, J. Lafferty, and D. Sleator, "*A robust parsing algorithm for link grammars*", Carnegie Mellon University Computer Science technical report CMU-CS-95-1
- [5] J. Ding, D. Berleant, Jun Xu, Andy W. Fulmer "*Extracting Biochemical Interactions from MEDLINE Using a Link Grammar Parser*", 2002.
- [6] Harsha V. Madhyastha, N. Balakrishnan, K. R. Ramakrishnan "*Event Information Extraction Using Link Grammar*", 0-7803-7868- IEEE, 2003.