Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 19 Issue 1 Version 1.0 Year 2019
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Protein and Other Biomedical Entity Name

By Md. Arif Rizvee, Md. Ashfakur Rahman Arju
& Saifuddin Mohammad Tareque

Abstract- Protein and other biomedical entities such as a gene, chromosome names are key elements in bioinformatics. Identifying them individually from the pdf file is very challenging. Because a text pdf document can contain lots of information, identifying them is not so much easy task. So the main focus in our project is converting the pdf file to humanreadable text file then we will have to find the gene and other entities from the GENIA tagger website database. Using natural language processing GENIA tagger will give us the name of all the protein, gene, and other biomedical entity name. After identifying them, we will save it to database. Then we will visualize the related data.

Keywords: tagging protein, gene, and other biomedical entities, natural language processing, GENIA tagger, data visualization.

GJCST-C Classification: J.3

Strictly as per the compliance and regulations of:

Protein and Other Biomedical Entity Name

Md. Arif Rizvee [α], Md. Ashfakur Rahman Arju [σ] & Saifuddin Mohammad Tareque [ρ]

Abstract- Protein and other biomedical entities such as a gene, chromosome names are key elements in bioinformatics. Identifying them individually from the pdf file is very challenging. Because a text pdf document can contain lots of information, identifying them is not so much easy task. So the main focus in our project is converting the pdf file to human-readable text file then we will have to find the gene and other entities from the GENIA tagger website database. Using natural language processing GENIA tagger will give us the name of all the protein, gene, and other biomedical entity name. After identifying them, we will save it to database. Then we will visualize the related data.

Keywords: tagging protein, gene, and other biomedical entities, natural language processing, GENIA tagger, data visualization.

I. Introduction

Protein and other biomedical entity name are used in various biomedical and other bioinformatics-related research. So we will have to work hard to identifying the entities. In text- based literature protein and other biomedical name are tagged with other text. Identifying such entities from text file is very difficult. So we will have to use any scientific approach to solve the problem. Natural language processing is a system which can be used to solve the problem. Using natural language processing we will extract the required info from a text file. We use 'GENIA tagger' database to extract the information from the pdf file and get our required biomedical name. Then we will use these names to make a relation between them and visualized them.

II. Related Work

Many researches are introduced in the field of biomedical and bioinformatics by using data extraction technique. All the work is currently done by text reading system. It is not possible to get the accurate data from manually text extraction technique. As a result protein and other

Biomedical entities are not possible to find out correctly. So it is huge drawback of these types of a research field. In the research we have tried to find out the protein, and gene name which is about 70%-80% correct.

Author α σ ρ: e-mails: arif25-627@diu.edu.bd, Muhammad25-631@diu.edu.bd, saifuddin25-630@diu.edu.bd

III. Our Proposed Work

In the research we try to identify the tagging problem and find a solution related to this type of work. Our work will follow the below procedure.

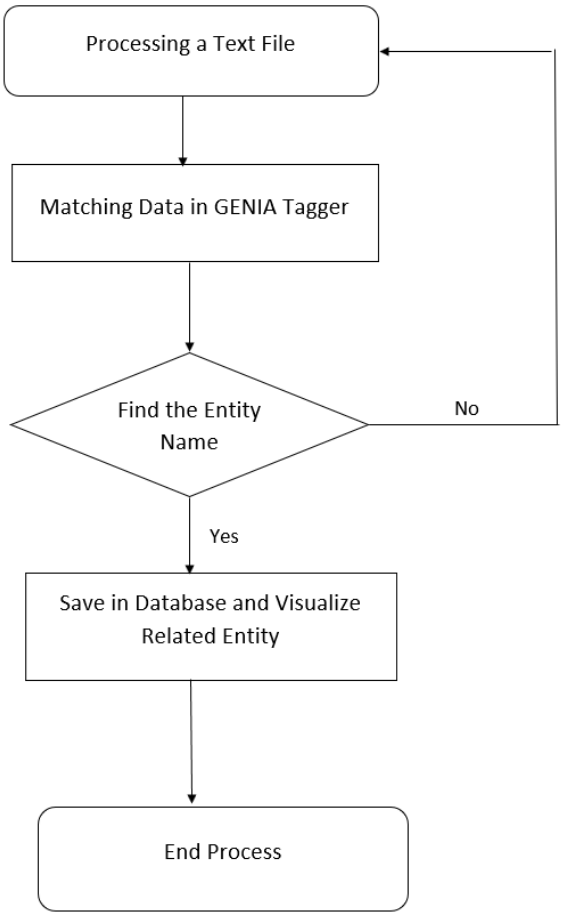

Fig. 1: Tagging & Visualize Protein and Other Biomedical Entity Name

IV. Conversion of PDF Files to Text File

Pdf to text file conversion is looks complex task, but we can convert it easily with the use of the algorithm and other tools.

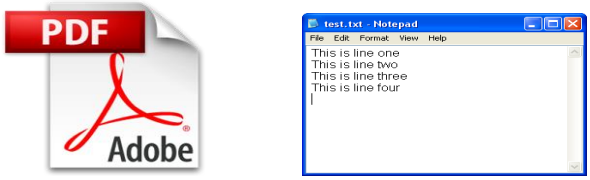

Fig. 2: Conversion of Pdf to Text File

V. Text-based Approach using Natural Language Processing

For our project work, we will have to work on the natural language based text extraction system, which will identify which type of the data is in the text file. Moreover, we can find out the required data and another type of system approach to find that entity. The natural language based system will help us on the text file to find out the necessary information for the system fulfillment of the data. By using this system, we can find out protein and its related entity.

VI. Using Genia Tagger for Data Extraction

GENIA tagger is a website that will help to find out the natural language based system for the protein name tagging from the text; we will use this website for the relevant data search. Moreover, we will use this information in the desired data analyze technique [5]. We also use these type of system for our data processing system [6].

VII. Data Tagging

First we will keep the data in the text file. These data will help us in accessing the information [7][8].

Protein name contains an acronym abbreviating the species name, e.g. Protein human growth hormone (hGH)/protein, but long-form human protein IGF-II / protein /long-form. Protein entities share common terms; there may be only one name entity that can be easily tagged. We tag such name as a protein. Long-form protein CSN subunits 4 /protein, 5,6 /long-form. Assessment of v2 the results on intercoder reliability using the revised guidelines are much better.

protein name contains an acronym abbreviating the species name , e.g . Protein human growth hormone (hGH) /pro- tein , but long-form human protein IGF-II / protein /long-form . protein entities share common terms , there may be only one name entity that can be easily tagged . We tag such an entity as a protein, while the list of enti- ties together are tagged as a long-form, e.g . Long-form protein CSN subunits 4 /protein, 5, 6 /long-form . Assessment of v2 The results on inter-coder reliability using the revised guidelines are much better . We present results for F-measure

Fig. 3: Data Tagging in a Text File

VIII. Tagging Performance with other Documents

GENIA tagger performance is better than other biomedical websites. GENIA tagger is trained in Wall Street Journal corpus, PennBioIE corpus so it performs well in various types of medical data.

Table I: Genia Tagger Comparison with Other Documents

	Wall Street Journal	GENIA corpus
A tagger trained on the WSJ corpus	97.45%	85.34%
A tagger trained on the GENIA corpus	78.4%	95.67%
GENIA tagger	94.67	97.45

IX. Retrieving Data and Save Database According to the Related Entity

We will have to save data according to the text file which we will get from GENIA tagger website. Then we will able to visualize them.

	id	entity_name	entity_tag
Edit Copy Delete	1	protein_complex	gene_promoters
Edit Copy Delete	2	La-related protein 6	LARP6
Edit Copy Delete	3	Haptoglobin- protein	HPR
Edit Copy Delete	4	Parathyroid Protein	HHM

Fig. 4: Saving Related Entity to Database

X. Data Visualization

After retrieving the data, we will analyze all the entities which are related to each other. We will categorize them according to the protein, gene, chromosome various entities. Then we will visualize

them according to their category so that we will able to know which data are the same group[10][11].

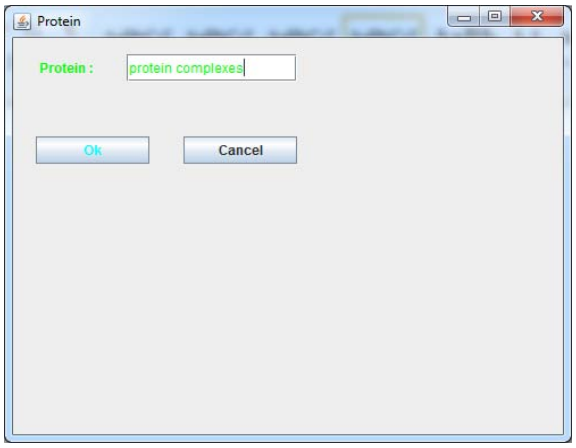

Fig. 5: Searching Related Protein Name

XI. Named Entity Recognition Performance

Our pdf file contains lots of entity of Protein, DNA, RNA, Cell Line, and Cell Type. Genia tagger provides us the flowing the final performance on the evaluation set is as follows [12].

Table II: Genia Tagger Performance

Entity Type	Recall	Precision	F-Score
Protein	75.89	68.89	90.89
RNA	72.56	65.56	67.34
DNA	69.34	73.78	78.90
Cell Line	63.56	85.35	76.34
Cell Type	56.78	65.45	82.98
Overall	73.45	68.67	78.79

XII. Conclusion

Natural Language Processing is a way to find out the similar relational data from a text or document. We try find out related protein and other biomedical entity name and visualize them. Our research makes the system fruitful for the data analysis process.

References Références Referencias

1. Jenny Rose Finkel and Christopher D. Manning, "Nested Named Entity Recognition"
2. Beatrice Alex, Barry Haddow and Claire Grover, "Recognizing Nested Named Entities in Biomedical Text", June 29 - 30, 2007
3. J¨org Tiedemann, "Improved Text Extraction from PDF Documents for Large-Scale Natural Language Processing", 2014
4. Matthew Lease, Eugene Charniak, 'Parsing Biomedical Literature'
5. Firat Tekiner, Yoshimasa Tsuruoka, Jun'ichi Tsujii, "Highly scalable Text Mining - parallel tagging application", Soft Computing, Computing with Words and Perceptions in System Analysis, Decision and Control, 2009. ICSCCW 2009. Fifth International Conference on IEEE, Famagusta, Cyprus.
6. Anni R. Codena Serguei V. PakhomovbRie K. Andoa Patrick H. Duffyb Christopher G.Chute, "Domain-specific language models and lexicons for tagging", Journal of Biomedical Informatics, December 2005.
7. Qian Wang , Huijun Xue , Siqi Li, Ying Chen, Xuelei Tian, Xin Xu, Wei Xiao, Yu Vincent Fu,"A method for labeling proteins with tags at the native genomic loci in budding yeast", Journal pone, May 1, 2017
8. Robert J. Latour, "Tagging methods and associated data analysis ", 2013
9. Ning Kang, Erik M.van MulligenJan, A.Kors, "Comparing and combining chunkers of biomedical text", 2010.
10. Jahiruddina, Muhammad Abulaisha, Lipika Dey, "A concept-driven biomedical knowledge extraction and visualization framework for conceptualization of text corpora", 2010.
11. Vivek N. Bhatia, David H. Perlman, Catherine E. Costello, and Mark E. McComb, "Software Tool for Researching Annotations of Proteins (STRAP): Open-Source Protein Annotation Software with Data Visualization", Journal of Biomedical Informatics, December 2009.
12. Jeffrey P Ferraro, Hal Daumé, III Scott L DuVall , Wendy W Chapman Henk Harkema, Peter J Haug, "Improving performance of natural language processing part-of-speech tagging on clinical narratives through domain adaptation", Journal of the American Medical Informatics Association, Volume 20, Issue 5, 1 September 2013.