Global Journals LaTeX JournalKaleidoscope™
Artificial Intelligence formulated this projection for compatibility purposes from the original article published at Global Journals. However, this technology is currently in beta. *Therefore, kindly ignore odd layouts, missed formulae, text, tables, or figures.*

Protein and Other Biomedical Entity Name

Md. Arif Rizvee[1], Md. Ashfakur Rahman Arju[2] and Saifuddin Mohammad Tareque[3]

[1] Daffodil International University

Received: 13 December 2018 Accepted: 2 January 2019 Published: 15 January 2019

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.

Index terms— tagging protein, gene, and other biomedical entities, natural language processing, GENIA tagger, data visualization.

1 Introduction

rotein and other biomedical entity name are used in various biomedical and other bioinformaticsrelated 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.

2 II.

3 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 ? ? ?: e-mails: arif25-627@diu.edu.bd, Muhammad25-631@diu.edu.bd, saifuddin25-630@diu.edu.bd III.

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

5 VI.

6 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].

7 VII.

8 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. 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.

9 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 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

10 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].

11 Conclusion

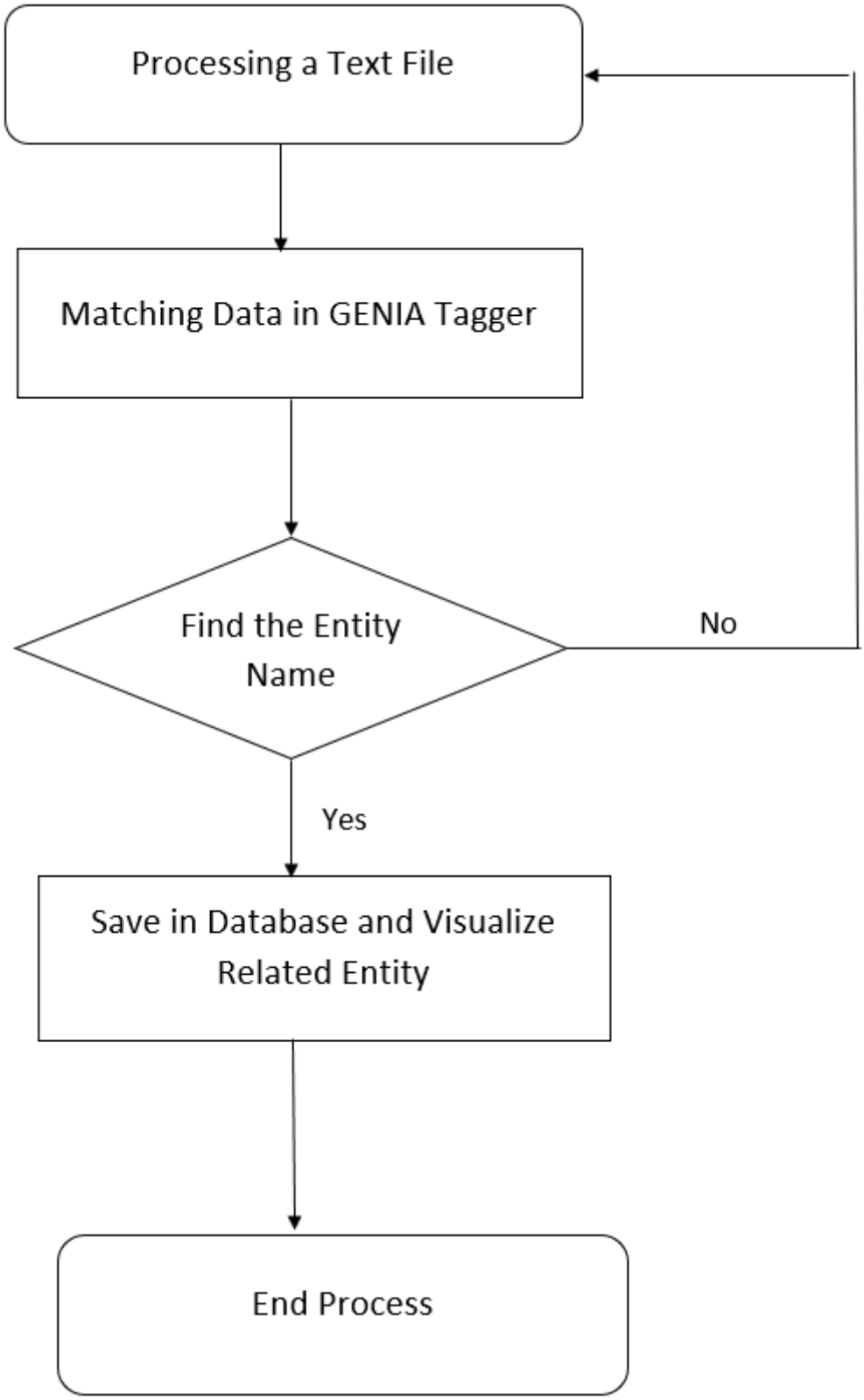
Natural Language Processing is a way to find out the similar relational data from a text or document. We try find out related protein and other biomedical entity name and visualize them. Our research makes the system fruitful for the data analysis process.

12 Entity

[1] [2]

[1] ( ) C © 2019 Global Journals

[2] © 2019 Global Journals



1

Figure 1: Fig. 1 :

2 

Figure 2: Fig. 2 :

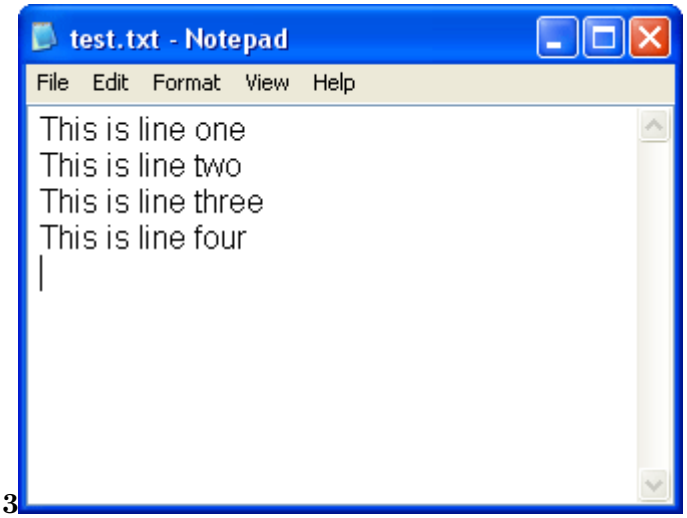
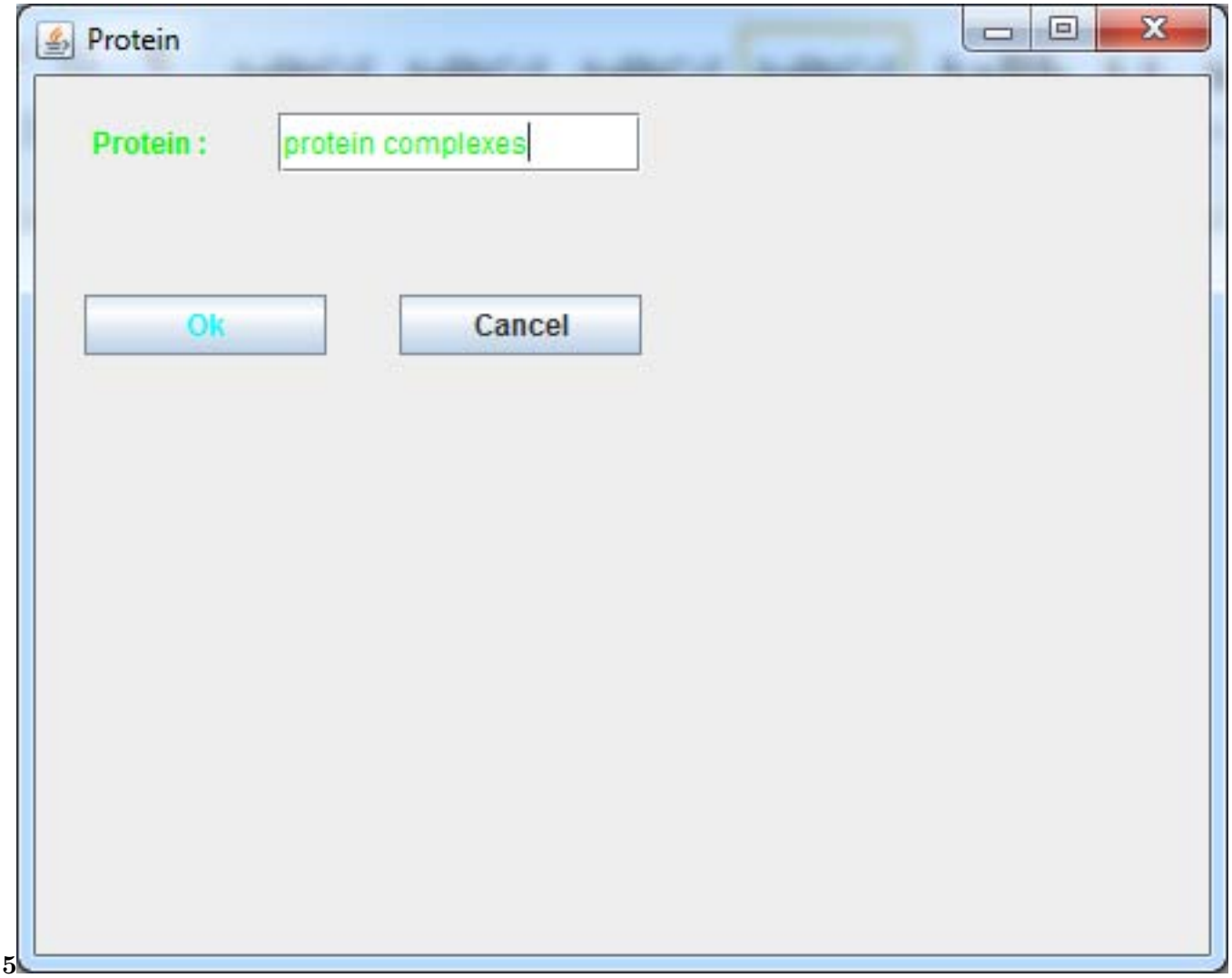
3 

Figure 3: Fig. 3 :

4

	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

Figure 4: Fig. 4 :



5

Figure 5: Fig. 5 :

I

Figure 6: Table I :

II

Figure 7: Table II :

[Lease and Charniak] , Matthew Lease , Eugene Charniak . (Parsing Biomedical Literature)

[Jahiruddina et al. ()] *A concept-driven biomedical knowledge extraction and visualization framework for conceptualization of text corpora*, Muhammad Jahiruddina , Lipika Abulaisha , Dey . 2010.

[Wang et al. (2017)] 'A method for labeling proteins with tags at the native genomic loci in budding yeast'. Qian Wang , Huijun Xue , Siqi Li , Ying Chen , Xuelei Tian , Xin Xu , Wei Xiao , Yu Vincent Fu . *Journal pone* May 1, 2017.

[Kang et al. ()] *Comparing and combining chunkers of biomedical text*, Ning Kang , Erik M Van Mulligenjan , A Kors . 2010.

[Codena Serguei et al. (2005)] 'Domainspecific language models and lexicons for tagging'. Anni R Codena Serguei , V Pakhomovbrie , K Patrick , H Duffyb Christopher , G Chute . *Journal of Biomedical Informatics* December 2005.

[Tiedemann ()] *Improved Text Extraction from PDF Documents for Large-Scale Natural Language Processing*, J¨org Tiedemann . 2014.

[Jeffrey et al. (2013)] 'Improving performance of natural language processing part-of-speech tagging on clinical narratives through domain adaptation'. P Jeffrey , Hal Ferraro , Daumé , L Scott , Wendy W Chapman Henk Duvall , Harkema , J Peter , Haug . *Journal of the American Medical Informatics Association* 1 September 2013. 20 (5) .

[Finkel and Manning] *Nested Named Entity Recognition*, Jenny Rose Finkel , Christopher D Manning .

[Alex et al. ()] *Recognizing Nested Named Entities in Biomedical Text*, Beatrice Alex , Barry Haddow , Claire Grover . June 29 -30, 2007.

[Tekiner et al. ()] 'Soft Computing, Computing with Words and Perceptions in System Analysis, Decision and Control'. Firat Tekiner , Yoshimasa Tsuruoka , ' Jun , Tsujii . *Fifth International Conference on IEEE*, (Famagusta, Cyprus) 2009. 2009. (Highly scalable Text Mining -parallel tagging application)

[Vivek et al. (2009)] 'Software Tool for Researching Annotations of Proteins (STRAP): Open-Source Protein Annotation Software with Data Visualization'. N Vivek , David H Bhatia , Catherine E Perlman , Mark E Costello , Mccomb . *Journal of Biomedical Informatics* December 2009.

[Latour ()] *Tagging methods and associated data analysis*, Robert J Latour . 2013.