G&I Genomics & Informatics

eISSN 2234-0742
Genomics Inform 2018;16(3):75-77
https://doi.org/10.5808/GI.2018.16.3.75

APPLICATION NOTE

GNI Corpus Version 1.0: Annotated Full-Text Corpus of *Genomics & Informatics* to Support Biomedical Information Extraction

So-Yeon Oh[1], Ji-Hyeon Kim[1], Seo-Jin Kim[1], Hee-Jo Nam[1], Hyun-Seok Park[1,2]*

[1]Bioinformatics Laboratory, ELTEC College of Engineering, Ewha Womans University, Seoul 03760, Korea,
[2]Center for Convergence Research of Advanced Technologies, Ewha Womans University, Seoul 03760, Korea

Genomics & Informatics (NLM title abbreviation: *Genomics Inform*) is the official journal of the Korea Genome Organization. Text corpus for this journal annotated with various levels of linguistic information would be a valuable resource as the process of information extraction requires syntactic, semantic, and higher levels of natural language processing. In this study, we publish our new corpus called GNI Corpus version 1.0, extracted and annotated from full texts of *Genomics & Informatics*, with NLTK (Natural Language ToolKit)-based text mining script. The preliminary version of the corpus could be used as a training and testing set of a system that serves a variety of functions for future biomedical text mining.

Keywords: biomedical text mining, corpus linguistics, text analytics

Availability: The datasets and software generated during the current study are publicly available through GitHub repository (https://github.com/Ewha-Bio/Genomics-Informatics-Corpus).

Introduction

Biomedical text mining (also known as BioNLP) refers to text mining applied to texts and literature of the biomedical and molecular biology domain [1, 2]. For biomedical text mining, a corpus is needed, wherein a corpus is a large and structured set of texts electronically stored and processed [3].

Full text of *Genomics & Informatics* in Portable Document Format (PDF) has been archived in *Genomics & Informatics* home pages since 2003 [4], where content of the journal is available immediately upon publication without an embargo period. As of July 18, 2018, 499 full text articles are available as a corpus resource, under the terms of the Creative Commons Attribution Non-Commercial license [5, 6]. To make the corpora more useful for conducting biomedical text mining, they are subjected to a process known as annotation, the practice of adding interpretative linguistic information to a corpus.

Therefore, in this study, we report on developing our new corpus called GNI Corpus, with statistics of annotated objects of the journal. The initial objective of developing GNI Corpus was to analyze counts frequencies of words, and to analyze current trends of the journal.

The Text Preprocess and Annotation Framework

Initially, we wrote a simple Python-based Web crawler, to browse and download PDF files from *Genomics & Informatics* archives. Then, we converted them into plain text files, using PDFMiner or other optical character recognition (OCR) tools [7]. The goal was to transform an image of a text into a readable text.

The next step was annotation. According to Bernardi *et al*. (2002) [8], the biological literature is characterized by heavy use of domain-specific terminology, wherein more than 12% of words found in biochemistry publications are technical terms. Therefore, we used NLTK module, for general text processing [9, 10], and GENIA tagger for recognizing

Received July 31, 2018; Revised August 23, 2018; Accepted August 23, 2018

***Corresponding author:** Tel: +82-2-3277-3513, Fax: +82-2-3277-2306, E-mail: neo@ewha.ac.kr

biological terms [11-13]. The annotation result used as an example sentence from our dataset, *"Most RFLP markers (80%) were pepper-derived clones and these markers were evenly distributed all over the genome."* is as follows:

('Most', 'Most', 'JJS', 'B-NP', 'O') ('RFLP', 'RFLP', 'NN', 'I-NP', 'B-DNA') ('markers', 'marker', 'NNS', 'I-NP', 'I-DNA') ('(','(', '(', 'O', 'O') ('80', '80', 'CD', 'B-NP', 'O') ('%', '%', 'NN', 'I-NP', 'O') (')',')', ')', 'O', 'O') ('were', 'be', 'VBD', 'B-VP', 'O') ('pepper-derived', 'pepper-derived', 'JJ', 'B-NP', 'B-cell_line') ('clones', 'clone', 'NNS', 'I-NP', 'I-cell_line') ('and', 'and', 'CC', 'O', 'O') ('these', 'these', 'DT', 'B-NP', 'O') ('markers', 'marker', 'NNS', 'I-NP', 'O') ('were', 'be', 'VBD', 'B-VP', 'O') ('evenly', 'evenly', 'RB', 'I-VP', 'O') ('distributed', 'distribute', 'VBN', 'I-VP', 'O') ('all', 'all', 'DT', 'B-ADVP', 'O') ('over', 'over', 'IN', 'B-PP', 'O') ('the', 'the', 'DT', 'B-NP', 'O') ('genome', 'genome', 'NN', 'I-NP', 'O') ('.', '.', '.', 'O', 'O').

Four different levels of tags are attached for each word in the example sentence: base form, POS tag, chunk tag, and named-entity tag. For example, (*'RFLP', 'RFLP', 'NN', 'I-NP', 'B-DNA'*) represent the part of speech of the word RFLP (restriction fragment length polymorphism) is a noun ('NN'), and that the word is internal to a noun phrase ('I-NP'), and a begin phrase of a DNA term ('B-DNA').

Specifically, the first tag is a morphological tag to represent a base form of a word. The second tag (based on Penn Treebank tag sets [14]) is a grammatical part-of-speech (POS) tag, needed for analysis of a sentence identifying constituent parts of sentences such as nouns, verbs, and adjectives. The third tag is a syntactic-level tag that links POS tag to higher order units termed chunks that have discrete grammatical meanings such as noun phrases, verb phrases, or other grammatical phrases [15]. For chunk tags, IOB notation was used, wherein the B/I/O terminology refers to begin phrase (B), internal to phrase (I), and outside of phrase (O) [16]. The last tag is a semantic-level tag to classify named entities in text into pre-defined categories such as proteins, DNAs, RNAs, cell lines, and cell types [17, 18].

The Current Status of the GNI Corpus

Presently, we have annotated 499 full texts of *Genomics & Informatics*. Among 2,867,430 words, we have marked up 88,629 names with different semantic classes, including 77,626 proteins, 7,293 DNAs, 1,436 RNAs, 226 cell lines, and 2,048 cell type tags.

Fig. 1 shows our GitHub repository (https://github.com/Ewha-Bio/Genomics- Informatics-Corpus) to host the study design, analysis plan, and data for our study. The tagged datasets and NLTK-based scripts written in Python generated and analyzed during this study are available.

GNI Corpus will be consistently updated in quantity and quality, by manually and automatically. Developing our own version of POS tagger is underway. Future work also includes enhancement of the existing GENIA ontology and co-reference structures.

ORCID: So-Yeon Oh: http://orcid.org/0000-0002-2779-4045; Ji-Hyeon Kim: http://orcid.org/0000-0002-4239-7021; Seo-Jin Kim: http://orcid.org/0000-0002-0621-8874; Hee-Jo Nam: http://orcid.org/0000-0001-6184-6737; Hyun-Seok Park: http://orcid.org/0000-0002-1237-8831

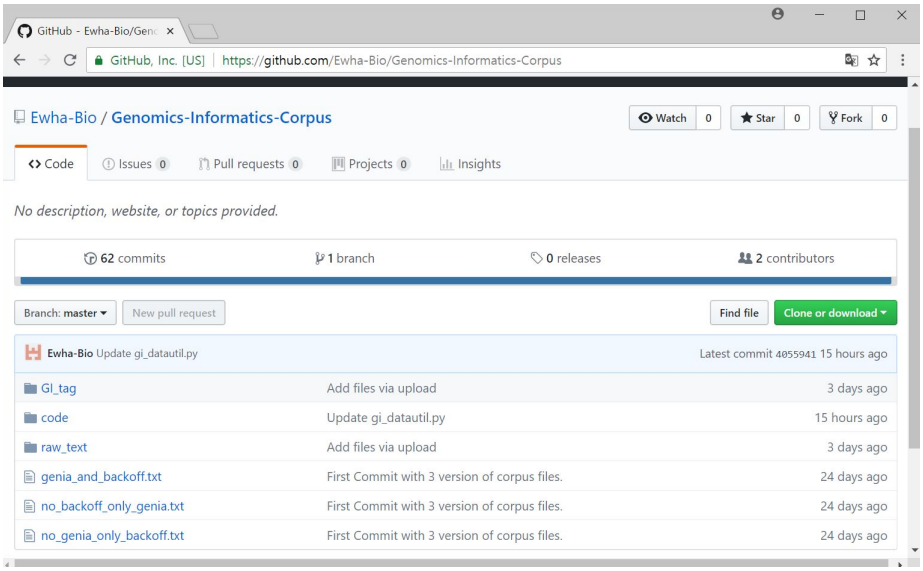

Fig. 1. The GNI Corpus 1.0 datasets and software generated during the current study (https://github.com/Ewha-Bio/Genomics-Informatics-Corpus).

Authors' contribution

Conceptualization: HSP
Data curation: SYO, JHK, SJK, HJN
Methodology: SYO, JHK
Writing – original draft: HSP

Acknowledgments

This work was supported by Ewha Womans University (1-2018-0698-001-1).

References

1. Cohen AM, Hersh WR. A survey of current work in biomedical text mining. *Brief Bioinform* 2005;6:57-71.
2. Zweigenbaum P, Demner-Fushman D, Yu H, Cohen KB. Frontiers of biomedical text mining: current progress. *Brief Bioinform* 2007;8:358-375.
3. Biber D, Conrad S, Reppen R. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press, 1998.
4. Genomics and Informatics archives. Seoul: Korea Genome Organization, 2018. Accessed 2018 Jul 29. Available from: https://genominfo.org/articles/archive.php.
5. Hagedorn G, Mietchen D, Morris RA, Agosti D, Penev L, Berendsohn WG, *et al*. Creative Commons licenses and the non-commercial condition: implications for the re-use of biodiversity information. *Zookeys* 2011;(150):127-149.
6. Creative Commons, Attribution-NonCommercial 4.0 International. Mountain View: Creative Commons, 2018. Accessed 2018 Jul 18. Available from: https://creativecommons.org/licenses/by-nc/4.0/.
7. Shinyama Y. PDFMiner.six: Python PDF parser and analyzer. San Francisco: GitHub Inc., 2018. Accessed 2018 July 17. Available from: https://github.com/pdfminer/pdfminer.six.
8. Bernardi L, Ratsch E, Kania R, Saric J, Rojas JH, Schatz BR, *et al*. Mining information for functional genomics. *IEEE Intell Syst* 2002;17:66-79.
9. Bird S, Klein E, Loper E. *Natural Language Processing with Python*. Sebastopol: O'Reilly Media Inc., 2009.
10. Perkins J. *Python Text Processing with NLTK 2.0 Cookbook*. Birmingham: Packt Publishing, 2010.
11. Collier N, Mima H, Lee SZ, Ohta T, Tateisi Y, Yakushiji A, *et al*. The GENIA project: knowledge acquisition from biology texts. *Genome Inform* 2000;11:448-449.
12. Kim JD, Ohta T, Teteisi Y, Tsujii J. GENIA corpus manual. Technical report TR-NLP-UT-2006-1. Tokyo: Tsujii Laboratory, University of Tokyo, 2006.
13. Tsuruoka Y. GENIA tagger: part-of-speech tagging, shallow parsing, and named entity recognition for biomedical text. Tokyo: University of Tokyo, 2006. Accessed 2018 Jul 27. Available from: http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger.
14. Marcus MP, Marcinkiewicz MA, Santorini B. Building a large annotated corpus of English: The Penn Treebank. *Comput Linguist* 1993;19:313-330.
15. Abney S. Parsing by chunks. In: *Principle-Based Parsing* (Berwick R, Abney S, Tenny C, eds.). Dordrecht: Springer, 1991. pp. 257-278.
16. Breckbaldwin. Coding chunkers as taggers: IO, BIO, BMEWO, and BMEWO+. Accessed 2018 Jul 27. Available from: https://lingpipe-blog.com/2009/10/14/coding-chunkers-as-taggers-io-bio-bmewo-and-bmewo/.
17. Chinchor NA. Overview of MUC-7. In: Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference, 1998 Apr 29-May 1, Fairfax, VA.
18. Nadeau D, Sekine S. A survey of named entity recognition and classification. *Lingvist Invest* 2007;30:3-26.