www.genominfo.org

eISSN 2234-0742

GENOMICS & INFORMATICS

Vol. 17 - No. 2, June 30 2019

Genomics & Informatics is indexed/tracked/covered by PubMed, PubMed Central, KoreaMed, KoMCI, ScienceCentral, CrossRef, BIOSIS Previews, DOAJ, and Google Scholar.

www.genominfo.org
eISSN 2234-0742

G&I
Genomics & Informatics

Volume 17 number 2, June 30, 2019

Aims and scope

Genomics & Informatics is the official journal of the Korea Genome Organization (http://kogo.or.kr). Its abbreviated title is Genomics Inform. It was launched in 2003 by the Korea Genome Organization. It aims at making a substantial contribution to the understanding of any areas of genomics or informatics. Its scope includes novel data on the topics of gene discovery, comparative genome analyses, molecular and human evolution, informatics, genome structure and function, technological innovations and applications, statistical and mathematical methods, cutting-edge genetic and physical mapping, next generation sequencing and de novo assembly, and other topics that present data where sequence information is used to address biological concerns. Especially, Clinical genomics section is for a short report of all kinds of genome analysis data from clinical field such as cancer, diverse complex diseases and genetic diseases. It encourages submission of the cancer panel analysis data for a single cancer patient or a group of patients. It also encourages deposition of the genome data into designated database. Genome archives section is for a short manuscript announcing the genetic information of recently sequenced prokaryotic and eukaryotic genomes. These genome archives data can make the rationale for sequencing a specific organism.

It is published and distributed quarterly at the last dates of March, June, September, and December. All submitted manuscripts will be reviewed and selected for publication after single blind review process. All manuscripts must be submitted online through the e-submission system available from:
http://submit.genominfo.org. It is an online-only peer reviewed open access journal. A free full text both in the XML and PDF formats is available from the journal homepage (https://genominfo.org). It has been indexed by or searchable from PubMed, PubMed Central, BIOSIS Previews, KoreaMed, KoMCI, Korea Citation Index, CrossRef metadata, DOAJ, and Google Scholar. This journal was supported by the Korean Federation of Science and Technology Societies Grant funded by the Korean Government.

- Manuscript Editing by InfoLumi Co., Seongnam, Korea.
- E-submission system by Inforang, Seoul, Korea
- PDF layout, XML production, and homepage management by M2Community Co., Seoul, Korea

Published by the Korea Genome Organization
Contact information
Park, Taesung, Editor-in-Chief

Editorial office of Genomics & Informatics
Room Number 1011, 22, Teheran-ro 7-gil, The Korea Science & Technology Center, Gangnam-gu, Seoul, 06130, Korea
Tel: +82-2-558-9394, Fax: +82-2-558-9434, email: kogo3@kogo.or.kr, URL address: https://genominfo.org

Disclaimer: The publisher, editors, and reviewers do not assume any legal responsibility for errors, omissions, or claims, nor do they provide any warranty, expressed or implied, with respect to information published in Genomics & Informatics

© Copyright 2019, the Korea Genome Organization
It is an open access journal. The articles are distributed under the terms of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Editorial Board

EDITOR IN CHIEF

Park, Taesung — *Seoul National University, Korea*

GUEST EDITORS

Kim, Jin-Dong — *DBCLS, ROIS, Japan*
Bretonnel Cohen, Kevin — *University of Colorado Denver, U.S.A.*
Collier, Nigel — *University of Cambridge, United Kingdom*
Lu, Zhiyong — *NCBI, NLM, U.S.A.*
Rinaldi, Fabio — *University of Zurich, Switzerland*
Park, Hyun Seok — *Ewha Womans University, Korea*

ADVISORY EDITORIAL BOARD

Batzer, Mark A. — *Louisiana State University, U.S.A.*
Church, George M. — *Harvard University, U.S.A.*
Lee, Byungkook — *National Institute of Health, U.S.A.*
Matsuda, Fumihiko — *Kyoto University, Japan*
Sakaki, Yoshiyuki — *RIKEN Genomic Science Center, Japan*
Seo, Jeong-Sun — *Seoul National University, Korea*

ASSOCIATE EDITORS

Cho, Soo Young — *National Cancer Center, Korea*
Choi, Murim — *Seoul National University, Korea*
Han, Kyudong — *Dankook University, Korea*
Huh, Sun — *Hallym University, Korea*
Kim, Sangsoo — *Soongsil University, Korea*
Oh, Bermseok — *Kyung Hee University, Korea*
Oh, S. June — *Inje University, Korea*
Park, Hyun Seok — *Ewha Womans University, Korea*
Yoon, Kyong-Ah — *Konkuk University, Korea*
Won, Sungho — *Seoul National University, Korea*
Woo, Hyun Goo — *Ajou University, Korea*

EDITORIAL BOARD

Ahn, Chul Woo — *University of Texas, U.S.A.*
Chen, Jyh-Yih — *Academia Sinica, Taiwan*
Cordaux, Richard — *University of Poitiers, France*
Divakar, Darshan Devang — *King Saud University, Saudi Arabia*
Hiroki, Yokota — *Indiana University, U.S.A*
Kim, Junhyong — *University of Pennsylvania, U.S.A.*
Kohane, Isaac S. — *Harvard University, U.S.A.*
Liang, Ping — *Brock University, Canada*
Marquardt, Jens — *Mainz University, Germany*
Mishra, Siddhartha K. — *Harisingh Gour Central University, India*
Ohno-Machado, Lucila — *Harvard University, U.S.A.*
Parine, Narasimha Reddy — *King Saud University, Saudi Arabia*
Pawan, K. Dhar — *RIKEN Genomic Science Center, Japan*
Salem, Abdel Halim — *Arabian Gulf University, Bahrain*
Shahik, Shah Md. — *University of Chittagong, Bangladesh*
Sree, N. Sreenath — *Case Western Reserve University, U.S.A.*
Srikulnath, Kornsorn — *Kasetsart University, Tailand*
Terwilliger, Joseph — *Columbia University, U.S.A.*
Valdes, Jorge — *Centro de Genómica yBioinformática, Chile*
Van, Steen — *Kristel University of Liége, Belgium*
Zhang, Feng — *Fudan University, China*

ETHICS EDITOR

Chung, Yeun-Jun — *The Catholic University, Korea*

LAYOUT EDITOR

Jeong, Eun Mi — *M2community, Korea*

STATISTICS EDITOR

Han, Buhm — *Seoul National University, Korea*

WEBSITE AND JATS XML FILE PRODUCER

Bae, Hyo-Jeong — *M2community, Korea*

MANUSCRIPT EDITOR

Chang, Soo Hee — *Infolumi, Korea*

G&I
Genomics & Informatics

Volume 17 number 2, June 30, 2019

Editorial

Introduction to BLAH5 special issue: recent progress on interoperability of biomedical text mining
Jin-Dong Kim, Kevin Bretonnel Cohen, Nigel Collier, Zhiyong Lu, Fabio Rinaldi

Mini reviews

A review of drug knowledge discovery using BioNLP and tensor or matrix decomposition
Mina Gachloo, Yuxing Wang, Jingbo Xia

Resources for assigning MeSH IDs to Japanese medical terms
Yuka Tateisi

Application note

Fully connecting the Observational Health Data Science and Informatics (OHDSI) initiative with the world of linked open data
Juan M. Banda

Biotea-2-Bioschemas, facilitating structured markup for semantically annotated scholarly publications
Leyla Garcia, Olga Giraldo, Alexander Garcia, Dietrich Rebholz-Schuhmann

PharmacoNER Tagger: a deep learning-based tool for automatically finding chemicals and drugs in Spanish medical texts
Jordi Armengol-Estapé, Felipe Soares, Montserrat Marimon, Martin Krallinger

Towards cross-platform interoperability for machine-assisted text annotation
Richard Eckart de Castilho, Nancy Ide, Jin-Dong Kim, Jan-Christoph Klie, Keith Suderman

Improving the CONTES method for normalizing biomedical text entities with concepts from an ontology with (almost) no training data
Arnaud Ferré, Mouhamadou Ba, Robert Bossy

Improving spaCy dependency annotation and PoS tagging web service using independent NER services
Nico Colic, Fabio Rinaldi

OryzaGP: rice gene and protein dataset for named-entity recognition
Pierre Larmande, Huy Do, Yue Wang

G&I
Genomics & Informatics

Editorial

eISSN 2234-0742
Genomics Inform 2019;17(2):e12
https://doi.org/10.5808/GI.2019.17.2.e12

*Corresponding author:
E-mail: jdkim@dbcls.rois.ac.jp

© 2019, Korea Genome Organization
This is an open-access article distributed under the terms of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Introduction to BLAH5 special issue: recent progress on interoperability of biomedical text mining

Jin-Dong Kim[1*], Kevin Bretonnel Cohen[2], Nigel Collier[3], Zhiyong Lu[4], Fabio Rinaldi[5,6,7]

[1]Database Center for Life Science, Research Organization of Information and Systems, Kashiwa 277-0871, Japan
[2]School of Medicine, University of Colorado, Aurora, CO 80045, USA
[3]Faculty of Modern and Medieval Languages, University of Cambridge, Cambridge CB3 9DP, UK
[4]National Center for Biotechnology Information (NCBI), U.S. National Library of Medicine (NLM), Bethesda, MD 20894, USA
[5]Institute of Computational Linguistics, University of Zurich, Zurich CH-8050, Switzerland
[6]IDSIA, Manno CH-6928, Switzerland
[7]Swiss Institute of Bioinformatics, Lausanne CH-1015, Switzerland

Position papers and keynote speeches have often pointed out the great potential of linked data as a mechanism for unifying biological databases with the biomedical literature, but there have been few long-term and large-scale explorations of the research and technical issues that such a unification would raise. To address that paucity of studies, the Biomedical Linked Annotation Hackathon (BLAH) has met for the past five years in order to bring together experts from a variety of relevant disciplines and to create an environment in which they can collaborate on highly focused projects for a short, intense period of colocation.

BLAH is organized annually by the Database Center for Life Science (DBCLS), Research Organization of Information and Systems (ROIS). The goal of the BLAH series is to enhance the interoperability of resources for biomedical text annotation and mining, which we believe is a key for the next breakthrough of biomedical text mining. This special issue delivers seven application notes and two mini reviews, under the theme, "biomedical text mining." They are outcomes from the 5th Biomedical Linked Annotation Hackathon (BLAH5), which was held from 12th through 15th February 2019 in Kashiwa, Japan.

The nine papers included in this special issue are the results of mini projects investigated before, during and/or after BLAH5, for which BLAH5 aimed to provide a good motivation and an ideal environment to make breakthroughs through intensive collaboration among various experts participated in the three days of hackathon. The papers represent various recent issues of biomedical text mining, ranging across sharing and interconnecting datasets, tools and platforms of biomedical text annotation, privacy issues, multilingualism, and machine learning. The application domains include, clinical domain, drug repurposing, rice biology, microbial biology, and so on. However, they were all investigated under the interoperability theme during BLAH5.

We hope it to be an opportunity for audience of the journal of *Genomics and Informatics* to be aware of the issues and progress of biomedical text mining from the perspective of interoperability.

ORCID

Jin-Dong Kim: https://orcid.org/0000-0002-8877-3248
Kevin Bretonnel Cohen: https://orcid.org/0000-0003-1749-8290
Nigel Collier: https://orcid.org/0000-0002-7230-4164
Zhiyong Lu: https://orcid.org/0000-0001-9998-916X
Fabio Rinaldi: https://orcid.org/0000-0001-5718-5462

Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

This work is supported by the ROIS Advanced Data Science Researcher Education Program, the ROIS International Network Formation and MOU Support Program, and the NIH Intramural Research Program, National Library of Medicine (ZL).

G&I
Genomics & Informatics

Mini review

eISSN 2234-0742
Genomics Inform 2019;17(2):e18
https://doi.org/10.5808/GI.2019.17.2.e18

Received: February 23, 2019
Revised: May 30, 2019
Accepted: May 30, 2019

*Corresponding author:
E-mail: xiajingbo.math@gmail.com, xjb@mail.hzau.edu.cn

© 2019, Korea Genome Organization
This is an open-access article distributed under the terms of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

A review of drug knowledge discovery using BioNLP and tensor or matrix decomposition

Mina Gachloo, Yuxing Wang, Jingbo Xia*

Hubei Key Laboratory of Agricultural Bioinformatics, College of Informatics, Huazhong Agricultural University, Wuhan 430070, China

Prediction of the relations among drug and other molecular or social entities is the main knowledge discovery pattern for the purpose of drug-related knowledge discovery. Computational approaches have combined the information from different resources and levels for drug-related knowledge discovery, which provides a sophisticated comprehension of the relationship among drugs, targets, diseases, and targeted genes, at the molecular level, or relationships among drugs, usage, side effect, safety, and user preference, at a social level. In this research, previous work from the BioNLP community and matrix or tensor decomposition was reviewed, compared, and concluded, and eventually, the BioNLP open-shared task was introduced as a promising case study representing this area.

Keywords: BioNLP, drug knowledge discovery, tensor decomposition

Background of Drug-Related Knowledge Discovery

Drug-related knowledge discovery is the process of discovering novel drug targets, drug-side effects, drug-drug interactions (DDIs), drug-disease or drug-indications interactions. The novel knowledge discovery has mainly led to better understanding of the molecular bases of drug efficacy, and with focus on the application scenario of new drug discovery, drug development or drug repurposing [1], i.e., search and replacement of compounds developed for specific diseases [2]. Drug discovery is usually initiated by an experimental method or computational method. Experimental methods, either in vivo or in vitro, are more acceptable by the clinical community. However, the disadvantages of these methods, such as cost and time-consumption, are also obvious. The computational way, also known as *in silico* method, is mainly to perform the knowledge discovery under data mining instead of experimental ("wet lab") manipulations. Early in 2009, a review in *Nature* [3] claimed that *in silico* predictions for drug discovery has come of age, and so far, PubMed has collected over 41 thousand papers about *in silico* drug knowledge discovery.

In Silico Methods for Drug Knowledge Discovery

Generally, *in silico* methods are a computational way to perform knowledge inference by using data mining, with less time-consumption and including machine learning, molecular docking, pharmacophore structure, structure-activity relationships (SAR), quantitative structure-activity relationship (QSAR), and combination methods. Drug-knowledge discovery with *in silico* methods mainly identified core molecular entities, including genes, proteins, therapeutic compounds, and other "omics" information and henceforth, to explore

the novel link between them [4]. Though chemoinformatics methods, such as SAR or QSAR, have made great success in screening chemical libraries, the huge body of candidate chemical compounds has led to overload calculation and made these methods far from perfect [5]. Thanks to the rapid emergence of deep neural network since the early 2010s, deep learning strategies have undoubtedly manifested their computational advantage over chemoinformatics strategies for drug screening [6], and made it another application field of deep neural networks. In the meantime, chemoinformatics strategies mainly focused on novel drug-target identification or DDIs prediction, instead of drug-side effects or drug-disease pairs. With higher odds of success, knowledge discovery tasks of recent ones relied heavily on structured knowledge entires came from bioinformatics-based data base searching or natural language processing (NLP)–aided automatical curation.

In this review, we mainly focus on two typical *in silico* methods of drug-related knowledge discovery. One method is text mining, i.e., Biomedical Natural Language Processing (BioNLP). Another one is a knowledge discovery method, with low rank approximation of drug data with a form of tensors or matrices. As structured knowledge entries were supportive to resolve most drug-related knowledge discovery tasks, NLP methods are regarded as a good addition to traditional in silico methods. In addition, the popularization of knowledge graphs, in recent years, has dramatically encouraged the promising application of knowledge inference in drug-related knowledge discovery. Tensors or matrices were treated as natural data structures, to contain drug knowledge entries, and tensor or matrix decomposition served as a rough approximation of novel link discovery.

Text resources and BioNLP methods for drug-related knowledge discovery

BioNLP is the application of NLP methods to biomedical entities such as macromolecules and relation extraction between protein-protein or drug-drug interactions. As a hyponym word for NLP, the definition of BioNLP appeared in the early 1990s [7], when distributed word representations and applications in BioNLP were introduced. With the fast accumulation of written material of scholarly publications and clinical narratives, the BioNLP community, formed in the late 1990s and various named entity recognition (NER) tools were developed for the purpose of biomedical applications such as DDIs, data base curation, ontology design, and so on [1].

In this section, we review the development of BioNLP in drug-related knowledge discovery by categorizing the resources for which type of research was performed. Three kinds of text resources, i.e., large-scale curation data, small-scale corpora, and heterogeneous data, were introduced, as well as drug-related discovery research approaches based on them. Here, PubMed and OMIM were introduced as two representatives of large-scale curated data, which as a tradition served for drug-related knowledge discovery for years; corpora emerged from small-scale data aiming for serving high quality text mining upon large text data; and finally, multi-omics data was introduced as heterogeneous data.

Large scale curation data and drug-knowledge discovery in a wide range

Released for the first time in 1996, PubMed has long been the main text resources for the BioNLP community to collect references and abstracts on life sciences and biomedical topics [8].

The 2014 version of PubMed Medline was explored by Yang et al. [9] through lexicon filtering and dependency parsing tree establishment. They used trigger word learning to extract relationships between diseases-genes and genes-drugs, After obtaining 114,381 disease-gene and 176,219 gene-drug link pairs, an ABC model was applied to extract the indirect link between disease and gene by considering disease-gene as A-B and gene-drug as B-C.

NER tools were developed, as well, among the BioNLP community, among dozens of popularized NER tools, including tmChem [10], DNorm [11], GNormPlus [12], and tmVar [13]. These were regarded as successful representative tools for recognizing chemicals, diseases, genes, and variations.

In the meantime, emergence of deep learning strategies in NLP propelled bio-NER dramatically, by introducing novel and sophisticated deep neural network training models, in the manner of classifier and word embedding. First, deep learning brought a new generation of neural networks as an effective classifier, i.e., long short-term memory (LSTM) neural networks; Second, deep learning introduced semantics consideration, like word embedding, as input, and enhanced the NER algorithms. For example, Habibi et al.'s work [14] was typical, which fully made use of CRF, LSTM, and word embedding, to extract entities including drugs from text, and the results of this work indicated that deep-learning methods performed better than other biomedical NER methods. The attempts of BioNLP community made the massive bioentity information retrieval more accessible.

As a user-friendly platform run by NCBI, PubTator [15] timely offered the PubMed-scale NER service to tag the above entities. By integrating the tagged entities of PubTator into the Stanford parsing tree, Percha and Altman [16] grouped PubMed sentences into semantically-related categories, to provide relations between entities and pairs, for each sentence. For instance, six groups of gene-chemical pairs were carefully defined in this work, i.e., drug target, metabolism, transport, inhibition, agonism, and antagonism. Finally, sophisticated semantic relations were mined out, such as DDIs, and varia-

https://doi.org/10.5808/GI.2019.17.2.e18

tions in drug responses.

Except PubMed, there were several text resources serving for drug-related knowledge discovery. Online Mendelian Inheritance in Man (OMIM, https://www.omim.org) [17] for drug mechanism, and ClinicalTrails.gov (https://www.clinicaltrials.gov) for drug usage.

OMIM, a popular knowledge base of human genes and genetic disorders, offers enriched text sets for addressing phenotypes of mutated genes. Wang and Zhang [18] manually curated the functional change mutations type, i.e., loss of function (LOF) and gain of function (GOF) mutations. It was stated that LOF and GOF recognition worked for novel drug discovery, in terms of core gene function change. Wang and Zhang [18] hypothesized that the "antagonist" chemical maps to a targeted gene with GOF, while another "agonist" chemical mapped to the gene with LOF. This hypothesis offered a straightforward rule for gene-drug pair filtering. Zhang et al. [19] employed OMIM and PubMed to gather GOF and LOF knowledge on the pathogenesis of antidiabetic targets, finding nine drugs for treating diabetes.

Besides PubMed text resources for published papers, and OMIM for curated heredity-centric knowledge text, ClinicalTrails.gov is a representative of an electronic health record (EHR) text resource, which was established in 1999 [20]. ClinicalTrails.gov contains various information about medical clinical studies in humans, and the open access policy made it widely used. For example, Su and Sanger [21] extracted serious adverse events SAEs data from the text in ClinicalTrials.gov, and ranked drugs by SAEs data, to find those with the least SAE. Then, new drugs could be predicted according to their SAEs. For example, Xu et al. [22] extracted gene alterations and identified cancer treatment trials by developing a semi-automatic framework on documents at ClinicalTrails.gov. In this research, they used three steps including: collect candidate trials about cancer treatment trials, score each candidate trials, and manually review trials with lower scores.

EHR data is a popular source information of clinical and transnational research for drug repurposing. Banda et al. [23] used four sources information from EHRs including public database, source of spontaneous reports, literature and non-EHRs DDIs predication methods to prioritize drug- drug-event association. It should be noted that the abundant clinical information in EHR data made it possible to serve for various precisional medical discovery. Denny et al.'s PheWAS [24] combined long temporal scale EHR data with genomics variation information, and proposed phenome-wide association study to trace core single nucleotide polymorphisms and disease trajectory. The emerging cross disciplinary research based on EHR as well propelled the research issues from Medical Natural Language Processing (MedNLP) [25,26].

In all, the development of large data resource knowledge discovery unveiled the following tendencies:

(1) PubMed is still the main open access resource for large scale resource, meanwhile, lack of other text resources with comparable level and restriction of full text access hinder the development of large scale knowledge discovery for bio-text miners.

(2) After years of development, NER of biomedical entities is not technical headache any longer, and make it possible to run comprehensive knowledge extraction tasks.

(3) As a result, a combination of full open access to PubMed-wide knowledge discovery and restricted access EHR data access for drug knowledge is a main research pat- tern in the next decade.

Corpora and purposes for drug-related text mining

Early attempts to apply BioNLP to knowledge discovery was propelled by the benchmark NLP dataset corpus. A well-structured corpus experiences a rigid evaluation procedure that ensures its usability. The steps included annotation guidelines design, annotation testing, and inter-annotator agreement computation.

The pioneer work was the corpus used in DDIs of DDI 2011 [27], DDI 2013 extraction challenge [28], and SemEval 2013 task 9 [29]. In early attempts, Segura-Bedmar et al. [27] used POS-tagging, lemmatization, and chunking as features of a shallow linguistic kernel method, to perform DDI extraction. To that end, Bui et al. [30] was among dozens of researchers that attended the DDI challenge, which manually created 292 relevant trigger words, converted sentences into semantic structures, extracted and fed features into a known classifier support vector machine (SVM) for DDI extraction. Afterward, Kim et al. [31] used SVMs, as well as performing DDI 2013 challenge, but with richly combined features, including word features, word pair features, dependency graph features, and parse tree features.

Corpora design, and its applications, gradually played substantial roles in drug-related knowledge discovery. In 2016, for the purpose of oncology knowledge discovery, Lee et al. [32] created a cancer and antitumor Biomedical entity Relation ONcology COrpus (BRONCO), which focused on the variant-centric entities including genes, diseases, drugs, and cell lines. Although BRONCO was a disease-oriented corpus, it focused on drugs, and Lee et al. [33] used this corpus to evaluate and develop a mutation-gene-drug discovery pipeline.

Focus on adverse reactions (ADRs) or side effects on drugs has attracted the attention of corpus designers. In that regard, Fang et al. [34] illustrated proper terminology discrimination upon ADR corpus design. A recent ADR-oriented corpus was created by an NCBI team Demner-Fushman et al. [35], i.e., Text Analysis Conference (TAC) 2017 drug labels corpus, which annotated labels of two hun-

dred Food and Drug Administration approved drugs. The mentioned topics they annotated covered "severity," "drug class," "adverse reaction," etc., which were fairly usable for ADR evaluation of drugs. ADR extraction was among the successfully held tasks of TAC 2017 and 2018 [36], and afterwards the same NCBI team constructed MEDIQA challenge, an Association for Computational Linguistics–community challenge for the question entailment of medical records [37] which expanded the drug-related ADR extraction to wider clinical scenarios, also known as MedNLP.

Another focus on drug-related corpus construction is on drug repurposing. Until now, the corpus working on drug repurposing was rare. Recent progress came from Wang et al.'s work [38], which designed an "active gene annotation corpus (AGAC)" to cultivate functional change of mutated genes. AGAC aimed to capture LOF- or GOF-mutated genes, and made it possible to find "agonist vs. LOF" and "antagonist vs. GOF" pairs for "drug vs. gene." This was a nice addition to a mutation-centric corpus for the purpose of drug repurposing [38].

The development of drug-oriented corpora design showed clear tendency as below.

(1) DDIs were a key focus in corpora design, and the DDI corpus has long been a tradition in drug-related corpus construction.
(2) Disease-oriented corpora covered drug- related knowledge curation, which served directly to specific disease and focused on tumors as targets.
(3) Drug-related ADR or side effect information was a focus in corpora design which served for drug effect, and as well led to expanded attention in medical and clinical applications.
(4) Mutation-centric corpus was a novel addition to the drug-related corpora, which was aimed to the application of drug repurposing.

Heterogeneous data for drug-related knowledge discovery

Unlike traditional text data, heterogeneous data is generally non-scientific text, like so cial media and various omics data, including genomic or proteomic data. While the non-scientific text enhanced research studies, with social concerns such as drug abuse, drug misuse, and drug safety, the various omics data achieved success under the collaboration of BioNLP and bioinformatics community.

Just like Twitter served well for drug prescription and drug abuse [39], social media allowed fast tracking of public opinion, and became popular resources for adverse drug reaction mining [40,41], drug misuse [42], drug safety [43], etc. It was worth noting that social media texts were mainly integrated into research with social issue topics, instead of drug knowledge in the molecular level.

With emergence of multi-omics data, the integration of text data with genome, or protome data attracted attention from a cross disciplinary view, for the purpose of drug-gene linking discovery. Early attempts of linking chemical to candidate genes was performed in late 2000s by Li et al. [44], who showed a significant combination of traditional bioinformatics and BioNLP approaches. This study used Online Predicted Human Interaction Database (OPHID), a predicted protein association network database, to obtain protein networks of Alzheimer disease, retrieved from disease-drug-protein links from PubMed, and formed a reliable connectivity map.

In most cases, multi-omics data integration led to indirect link discovery between drugs and their targeted proteins or candidate loci. Zhang et al. [45] obtained a colorectal cancer-related gene list by text mining from PubMed and then integrated genomics data and proteomics data to identify the more risky loci associated with colorectal cancer. Barupal et al. [46] investigated metabolic genes as therapeutic targets in breast tumors by using multi-omics data and text mining. Meanwhile, Long et al. [47] identified and validated oncogenic biomarkers of pancreatic cancer, through integrative text mining and omics-based translational modelling. Such progress also reflects the mainstream data fusion research idea within the bioinformatics community.

To conclude, the availability of the heterogeneous data propelled drug-related knowledge discovery both in social and bioinformatics domains.

(1) Social media data became an exclusively important resources for collecting public opinion, helping to resolve several drug-related topics, such as drug safety, drug usage, or drug side effects.
(2) Integration of text data with multi-omics data became a tendency upon drug-gene linking or therapeutic target.

discovery, and huge text data was regarded as one member of omics data from the view of the bioinformatics community.

Matrix or tensor decomposition methods for drug-related knowledge discovery

Matrix factorization or decomposition are important techniques for extracting information from a matrix or a tensor [48]. Basically, a matrix $A \in R^{m \times n}$ refers to a $m \times n$ data array, which is suitable for storing and linking two entities. Meanwhile, an M-way tensor $R^{n1 \times n2 \times \cdots \times nM}$ provides a higher ordered structure, which is capable of storing M different kinds of entities. The computational decomposition (sometimes called factorization) result of a matrix or a tensor led to a so-called low rank approximation of the original structure, and made a basis for novel link discovery.

If compared with great amount and various patterns of BioNLP research on drug-related knowledge discovery, the research of matrix or tensor decomposition was comparatively less, and more topic-specific. In general, the adaptable data structure made it possible to illustrate higher order links, while the lower rank approximation

https://doi.org/10.5808/GI.2019.17.2.e18

made it a suitable one for novel link discovery. A comprehensive review of mathematical illustration of the matrix decomposition ("also known as matrix factorization") by Wang and Zhang's work [49] listed basic notations, definitions and detailed ideas, while Kolda and Bader [50] provided another one for tensor decomposition, including the classic CANDE-COMP/PARAFAC (CP) and Tucker decomposition. To trace the rapid development of knowledge inference in the years of knowledge graph, Nickel et al.'s RESCAL algorithm [51] made good use of tensor structure for triple knowledge. In addition, Nimishakavi's series work on higher order relation schema [52,53], and side information integration, were representative issues in higher order link discovery and multi-resource data fusion. Since this review is mainly for concluding research on drug-related knowledge discovery, the following section reviews the matrix and tensor level decomposition and applications on drug, respectively.

Matrix decomposition method

Matrix decomposition obtains a sum of lower-rank matrices, and then models a small number of factors [54]. A matrix decomposition models known associations, which predicts novel drug indications.

In 2013, Zheng et al. [55] predicted new drug-target interactions by using collaborative matrix factorization. In their work, three different datasets, such as drug-target interactions, DDIs, and target-target interactions were input to build three matrices. After matrix factorization, two low-rank matrices were obtained, which approximate to the known drug-target interaction matrix, and novel relationship predictions were able to perform by the new number in the approximate matrix. In this research, three kinds of data such as drug-target interaction, drug similarity, and target similarity were input into the three matrices separately.

Similarly, Liu et al.'s work [56] presented an integrated framework to create new therapeutic associations between drug-drug, drug-disease, and disease-disease by matrix decomposition. Zhang et al. [57] proposed two projections including low-dimensional drug projection and disease projection matrix, and utilized them to factorize the drug-disease matrix. Dai et al. [58] used three interaction data, including drug-disease, disease-gene, and drug-gene interactions, to predict drug-disease association. In their work, they clustered genes by using gene-drug interactions and gene-disease interactions, respectively, and two different clustering results consisted of two axes of matrix. After matrix factorization, the novel relations between the clusters were predicted. Through tracing backing to the corresponding disease and drug of the clusters, newly drug-disease relations were obtained.

Tensor decomposition method

Tensor decomposition appeared early in 1927 [59] and emerged into computer science applications in the 2000s. Tensors incorporate a multidimensional array of numerical data and are applied to various machine-learning tasks [48]. Similar to matrix factorization, tensor decomposition extracted a low rank approximation of drug data, while withholding more complex data structure. To that end, Ho et al. [60] utilized tensor decomposition, in an unsupervised manner, for EHR data, and extracted candidate phenotype generation through checking interactions of diagnoses and drugs among patients. Arany et al. [61] similarly used tensor decomposition to infer drug-protein interaction types: competitive or non-competitive. This was a novel idea in this research to design a 3-way tensor with cell ijk represented inhibition of the j-th protein with i-th drug for the k-th given inhibition measure, and to decompose the tensor by using side information of chemical features.

Basically, it was a natural idea to incorporate various drug-related information into the axes of a tensor, and achieve an imaginary knowledge structure. Khan et al. [62] proposed structural toxicogenomics complex tensors by creating structure matrices with drugs and structural descriptors, respectively, a gene tensor for diseases, and post-treatment gene expression, and a toxicity tensor with drug toxicity measurements. Decomposition of the complex tensors led to predictions of toxicity of unseen drugs. Afterward, Taguchi et al. [4] performed a series of studies on identified candidate drugs, especially drugs for heart failure [63], by integrating gene expression data into a tensor decomposition model. Unlike linking drug-protein, drug-toxicity, or drug-disease pairs, linking drug-consumer led to different applications, like drug recommendations. Wang et al. [64] likewise designed a 3-way tensor with "user," "drug" and "label," and constructed a precise drug recommendation model.

The above methods mainly fulfilled tensor axes with various drug-related domain data like gene expression or chemical features, and then a novel link discovery was mined out from the decomposed tensor. Meanwhile, a hybrid strategy of BioNLP and tensor decompostion came from Zhou et al. [65], who used AGAC corpus [38] as a training set to perform OMIM-wide text mining, and predict novel higher order links among five entities, including genes, mutations, functions, diseases, and functional changes. In this work, new nonzero cells in the decomposed tensors were treated as novel links, among five entities, and infer the functional change of a mutated gene. Finally, agonist/antagonist drug information was extracted from DrugBank [66], and applied to help linking "agonist vs. LOF" and "antagonist vs. GOF" pairs, for the purpose of drug repurposing.

Research pattern of novel drug-related knowledge link discovery in the form of matrix or tensor decomposition

Among the above research studies, the characteristics of matrix or tensor decomposition method enabled investigators to input multi-

ple data, and thus provide more comprehensive information for prediction, which may elevate knowledge prediction accuracy. The research tendency of matrix or tensor decomposition on drug-related knowledge discovery is listed below.

(1) Matrix or Tensors are natural data structures to contain multiple arrays of drug-related entries. Paired knowledge entries are mapped into a matrix element, such as a drug-target, drug-drug pair, while three linked entities are mapped into a cell in tensor, such as "drug", "user," and "label," in drug recommendations. Furthermore, higher order links are mapped into higher order tensors.

(2) Generally, novel link discovery is inferred from the novel nonzero cells in the decomposed matrix or tensor. Methods differ according to the chosen decomposition algorithm. For example, a new link is inferred from a core tensor after decomposition in a RESCAL-based tensor decomposition, while a nonzero cell in the approximated tensor counts as a novel link in a CP decomposition.

(3) Three way tensors were the most popular choice in the knowledge inference applications. As shown in Fig. 1, a 3-way tensor is favored more in a triple data structure than that in two matrices, thus making it convenient for high-order link data representation. It is straightforward to claim that a M-way tensor can provide a natural data structure to store higher-order links mentioning M entities. However, the higher the reach of the level, the more sparse the tensor is. This creates a computational bottleneck.

(4) Knowledge inference algorithms such as jointly decomposed matrices and tensors, bring the data fusion idea into the matrix or tensor decomposition strategy, and make it possible to perform a drug-related knowledge discovery, by incorporating various kinds of heterogeneous data.

Conclusion

Trends in BioNLP and drug-related knowledge discovery

The goal of drug-related discovery is to find novel knowledge for extracting drugs, and use the newly identified drugs for disease therapy. In this review, we focused on BioNLP and tensor or matrix decomposition methods to predict novel alternative therapeutic symptoms.

Recent progress in drug-related knowledge discovery led to a couple of research trends:

(1) Well-annotated corpora are a core gold standard dataset. Annotation corpora are crucial to BioNLP, and could help to retrieve and extract information from biomedical text, and also provide standard data for repeatable training and evaluation of BioNLP.

(2) NER tasks are replaced by more complicated knowledge curation tasks, in the BioNLP community. Information from text can be extracted by BioNLP, which could be the original data to find novel knowledge through prediction models. With the recent development of PubTator, NER, and term normalization, are properly solved, while aiming to curate all of PubMed.

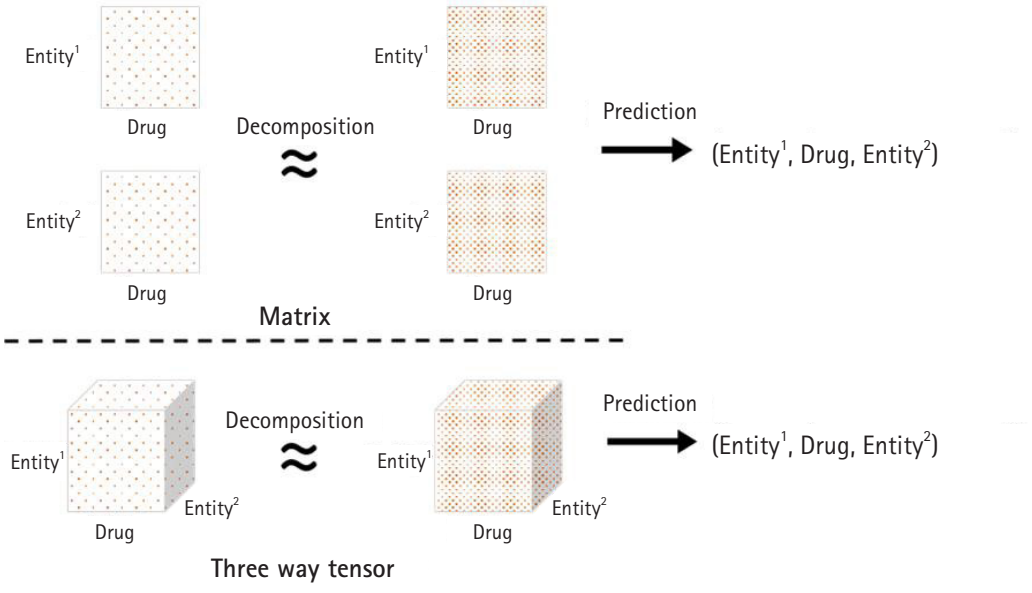

Fig. 1. Structure of a matrix and a three way tensor.

https://doi.org/10.5808/GI.2019.17.2.e18

(3) The application of BioNLP in drug-related knowledge discovery requires deepened integration of multi-omics data. Cross-disciplinary collaboration among BioNLP, MedNLP, and bioinformatics communities is a promising approach.
(4) Knowledge inference, based on tensor or matrix decomposition, is regarded as a reliable prediction model. The integration of algorithms and theorems, developed in knowledge graphs, is a promising approach to resolve various drug-related knowledge discoveries.

BioNLP Open Shared Task: AGAC track

To encourage cross-disciplinary collaboration from various drug-related knowledge discoveries, shared tasks have long been a stage to gather researchers with different backgrounds, e.g., the series of BioNLP Shared Task (BioNLP-ST) workshops [67-71].

Aiming to gather text mining approaches among the BioNLP community to propel drug-oriented knowledge discovery, BioNLP Open Shared Task workshop (https://2019.bionlp-ost.org/tasks) proposed five sub-tasks (tracks). Among the five tasks, we propose a AGAC track (https://sites.google.com/view/bionlp-ost19-agac-track), for the goal of drug repurposing.

AGAC track provides an AGAC and aims to extract mutation-disease knowledge from PubMed. The mutation-disease knowledge in this track links gene-mutation-function change to disease, which not only contains the relationship between mutation and disease, but also indicates the functional change of the mutation, i.e., GOF or LOF. One application of this track is to elevate the efficiency of drug discovery, since matching drugs with their target mutated genes must consider the corresponding of the function change of mutated gene and the pharmacological activities of drugs.

AGAC track contains three different tasks.

(1) Trigger words NER: This task requires participants to recognize trigger words from PubMed abstracts, and annotate them with their corresponding AGAC labels or entities (Var, MPA, Interaction, Pathway, CPA, Reg, PosReg, NegReg, Disease, Gene, Protein, and Enzyme).
(2) Themetic roles identiftcation: Identification of AGAC themetic roles (e.g., Theme Of, Cause Of), between trigger words.
(3) Gene-function mutation-disease link discovery: Extract the gene-(mutation)-function change-biology function or disease link. For example, "Mutations in SHP-2 phosphatase that cause hyperactivation of its catalytic activity have been identified in human leukemias, particularly juvenile myelomonocytic leukemia." From this sentence, the participants need to extract (SHP-2–GOF–juvenile myelomonocytic leukemia).

The baseline methods for task 1 or 2 was performed in Zhou et al.'s work [65], while the "agonist vs. LOF" and "antagonist vs. GOF" hypothesis for the support of drug repurposing was proposed in Wang and Zhang's work [18]. The development of the AGAC corpus [38] laid the basis for the data availability, while PubAnnotation [72] served as the evaluation platform.

ORCID

Mina Gachloo: https://orcid.org/0000-0002-8902-5268
Yuxing Wang: https://orcid.org/0000-0003-4510-2783
Jingbo Xia: https://orcid.org/0000-0002-7285-588X

Authors' Contribution

Conceptualization: JX. Formal analysis: MG, YW. Writing - original draft: MG. Writing - review & editing: JX.

Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

The authors would like to express their gratitudes to Dr. Kevin Bretonnel Cohen for many interesting discussions and nice suggestions about the paper, as well as Dr. Jin-Dong Kim for many illuminative discussions during the BLAH5 workshop. Gratitude is also expressed to Mr. Kaiyin Zhou, Ms. Xuan Qin, Ms. Yuxin Ren, Ms. Shanghui Nie, and all of the audiences who attended the AGAC discussion in the HZAU BioNLP seminar. This work is funded by the Fundamental Research Funds for the Central Universities of China (Project No. 2662018PY096).

References

1. Huang CC, Lu Z. Community challenges in biomedical text mining over 10 years: success, failure and the future. Brief Bioinform 2016;17:132-144.
2. Medina-Franco JL, Giulianotti MA, Welmaker GS, Houghten RA. Shifting from the single to the multitarget paradigm in drug discovery. Drug Discov Today 2013;18:495-501.
3. Hopkins AL. Drug discovery: predicting promiscuity. Nature 2009;462:167-168.
4. Taguchi YH. Identification of candidate drugs using tensor-decomposition-based unsupervised feature extraction in integrated analysis of gene expression between diseases and DrugMatrix datasets. Sci Rep 2017;7:13733.
5. Danishuddin, Khan AU. Descriptors and their selection methods

in QSAR analysis: paradigm for drug design. Drug Discov Today 2016;21:1291-1302.
6. Ghasemi F, Mehridehnavi A, Perez-Garrido A, Perez-Sanchez H. Neural network and deep-learning algorithms used in QSAR studies: merits and drawbacks. Drug Discov Today 2018;23:1784-1790.
7. Zheng J, Yu H. Learning distributed word representations and applications in biomedial natural language processing. Language 1992;18:467-479.
8. Canese K. PubMed celebrates its 10th anniversary. NLM Tech Bull 2006;352:e5.
9. Yang HT, Ju JH, Wong YT, Shmulevich I, Chiang JH. Literature-based discovery of new candidates for drug repurposing. Brief Bioinform 2017;18:488-497.
10. Leaman R, Wei CH, Lu Z. tmChem: a high performance approach for chemical named entity recognition and normalization. J Cheminform 2015;7:S3.
11. Leaman R, Islamaj Dogan R, Lu Z. DNorm: disease name normalization with pairwise learning to rank. Bioinformatics 2013;29:2909-2917.
12. Wei CH, Kao HY, Lu Z. GNormPlus: an integrative approach for tagging genes, gene families, and protein domains. Biomed Res Int 2015;2015:918710.
13. Wei CH, Harris BR, Kao HY, Lu Z. tmVar: a text mining approach for extracting sequence variants in biomedical literature. Bioinformatics 2013;29:1433-1439.
14. Habibi M, Weber L, Neves M, Wiegandt DL, Leser U. Deep learning with word embeddings improves biomedical named entity recognition. Bioinformatics 2017;33:i37-i48.
15. Wei CH, Kao HY, Lu Z. PubTator: a web-based text mining tool for assisting biocuration. Nucleic Acids Res 2013;41:W518-W522.
16. Percha B, Altman RB. A global network of biomedical relationships derived from text. Bioinformatics 2018;34:2614-2624.
17. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 2005;33:D514-D517.
18. Wang ZY, Zhang HY. Rational drug repositioning by medical genetics. Nat Biotechnol 2013;31:1080-1082.
19. Zhang M, Luo H, Xi Z, Rogaeva E. Drug repositioning for diabetes based on 'omics' data mining. PLoS One 2015;10:e0126082.
20. Bourgeois FT, Murthy S, Mandl KD. Outcome reporting among drug trials registered in ClinicalTrials.gov. Ann Intern Med 2010;153:158-166.
21. Su EW, Sanger TM. Systematic drug repositioning through mining adverse event data in ClinicalTrials.gov. PeerJ 2017;5:e3154.
22. Xu J, Lee HJ, Zeng J, Wu Y, Zhang Y, Huang LC, et al. Extracting genetic alteration information for personalized cancer therapy from ClinicalTrials.gov. J Am Med Inform Assoc 2016;23:750-757.
23. Banda JM, Callahan A, Winnenburg R, Strasberg HR, Cami A, Reis BY, et al. Feasibility of prioritizing drug-drug-event associations found in electronic health records. Drug Saf 2016;39:45-57.
24. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics 2010;26:1205-1210.
25. Barrett N, Weber-Jahnke JH. Applying natural language processing toolkits to electronic health records: an experience report. Stud Health Technol Inform 2009;143:441-446.
26. Dalianis H. Clinical Text Mining: Secondary Use of Electronic Patient Records. Cham: Springer, 2018. pp. 109-148.
27. Segura-Bedmar I, Martinez P, de Pablo-Sanchez C. Using a shallow linguistic kernel for drug-drug interaction extraction. J Biomed Inform 2011;44:789-804.
28. Segura-Bedmar I, Martinez P, Herrero Zazo M. Semeval-2013 task 9: extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). In: Second Joint Conference on Lexical and Computational Semantics (* SEM), Vol. 2. Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013) (Manandhar S, Yuret D, eds.), 2013 Jun, Atlanta, GA, USA. Stroudsburg: Association for Computational Linguistics, 2013. pp. 341-350.
29. Herrero-Zazo M, Segura-Bedmar I, Martinez P, Declerck T. The DDI corpus: an annotated corpus with pharmacological substances and drug-drug interactions. J Biomed Inform 2013;46:914-920.
30. Bui QC, Sloot PM, van Mulligen EM, Kors JA. A novel feature-based approach to extract drug-drug interactions from biomedical text. Bioinformatics 2014;30:3365-3371.
31. Kim S, Liu H, Yeganova L, Wilbur WJ. Extracting drug-drug interactions from literature using a rich feature-based linear kernel approach. J Biomed Inform 2015;55:23-30.
32. Lee K, Lee S, Park S, Kim S, Kim S, Choi K, et al. BRONCO: Biomedical entity Relation ONcology COrpus for extracting gene-variant-disease-drug relations. Database (Oxford) 2016;2016:baw043.
33. Lee K, Kim B, Choi Y, Kim S, Shin W, Lee S, et al. Deep learning of mutation-gene-drug relations from the literature. BMC Bioinformatics 2018;19:21.
34. Fang AC, Liu Y, Lu Y, Cao J, Xia J. A corpus-oriented perspective on terminologies of side effect and adverse reaction in support of text retrieval for drug repurposing. Int J Data Min Bioinform 2018;21:269-286.

https://doi.org/10.5808/GI.2019.17.2.e18

35. Demner-Fushman D, Shooshan SE, Rodriguez L, Aronson AR, Lang F, Rogers W, et al. A dataset of 200 structured product labels annotated for adverse drug reactions. Sci Data 2018;5:180001.
36. Roberts K, Demner-Fushman D, Tonning JM. Overview of the TAC 2017 adverse reaction extraction from drug labels track. In: Proceedings of the 2017 Text Analysis Conference, 2017 Nov 13-14, Gaithersburg, MD, USA. Gaithersburg: National Institute of Standards and Technology, 2017.
37. Abacha AB, Demner-Fushman D. A question-entailment approach to question answering. Ithaca: arXiv, Cornell University, 2019. Accessed 2019 May 10. Available from: https://arxiv.org/abs/1901.08079.
38. Wang Y, Yao X, Zhou K, Qin X, Kim JD, Cohen KB, et al. Guideline design of an active gene annotation corpus for the purpose of drug repurposing. In: 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI) (Li W, Li Q, Wang L, eds.), 2018 Oct 13-15, Beijing, China. Piscataway: Institute of Electrical and Electronics Engineers, 2018. pp. 1-5.
39. Hanson CL, Cannon B, Burton S, Giraud-Carrier C. An exploration of social circles and prescription drug abuse through Twitter. J Med Internet Res 2013;15:e189.
40. Nikfarjam A, Sarker A, O'Connor K, Ginn R, Gonzalez G. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. J Am Med Inform Assoc 2015;22:671-681.
41. Lardon J, Abdellaoui R, Bellet F, Asfari H, Souvignet J, Texier N, et al. Adverse drug reaction identification and extraction in social media: a scoping review. J Med Internet Res 2015;17:e171.
42. Bigeard E, Grabar N, Thiessard F. Detection and analysis of drug misuses: a study based on social media messages. Front Pharmacol 2018;9:791.
43. Sinha MS, Freifeld CC, Brownstein JS, Donneyong MM, Rausch P, Lappin BM, et al. Social media impact of the Food and Drug Administration's drug safety communication messaging about zolpidem: mixed-methods analysis. JMIR Public Health Surveill 2018;4:e1.
44. Li J, Zhu X, Chen JY. Building disease-specific drug-protein connectivity maps from molecular interaction networks and PubMed abstracts. PLoS Comput Biol 2009;5:e1000450.
45. Zhang J, Jiang K, Lv L, Wang H, Shen Z, Gao Z, et al. Use of genome-wide association studies for cancer research and drug repositioning. PLoS One 2015;10:e0116477.
46. Barupal DK, Gao B, Budczies J, Phinney BS, Perroud B, Denkert C, et al. Prioritization of metabolic genes as novel therapeutic targets in estrogen-receptor negative breast tumors using multi-omics data and text mining. Oncotarget 2019;10:3894-3809.
47. Long NP, Jung KH, Anh NH, Yan HH, Nghi TD, Park S, et al. An integrative data mining and omics-based translational model for the identification and validation of oncogenic biomarkers of pancreatic cancer. Cancers (Basel) 2019;11:E155.
48. Rabanser S, Shchur O, Gunnemann S. Introduction to tensor decompositions and their applications in machine learning. Ithaca: arXiv, Cornell University, 2017. Accessed 2019 May 10. Available from: https://arxiv.org/abs/1711.10781.
49. Wang YX, Zhang YJ. Nonnegative matrix factorization: a comprehensive review. IEEE Trans Knowl Data Eng 2013;25:1336-1353.
50. Kolda TG, Bader BW. Tensor decompositions and applications. SIAM Rev 2009;51:455-500.
51. Nickel M, Tresp V, Kriegel HP. A three-way model for collective learning on multi-relational data. In: ICML'11 Proceedings of the 28th International Conference on International Conference on Machine Learning (Getoor L, Scheffer T, eds.), 2011 Jun 28-Jul 2, Bellevue, WA, USA. Madison: Omnipress, 2011. pp. 809-816.
52. Nimishakavi M, Saini US, Talukdar P. Relation schema induction using tensor factorization with side information. Ithaca: arXiv, Cornell University, 2016. Accessed 2019 May 10. Available from: https://arxiv.org/abs/1605.04227.
53. Nimishakavi M, Talukdar P. Higher-order relation schema induction using tensor factorization with back-off and aggregation. Ithaca: arXiv, Cornell University, 2017. Accessed 2019 May 10. Available from: https://arxiv.org/abs/1707.01917.
54. Hodos RA, Kidd BA, Shameer K, Readhead BP, Dudley JT. in silico methods for drug repurposing and pharmacology. Wiley Interdiscip Rev Syst Biol Med 2016;8:186-210.
55. Zheng X, Ding H, Mamitsuka H, Zhu S. Collaborative matrix factorization with multiple similarities for predicting drug-target interactions. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Ghani R, Senator TE, Bradley P, Parekh R, He J, eds.), 2013 Aug 11-14, Chicago, IL, USA. New York: Association for Computing Machinery, 2013. pp. 1025-1033.
56. Liu Y, Wu M, Miao C, Zhao P, Li XL. Neighborhood regularized logistic matrix factorization for drug-target interaction prediction. PLoS Comput Biol 2016;12:e1004760.
57. Zhang P, Wang F, Hu J. Towards drug repositioning: a unified computational framework for integrating multiple aspects of drug similarity and disease similarity. AMIA Annu Symp Proc 2014;2014:1258-1267.
58. Dai W, Liu X, Gao Y, Chen L, Song J, Chen D, et al. Matrix factorization-based prediction of novel drug indications by integrating genomic space. Comput Math Methods Med 2015;2015:275045.
59. Hitchcock FL. The expression of a tensor or a polyadic as a sum of products. J Math Phys 1927;6:164-189.

60. Ho JC, Ghosh J, Steinhubl SR, Stewart WF, Denny JC, Malin BA, et al. Limestone: high-throughput candidate phenotype generation via tensor factorization. J Biomed Inform 2014;52:199-211.
61. Arany A, Simm J, Zakeri P, Haber T, Wegner JK, Chupakhin V, et al. Highly scalable tensor factorization for prediction of drug-protein interaction type. Ithaca: arXiv, Cornell University, 2015. Accessed 2019 May 10. Available from: https://arxiv.org/abs/1512.00315.
62. Khan SA, Leppaaho E, Kaski S. Bayesian multi-tensor factorization. Mach Learn 2016;105:233-253.
63. Taguchi YH. Identification of candidate drugs for heart failure using tensor decomposition-based unsupervised feature extraction applied to integrated analysis of gene expression between heart failure and DrugMatrix datasets. In: Intelligent Computing Theories and Application: 13th International Conference (ICIC 2017) (Huang DS, Bevelacqua V, Premaratne P, Gupta P, eds.), 2017 Aug 7-10, Liverpool, UK. Cham: Springer, 2017. pp. 517-528.
64. Wang L, Wang JL, Cheng ZL, Ran L, Yin Z. Personalized medicine recommendation based on tensor decomposition. Comput Sci 2015;42:225-229.
65. Zhou KY, Wang YX, Zhang S, Gachloo M, Kim JD, Luo Q, et al. GOF/LOF knowledge inference with tensor decomposition in support of high order link discovery for gene, mutation and disease. Math Biosci Eng 2019;16:1376-1391.
66. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res 2006;34:D668-D672.
67. Kim JD, Ohta T, Pyysalo S, Kano Y, Tsujii J. Overview of BioNLP'09 Shared Task on event extraction. In: Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task (Tsujii J, ed.), 2009 Jun, Boulder, CO. Stroudsburg: Association for Computational Linguistics, 2009. pp. 1-9.
68. Kim JD, Wang Y, Takagi T, Yonezawa A. Overview of genia event task in BioNLP Shared Task 2011. In: BioNLP Shared Task '11 Proceedings of the BioNLP Shared Task 2011 Workshop (Tsujii J, Kim JD, Pyysalo S, eds.), 2011 Jun 24, Portland, OR. Stroudsburg: Association for Computational Linguistics, 2011. pp. 7-15.
69. Nedellec C, Bossy R, Kim JD, Kim JJ, Ohta T, Pyysalo S, et al. Overview of BioNLP Shared Task 2013. In: Proceedings of the BioNLP Shared Task 2013 Workshop (Nedellec C, Bossy R, Kim JD, Kim JJ, Ohta T, Pyysalo S, et al., eds.), 2013 Aug, Sofia, Bulgaria. Stroudsburg: Association for Computational Linguistics, 2013. pp. 1-7.
70. Deleger L, Bossy R, Chaix E, Ba M, Ferre A, Bessieres P, et al. Overview of the Bacteria Biotope Task at BioNLP Shared Task 2016. In: Proceedings of the 4th BioNLP Shared Task Workshop (Nedellec C, Bossy R, Kim R, Kim JD, eds.), 2016 Aug, Berlin, Germany. Stroudsburg: Association for Computational Linguistics, 2016. pp. 12-22.
71. Chaix L, Dubreucq B, Fatihi A, Valsamou D, Bossy R, Ba M, et al. Overview of the regulatory network of plant seed development (SeeDev) task at the BioNLP Shared Task 2016. In: Proceedings of the 4th BioNLP Shared Task Workshop (Nedellec C, Bossy R, Kim JD, eds.), 2016 Aug, Berlin, Germany. Stroudsburg: Association for Computational Linguistics, 2016. pp. 1-11.
72. Kim JD, Wang Y. Pubannotation: a persistent and sharable corpus and annotation repository. In: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (Cohen KB, Demner-Fushman D, Ananiadou S, Webber B, Tsujii J, Pestian J, eds.), 2012 Jun, Montreal, Canada. Stroudsburg: Association for Computational Linguistics, 2012. pp. 202-205.

https://doi.org/10.5808/GI.2019.17.2.e18

G&I
Genomics & Informatics

Mini review

eISSN 2234-0742
Genomics Inform 2019;17(2):e16
https://doi.org/10.5808/GI.2019.17.2.e16

Received: March 16, 2019
Revised: May 24, 2019
Accepted: May 24, 2019

*Corresponding author:
E-mail: tateisi@biosciencedbc.jp

© 2019, Korea Genome Organization
This is an open-access article distributed under the terms of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Resources for assigning MeSH IDs to Japanese medical terms

Yuka Tateisi*

National Bioscience Database Center, Japan Science and Technology Agency, Tokyo 102-8666, Japan

Medical Subject Headings (MeSH), a medical thesaurus created by the National Library of Medicine (NLM), is a useful resource for natural language processing (NLP). In this article, the current status of the Japanese version of Medical Subject Headings (MeSH) is reviewed. Online investigation found that Japanese-English dictionaries, which assign MeSH information to applicable terms, but use them for NLP, were found to be difficult to access, due to license restrictions. Here, we investigate an open-source Japanese-English glossary as an alternative method for assigning MeSH IDs to Japanese terms, to obtain preliminary data for NLP proof-of-concept.

Keywords: Japanese language resource, medical vocabulary, MeSH

Introduction

Tokenization, or word segmentation, is an indispensable first step in natural language processing (NLP). Unlike languages such as English, in which sentences are already segmented into words by spaces, tokenization is not a trivial process in Japanese, where usually no word delimiters are provided.

Systems called morphological analyzers, which tokenize sentences and assign parts of speech to the tokens, are used for this purpose, and most current morphological analyzers require dictionaries. In this process, biomedical text requires special dictionaries, because technical terms often cause problems as out-of-vocabulary terms.

In BLAH5 (Biomedical Linked Annotation Hackathon 5), the current status of Japanese medical vocabularies, especially the Japanese version of Medical Subject Headings (MeSH) is reviewed, and the creation of an open-source alternative is attempted.

Medical Subject Headings (MeSH)

MeSH [1] is a medical thesaurus developed and maintained by the U.S. National Library of Medicine (NLM) for indexing and cataloging biomedical research articles. MeSH consists of descriptors, qualifiers, and supplementary concept records (SCRs). As of 2018, MeSH contained 28,939 descriptors, 79 qualifiers, and 244,778 SCRs, with each mapped to a unique ID (MeSH ID). Descriptors are hierarchically organized sets of medical concepts, and a descriptor record has a heading (a term in a controlled vocabulary for describing the subjects of articles), entry terms (synonyms to the heading), tree numbers (positions in the hierarchy), scope note (the definition of the subject, in natural language text), and other information. A descriptor may have more than one tree numbers (thus, the MeSH hierarchy is not a tree but a lattice). Qualifiers are terms that are used with a descriptor to define the

domain of the study for which the descriptor is applied. SCRs consist mostly of names of substances, having headings and synonyms, but not tree numbers. Instead, SCRs are mapped to the headings in descriptor records.

Due to synonyms, the number of terms that can be mapped to records far exceeds the number of records. For example, there are 234,842 unique terms (headings and synonyms) in MeSH descriptors, and 609,418 unique terms in supplementary concepts. As the synonyms are mapped to headings, via MeSH IDs, and are hierarchically organized, via tree numbers, MeSH represents synonym- and hypernym-hyponym relationships between biomedical terms, and is thus a useful resource for text mining.

MeSH is openly available (downloadable) from the NLM site, and is also included as a part of the Unified Medical Language System (UMLS), which also includes translation of MeSH into other languages, including Japanese. The Japanese translation of MeSH, included in the UMLS, was created by NPO Japan Medical Abstracts Society (JAMAS, *Igaku-Chuo-Zasshi*). The current version, as of February 2019, was released in 2015, and is based on MeSH 2014.

Although the original English MeSH is openly available, other MeSH translations in the UMLS (Czech, Dutch, Finnish, French, German, Italian, Japanese, Latvian, Norwegian, Polish, Portuguese [Brazilian], Russian, Croatian, Spanish, and Swedish) carry the Category 3 License Restriction (https://uts.nlm.nih.gov/license.html); that is, they can only be used at the licencee's site, and cannot be incorporated into publicly accessible computer-based information systems, thus prohibiting creation of open-source, derivative works. These restrictions significantly complicate adapting MeSH for dictionaries used for various NLP tasks.

Japanese Language Resources with Mappings to MeSH

There are several online resources in Japan which have MeSH information. Unfortunately, none are fully open for NLP. For example, the *Igaku-Chuo-Zasshi* Thesaurus (https://www.jamas.or.jp/database/thesaurus.html), created by JAMAS, organizes Japanese medical terms, based on MeSH tree structures. The latest version (version 9), released in 2019, is freely available for searching and browsing, without registration, but is not downloadable. This thesaurus is based on MeSH 2018, but is not its direct translation. It has 28,247 of 28,939 MeSH 2018 descriptors, and 3,513 terms outside the original MeSH, such as names of drugs sold in Japan and names of Japanese locations and institutions. The terms excluded from MeSH 2018 mainly consist of the names of foreign locations and institutions. The previous version (version 8), released in 2015, was based on MeSH 2014, and claimed to be the Japanese MeSH included in the UMLS. However, while the data download is not free, and the price is not made public.

The Japanese Association of Medical Sciences (JAMS, *Nihon Igakkai*) provides an English-Japanese/Japanese-English medical dictionary online (http://jams.med.or.jp/dic/mdic.html), to registered users, for searching. Download of data, however, is not available. The latest version of the online dictionary (November 2016) has 71,067 Japanese terms and 70,103 English terms, corresponding to 50,088 concepts. The terms in MeSH (and their translation to Japanese) are marked as such, and the top of the MeSH tree numbers (e.g., C14 for *abdominal aortic aneurysm*, whose tree numbers are C14.907.055.239.075 and C14.907.109.139.075) are provided. This MeSH coding is based on MeSH 2014.

The Life Science Dictionary (https://lsd-project.jp/cgi-bin/lsd-proj/ejlookup04.pl) [2] is developed and regularly updated at Kyoto University. The dictionary actually consists of three mutually linked components: the English-Japanese/Japanese-English dictionary, the thesaurus, and the corpus. The English-Japanese/Japanese-English dictionary has 117,857 English terms and 132,100 Japanese terms. The entries have been assigned ICD-10 (International Statistical Classification of Diseases, version 10) codes, and a link to the thesaurus, where applicable. The thesaurus is a subset of English terms in the dictionary which have corresponding entries in MeSH (the current thesaurus is based on MeSH 2018), organized in the MeSH hierarchy, with synonyms both in English and Japanese, and related terms, based on co-ocurrences in PubMed abstracts. The corpus is a KWIC (keyword in context) of English terms in PubMed articles, generated on-demand. The dictionary, the thesaurus, and the corpus can all be searched freely, without registration, but downloading them requires an extra license agreement. In addition, the Life Science Dictionary converted to a Resource Description Framework (RDF) format is maintained by the Database Center for Life Science (DBCLS, http://lsd.dbcls.jp/portal/), available under the Creative Commons Attribution-NoDerivs 3.0 Unported license (https://creativecommons.org/licenses/by-nd/3.0/).

The Interlinking Ontology for Biological Concepts (IOBC) [3] is a derivative of the JST-thesaurus (i.e., the thesaurus for indexing general science and technology publications used by the Japan Science and Technology Agency [JST]). This derivative was developed by the National Bioscience Database Center of JST, and is available from the NCBO BioPortal (https://bioportal.bioontology.org/ontologies/IOBC) under the Creative Commons Attribution-NonCommercial 4.0 International license (https://creativecommons.org/licenses/by-nc/4.0/). This IOBC has 153,160 concepts, from the life sciences and related categories (such as chemistry), from the JST thesaurus 2015 version, written in both English and Japanese.

https://doi.org/10.5808/GI.2019.17.2.e16

Among them, 12,780 concepts also have corresponding MeSH UIDs. However, the MeSH information in the IOBC has not been updated from the one in the JST thesaurus 2015 version. The JST thesaurus itself is updated, and can be freely searched via web interface, although data download is not allowed. The JST also has the RDF version of the thesaurus, and its SPARQL endpoint was available to registered users of the service "J-Global Knowledge," which unfortunately, has been discontinued.

Of these dictionaries, the JAMAS Thesaurus, JAMS Dictionary, and Life Science Dictionary are targeted to readers to seek information on already known terms, but cannot be searched by applications (such as morphological analysis), where the systems must find unspecified terms, over a broader context. Since creating derivatives of these dictionaries is not allowed for free, adapting them for NLP processes (such as morphological analysis), and especially making the results open, is very difficult. The IOBC does allow creation of derivatives, but only for non-commercial purposes, and its MeSH information is not updated, at least in the current version.

Toward an Open Alternative of Japanese MeSH

In BLAH5, a small experiment was conducted as an attempt to map Japanese medical terms to MeSH, using open English-Japanese vocabularies, by first mapping Japanese terms to English terms, and the resulting English terms mapped to concepts in MeSH, which have the term as either its heading or one of its entry terms.

In our endeavor, we first used the English Wikipedia. From this, we retrieved Wikipedia entries (via DBpedia), with MeSH UIDs in the infobox, and Japanese language links were retrieved. The number of English Wikipedia entries with the MeSH UIDs was 2,719, of which 1,136 had links to Japanese Wikipedia entries. The number of distinct MeSH UIDs was 1,102, of which 1,075 were assigned to one Wikipedia entry. The remaining 27 UIDs were mapped to more than one term, but the relationships of the terms assigned the same UID were not always synonymous. Those other undesirable relationships included hypernym-hyponym (e.g., D003117 for *Dichromacy and Color blindness*; dichromacy is a type of color blindness) and cause-results (e.g., D006471), for *Gastrointestinal bleeding* and *Hematochezia*; hematochezia is a result of gastrointestinal bleeding). Further investigation is left for future work.

Second, use of open English-Japanese glossaries, for Japanese-English mapping and matching of English terms to the MeSH UID, was attempted. For this preliminary experiment, the MeSpEN English-Japanese glossary, a part of the multilingual medical glossary [4], developed by the Spanish National Cancer Research Center and the Barcelona Supercomputing Center (the version in this experiment was downloaded from http://temu.bsc.es/mespen/downloads/glossaries.tar.gz), was used. The English-Japanese glossary had 27,668 entries, although 18,325 unique English-Japanese ones were duplicate entries. Simple (lower-cased) matching against MH (heading) and MN (entry terms) of the MeSH 2018 descriptors, assigned the MeSH UIDs to 5,866 English-Japanese pairs. From this process, the number of unique MeSH UIDs assigned to these pairs was 2,955, and matched the remaining pairs to MeSH 2018 supplementary concepts (having assigned UIDs), leaving a remaining 1,725 pairs (1,173 unique UIDs).

Conclusion

In this study, dictionaries for assigning MeSH information to Japanese medical terms was investigated. Although there are limited numbers of Japanese-English dictionaries that assign MeSH IDs (i.e., tree numbers) to Japanese terms, using them for NLP applications is not simple, due to license restrictions. An alternative approach, using open-source resources, was attempted, and yielded a partial success. Using other open resources, and qualitative evaluation of our results, will be the subject of future work.

ORCID

Yuka Tateisi: https://orcid.org/0000-0002-3813-5782

Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

I am grateful to Dina Demner-Fushman, Nigel Collier, Martin Krallinger, Mizuki Morita, and Yasunori Yamamoto for their suggestions.

References

1. Rogers FB. Communication to the editor. Bull Med Libr Assoc 1963;51:114-116.
2. Kaneko S, Fujita N, Ugawa Y, Kawamoto T, Takeuchi H, Takekoshi M, et al. Life science dictionary: a versatile electronic database of medical and biological terms. In: Dictionaries and Language Learning: How can Dictionaries Help Human & Machine Learning. The 3rd Asialex Biennial International Conference (Murata M, Yamada S, Tono Y, eds.), 2003 Aug 27-29, Meikai University, Uruyasa, Chiba, Japan. Tokyo: The Asian Association for Lexicography, 2003. pp. 434-439.
3. Kushida T, Kozaki K, Tateisi Y, Watanabe K, Masuda T, Matsumu-

ra K, et al. Efficient construction of a new ontology for life sciences by sub-classifying related terms in the Japan Science and Technology Agency thesaurus. In: Proceedings of the 8th International Conference on Biomedical Ontology (ICBO 2017) (Horridge M, Lord P, Warrender JD, eds.), 2017 Sep 13-15, Newcastle upon Tyne, UK. Newcastle upon Tyne: Newcastle University, 2017. pp. 1-6.

4. Villegas M, Intxaurrondo A, Gonzalez-Agirre A, Marimon M, Krallinger M. The MeSpEN resource for English-Spanish medical machine translation and terminologies: census of parallel corpora, glossaries and term translations. In: LREC MultilingualBIO: Multilingual Biomedical Text Processing (Malero M, Krallinger M, Gonzalez-Agirre A, eds.), 2018 May 8, Miyazaki, Japan. Paris: European Language Resources Association, 2018.

https://doi.org/10.5808/GI.2019.17.2.e16

G&I
Genomics & Informatics

Application note

eISSN 2234-0742
Genomics Inform 2019;17(2):e13
https://doi.org/10.5808/GI.2019.17.2.e13

Received: December 13, 2018
Revised: May 20, 2019
Accepted: May 20, 2019

*Corresponding author:
E-mail: jbanda@gsu.edu

© 2019, Korea Genome Organization
This is an open-access article distributed under the terms of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Fully connecting the Observational Health Data Science and Informatics (OHDSI) initiative with the world of linked open data

Juan M. Banda*

Panacea Laboratory, Department of Computer Science, Georgia State University, Atlanta, GA 30303, USA

The usage of controlled biomedical vocabularies is the cornerstone that enables seamless interoperability when using a common data model across multiple data sites. The Observational Health Data Science and Informatics (OHDSI) initiative combines over 100 controlled vocabularies into its own. However, the OHDSI vocabulary is limited in the sense that it combines multiple terminologies and does not provide a direct way to link them outside of their own self-contained scope. This issue makes the tasks of enriching feature sets by using external resources extremely difficult. In order to address these shortcomings, we have created a linked data version of the OHDSI vocabulary, connecting it with already established linked resources like bioportal, bio2rdf, etc. with the ultimate purpose of enabling the interoperability of resources previously foreign to the OHDSI universe.

Keywords: clinical informatics, common data model, controlled vocabularies, linked open data, RDF, semantic web
Availability: The resource described in this article is available in two different ways: fully constructed RDF graph (GRAPHcompressed files), https://github.com/thepanacealab/OHDSI2RDF; scripts to generate RDF graph, https://github.com/thepanacealab/OHDSI2RDF.

Introduction

The Observational Health Data Science and Informatics (OHDSI) is a world-wide initiative, which over the course of five years has managed to bring groups of researchers all over the world together in converting their clinical patient data (electronic health records, claims, clinical registries) into the Observational Medical Outcomes Partnership (OMOP) common data model (CDM). This initiative has built a large set of publicly available tools which allow researchers to standardize the way they build patient cohorts, characterize their data [1], perform large scale patient level prediction studies [2], and perform electronic phenotyping [3]. In just a few years the OHDSI initiative has managed to perform large-scale studies involving over 200 million patients [4], answer drug safety questions by analyzing the association of the anticonvulsant levetiracetam with increased risk for angioedema in 10 international databases [5], and has characterized the effectiveness of second-line treatment of type 2 diabetes after initial therapy with metformin in over 246 million patients [6]. All of these massive studies have been made possible thanks to the use of a CDM and a standardized vocabulary. This strength becomes a weakness as the vocabulary standardizes multiple external vocabularies, ontologies and term sets, such as SNOMED, RxNorm, MeSH, and 90+ others, but it does not provide an easy way to link

them to additional resources such as the Unified Medical Language System (UMLS) [7] and other linked open data resources like Bio2rdf [8] and BioPortal [9]. During our time at the Biomedical Link Data Hackathon 5 in Kashiwa, Japan we developed the first attempt to create an RDF version of the OHDSI vocabulary with linkages to UMLS and BioPortal.

Methods

In order to link the OHDSI vocabulary with UMLS, we will leverage Ananke [10], a resource built for the mapping of UMLS Concept Unique Identifiers (CUIs) into OHDSI concept_id's, which are the unique identifiers assigned to all concepts in the vocabulary. This will allow us to use BioPortals URI's for the CUIs and make the necessary connections when using their SPARQL endpoints for federated queries. All other Python 2.7 code just iterates through the vocabulary concepts, find proper UMLS matches and writes out each entry using a predefined schema. The conversion process assumes the OHDSI vocabulary files are in the same folder, as well as the Ananke mappings. If the researcher does not have a full copy of the OHDSI vocabulary, we provide an already built RDF graph for Vocabulary version v5.0 11-FEB-19.

Results and Discussion

The RDF conversion results in a total of 24 million triples and takes around 15 minutes. Our resource links a total of 861,732 OHDSI concept_id's from SNOMED, 286,256 concept_id's from RxNORM, 109,706 concept_id's from ICD10, and 22,029 concept_id's from ICD9, all linked directly to bioportal. We also include 1,321,986 mappings to UMLS via Ananke [10].

Our initial goals for this resource were to bring into the OHDSI context semantic enrichment of longitudinal clinical study data, as it has been shown to be quite effective in the past [11,12]. Our particular practical application of interest is taking advantage of the resource for electronic phenotyping purposes. As authors of the Automated PHenotype Routine for Observational Definition, Identification, Training and Evaluation (APHRODITE) R package [3], our goals were as follows.

(1) Be able to expand and enrich our feature sets for phenotyping. With one of the main feature spaces of APHRODITE being clinical narratives, these are annotated using the OHDSI vocabulary. Having a linked version of it will allow us to expand any particular feature domain with other linked resources to SNOMEDCT, RxNORM, etc. Fig. 1 shows a sample query were we expand the SNOMED concept for "Type 2 diabetes mellitus" with all its available parents in BioPortal via a federated query.

(2) One of the outputs of APHRODITE, besides a machine learning model for the target phenotype, is a list of relevant features that add interpretability to any model. This list of features covers the most important domains in the OHDSI CDM and vocabulary. We want to be able to produce this list as a linked resource that will allow researchers to enhance their understanding by being able to semantically link them to other resources like the Human Phenotype Ontology [13] among others.

We believe that such interoperability will enable other researchers to generate enhanced evidence by linking outside of the OHDSI CDM and vocabulary with additional resources available, such as phenotype annotations from PubMed abstracts automatically [14], provide extra context for word embeddings models built from clini-

```
PREFIX map: <http://protege.stanford.edu/ontologies/mappings/mappings.rdfs#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
SELECT DISTINCT ?mappedParent WHERE {
    SERVICE <http://deepml.gsu.edu/sparql> {
        ?mapping map:target <http://athena.ohdsi.org/search-terms/terms/201826> .
        ?mapping map:source ?source .
    }
    SERVICE <http://sparql.bioontology.org/ontologies/sparql/?apikey=YOUR API KEY HERE> {
        ?source rdfs:subClassOf ?mappedParent .
    }
}
```

Fig. 1. Sample federated SPARQL query to retrieve parent elements for a specific SNOMED concept.

https://doi.org/10.5808/GI.2019.17.2.e13

cal narratives [15], which in theory can help the embeddings be more specific by providing additional context [16], and many additional applications. This resource brings us one step closer to enrich EHR, claims, and registry patient data with the world of linked open data.

ORCID

Juan M. Banda: https://orcid.org/0000-0001-8499-824X

Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

We would like to acknowledge DBCLS and ROIS for making our participation in BLAH (Biomedical Linked Annotation Hackathon) possible.

References

1. Atlas. San Franciso: GitHub, 2019. Accessed 2019 Apr 2. Available from: https://github.com/OHDSI/Atlas.
2. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. J Am Med Inform Assoc 2018;25:969-975.
3. Banda JM, Halpern Y, Sontag D, Shah NH. Electronic phenotyping with APHRODITE and the Observational Health Sciences and Informatics (OHDSI) data network. AMIA Jt Summits Transl Sci Proc 2017;2017:48-57.
4. Hripcsak G, Ryan PB, Duke JD, Shah NH, Park RW, Huser V, et al. Characterizing treatment pathways at scale using the OHDSI network. Proc Natl Acad Sci U S A 2016;113:7329-7336.
5. Duke JD, Ryan PB, Suchard MA, Hripcsak G, Jin P, Reich C, et al. Risk of angioedema associated with levetiracetam compared with phenytoin: findings of the observational health data sciences and informatics research network. Epilepsia 2017;58:e101-e106.
6. Vashisht R, Jung K, Schuler A, Banda JM, Park RW, Jin S, et al. Association of hemoglobin A1c levels with use of sulfonylureas, dipeptidyl peptidase 4 inhibitors, and thiazolidinediones in patients with type 2 diabetes treated with metformin: analysis from the observational health data sciences and informatics initiative. JAMA Netw Open 2018;1:e181755.
7. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004;32:D267-D270.
8. Callahan A, Cruz-Toledo J, Ansell P, Dumontier M. Bio2RDF release 2: improved coverage, interoperability and provenance of life science linked data. In: The Semantic Web: Semantics and Big Data (Cimiano P, Corcho O, Presutti V, Hollink L, Rodolph S, eds.), 2013 May 26-30, Montpellier, France. Berlin: Springer, 2013. pp. 200-212.
9. Salvadores M, Alexander PR, Musen MA, Noy NF. BioPortal as a dataset of linked biomedical ontologies and terminologies in RDF. Semant Web 2013;4:277-284.
10. OHDSIananke. San Franciso: GitHub, 2019. Accessed 2019 Apr 2. Available from: https://github.com/thepanacealab/OHDSI-ananke.
11. Leroux H, Lefort L. Semantic enrichment of longitudinal clinical study data using the CDISC standards and the semantic statistics vocabularies. J Biomed Semantics 2015;6:16.
12. Blair DR, Wang K, Nestorov S, Evans JA, Rzhetsky A. Quantifying the impact and extent of undocumented biomedical synonymy. PLoS Comput Biol 2014;10:e1003799.
13. Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. Am J Hum Genet 2008;83:610-615.
14. Henderson J, Bridges R, Ho JC, Wallace BC, Ghosh J. PheKnow-Cloud: a tool for evaluating high-throughput phenotype candidates using online medical literature. AMIA Jt Summits Transl Sci Proc 2017;2017:149-157.
15. Ling Y, An Y, Liu M, Hasan SA, Fan Y, Hu X. Integrating extra knowledge into word embedding models for biomedical NLP tasks. In: 2017 International Joint Conference on Neural Networks (IJCNN), 2017 May 14-19, Anchorage, AK, USA. Hoffman Estates: International Joint Conference on Neural Networks, 2017. pp. 968-975.
16. Dubois S, Romano N, Kale DC, Shah N, Jung K. Effective representations of clinical notes. Ithaca: arXiv, Cornell University, 2017. Accessed 2019 Apr 2. Available from: http://arxiv.org/abs/1705.07025.

G&I
Genomics & Informatics

Application note

eISSN 2234-0742
Genomics Inform 2019;17(2):e14
https://doi.org/10.5808/GI.2019.17.2.e14

Received: March 15, 2019
Revised: May 20, 2019
Accepted: May 20, 2019

*Corresponding author:
E-mail: ljgarcia@ebi.ac.uk

© 2019, Korea Genome Organization
This is an open-access article distributed under the terms of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Biotea-2-Bioschemas, facilitating structured markup for semantically annotated scholarly publications

Leyla Garcia[1,2*], Olga Giraldo[3], Alexander Garcia[4], Dietrich Rebholz-Schuhmann[5]

[1]EMBL-EBI, Wellcome Genome Campus, Hinxton, CB10 1SD, UK
[2]ELIXIR Hub, Wellcome Genome Campus, Hinxton, CB10 1SD, UK
[3]Ontology Engineering Group, Campus de Montegancedo, Boadilla del Monte, Universidad Politécnica de Madrid, Madrid, Spain
[4]BASF SE, G-FDR/BI-G200, 67056 Ludwigshafen am Rhein, Germany
[5]ZB MED-Information Centre for Life Sciences, 50931 Köln, Germany

The total number of scholarly publications grows day by day, making it necessary to explore and use simple yet effective ways to expose their metadata. Schema.org supports adding structured metadata to web pages via markup, making it easier for data providers but also for search engines to provide the right search results. Bioschemas is based on the standards of schema.org, providing new types, properties and guidelines for metadata, i.e., providing metadata profiles tailored to the Life Sciences domain. Here we present our proposed contribution to Bioschemas (from the project "Biotea"), which supports metadata contributions for scholarly publications via profiles and web components. Biotea comprises a semantic model to represent publications together with annotated elements recognized from the scientific text; our Biotea model has been mapped to schema.org following Bioschemas standards.

Keywords: biomedical text mining, literature metadata, semantic annotations, structured data, web page markup
Availability: Data and software used are available at https://doi.org/10.5281/zenodo.2595281.

Introduction

The high and sustained growth rate in scholarly publications requires ever more efficient ways to identify the most relevant documents for the own research work and increasingly, the scientific text is exploited to guide this search. This requires Named Entity Recognition and Text-Mining approaches, but also effective ways to expose identified data as structured metadata and to interlink to data and metadata with external sources such as ontologies and fact repositories, which is still a challenging research domain. The Journal Article Tag Suite (JATS) serves as the preferred (semi)structured approach for publishers to provide machine-readable access to scholarly articles [1]. Although XML is a well-established and shared format, it falls behind in the efficient use of semantics, where Resource Data Framework (RDF)/XML or JSON-LD deliver better results. However, deploying a RDF infrastructure approach can only be achieved with significant overheads in comparison to offering RESTful services for the reason that data (and metadata) is more readily available through RESTful services in contrast to data provision through XML repositories. Schema.org forms a lightweight alternative, i.e., allows semantic annotation at low development overheads, in comparison to complete semantic environments based on ontologies and

RDF approaches.

Schema.org is a collaborative initiative offering a simple yet effective way to add structured metadata to web pages via markup: making pages more findable and improving search results as search engines can identify better the clues what a page is about. Schema.org also eases the ways to interlink related resources; for instance, a movie can be linked to its actors, directors or similar movies. Such interlinking of web resources can boost interoperability, allowing search engines to move from mainly presenting search only to providing comprehensive summaries. Scholarly literature repositories such as Zenodo and EuropePMC have recognized the potential of schema.org and currently support markup on their pages. However, they mostly limit to metadata, i.e., authors, title, abstract and journal, leaving aside any entity of interest from the life science domain that could be recognized from the text. Despite its popularity, the high diversity of properties in schema.org makes it still difficult for not well prepared data scientists to adopt such markup. Bioschemas, a community and collaborative project, focuses on integrating Life Science types to schema.org and, at the same time, makes its adoption easier via profiles tailored to the Life Science domain [2]. A profile comprises a fixed combination (a "set") of guidelines regarding minimum, recommended and optional properties, well-known ontology terms to be used together with the properties and examples, so that users can use them as a template to markup their own resources.

Biotea proposes a conceptual model to represent scholarly publications as Linked Data [3,4], covering not only the metadata but also the article structure, content and semantic annotations, where a semantic annotation corresponds to a named entity recognized from the text. Biotea currently supports annotations obtained from the National Center for Biomedical Ontology (NCBO) annotator service [5]; however, additional annotators could be added. In order to facilitate the adoption of schemas.org markup for scholarly articles as well as to add support to semantic annotations, Biotea has joined the Bioschemas effort by adding five scholarly publication related profiles. Here we report on the Biotea draft profiles proposed to Bioschemas and on two web components capable to render schema.org markup for scholarly articles metadata together with annotations, using the PubMed Central Open Access subset (PMC-OA) as main input. As a result, we anticipate a knowledge graph derived from the scientific literature and based on Bioschema. This type of graph eases the way towards co-citation networks, author-expertise networks, and ontology-based associations among others.

Biotea Draft Profiles for Bioschemas

Biotea profiles proposed to Bioschemas (as a draft) include journal, volume, issue, scholarly article and semantic scholarly annotation (http://bioschemas.org/groups/Biotea2Bioschemas). These profiles corresponding to journal, volume, issue and scholarly article are reasonably stable as they correspond to regular metadata and data that have been well-established for publications, e.g., title, authors, abstract, journal, publisher, pages, dates and citations. Biotea supports two ways to represent annotations, one based on the Annotation Ontology [6] and another one on the Open Annotation Ontology [7]. None of them can be fully mapped to schema.org, therefore modifications are more likely to be introduced to our SemanticAnnotation profiles, so it can adapt well to a variety of annotators and annotated documents. Currently we are using such profiles for annotations recognized in scholarly documents, but it could as well be deployed, for instance, to annotate chemicals in patent documents. As more use cases will be explored, further adjustments will be needed and should be addressed before moving from a draft to a supported profile.

The publication is represented as a schema:ScholarlyArticle part of a periodical publication, i.e., a journal with possibly a volume and issue. The property schema:about is used to link the publication to its annotations which link back to the article via schema:subjectOf. An overview of our Biotea mapping to Bioschemas is shown in Fig. 1.

One of the key aspects in Biotea's mapping to schema.org is the separation between the structured data and the publication itself. Schema.org includes some properties in schema:CreativeWork, that should be used only for a structured data representation. However, those properties are still limited, i.e., do not cover all the needs regarding structured metadata on top of scholarly publications. For instance, structure metadata can be split into different parts, e.g., linksets, but there is only one schema:hasPart property for schema:CreativeWork. In order to allow for more input, we therefore keep the structured data model separated from the publication, the former modelled as a schema:CreativeWork and the latter as a schema:ScholarlyPublication; the structured data is linked to the publication via schema:mainEntity. Through this separation in combination with the use of a lightweight semantic approach such as schemas.org, we aim to improve our approaches for FAIRability for publications, since this type of data gives better compliance (in comparison to JATS/XML) with the Findability, Accessibility, Interoperability and Reusability (FAIR) principles [8].

Biotea-2-Bioschemas Web Components

Web components make it easy to add customized elements to web pages so they can easily be embedded into any web page. In order to parse and render scholarly publication metadata as well as semantic annotations from their text, we have created two web com-

https://doi.org/10.5808/GI.2019.17.2.e14

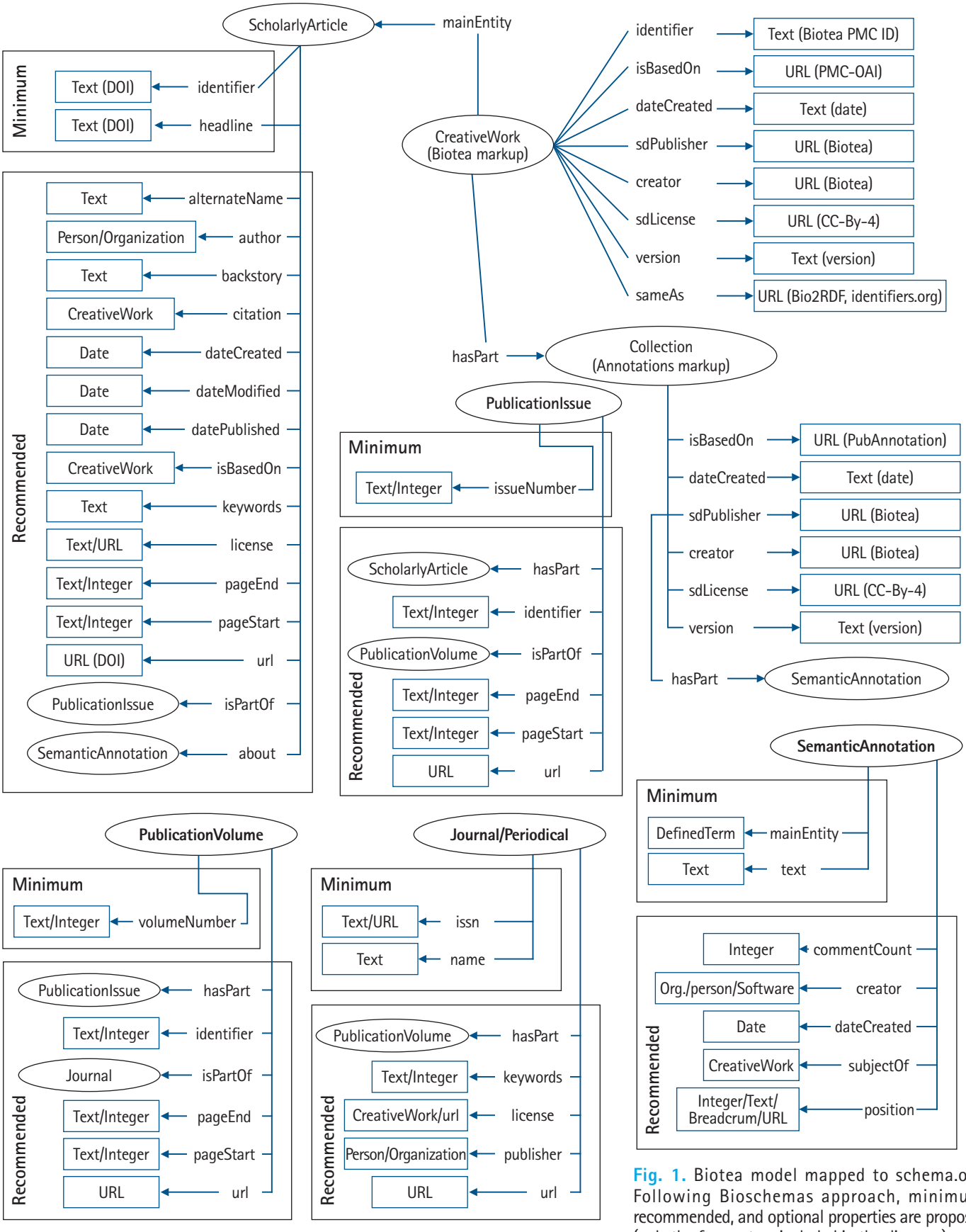

Fig. 1. Biotea model mapped to schema.org. Following Bioschemas approach, minimum, recommended, and optional properties are proposed (only the former two included in the diagram).

@id	https://doi.org/10.1371/journal.pbio.1000016
@type	ScholarlyArticle
identifier	doi:10.1371/journal.pbio.1000016
url	https://doi.org/10.1371/journal.pbio.1000016
alternateName	accession:PMC2628407
	pmcid:PMC2628407
	pmc-uid:2628407
	pmid:19166269
	publisher-id:08-PLBI-RA-2370R3
	sici:plbi-07-01-09
headline	A Novel Function for Fragile X Mental Retardation Protein in Translational Activation
datePublished	2009-1-20
sameAs	http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2628407
	http://info.identifiers.org/pubmed/19166269
	https://www.ncbi.nlm.nih.gov/pubmed/19166269
license	http://creativecommons.org/licenses/by/4.0/

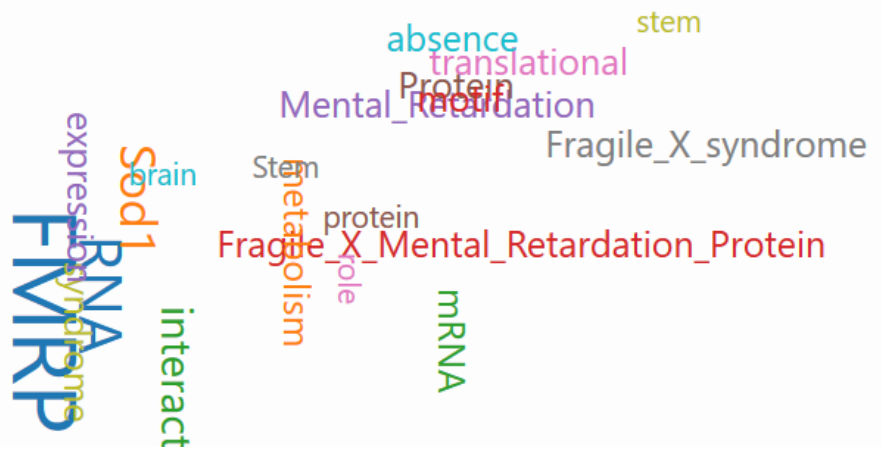

Fig. 2. Overview of the metadata markup showed as a table, annotations with more than 1 occurrence are shown as a cloud of words.

ponents. Our web components take their input, parse it and render it following the corresponding Bioschemas markup. The markup is added at the end of the web page as a new element, such element corresponds to a script with the MIME-type "application-json". We have chosen this asynchronous approach as it makes it easier for web page providers to optimize the end-user content, leaving any Biotea-Bioschemas machine-readable bits to the end. Rendering the markup as JSON-LD also provides readable HTML, as it keeps the content separated from the markup. A possible drawback in this approach is the uncertainty regarding the capacity of crawlers to support asynchronous loading of the structured metadata. Asynchronous loading is nowadays a common approach on web page development and therefore should be supported by main search engines when it comes to structured metadata. Bioschemas has already recognized this and therefore has initiated efforts to contribute with a crawler capable to deal with delayed loading (https://github.com/ricardoaat/bioschemas-gocrawlit).

The Biotea-bioschemas-metadata web component requires as input a JATS XML response from the PMC-OA Interface service (PMC-OAI, https://www.ncbi.nlm.nih.gov/pmc/tools/oai/). It converts such a response into a JSON object and then selects elements of interest such as title authors, publication data and abstract and map them to the Biotea-Bioschemas profiles. For our second web component, biotea-bioschemas-annotations, we use PubAnnotation [9] annotations as main input. Although it is possible to use the NCBO annotator service on the fly, doing so poses a prob-

https://doi.org/10.5808/GI.2019.17.2.e14

lem regarding the obtained annotations. The NCBO annotator works with the current version of the ontologies hosted by the NCBO ontology portal (https://bioportal.bioontology.org), as ontologies are updated, so could annotations using such ontologies, meaning that a different set of annotations for the same article could be recovered at different times. In order to avoid this, we have created a Biotea project in PubAnnotation so we can host there a set of frozen annotations used to provide the corresponding Bioschemas markup.

As a proof of concept, we have uploaded 2,596 full text articles, annotated their abstracts with 13 ontologies from Bioportal, covering gene, protein, drug, disease and symptom areas, for a total of 894,926 annotations, and exposed them via PubAnnotation. Our web components can be seen in action in our Biotea-Bioschemas main page (http://biotea.github.io/bioschemas). Although it is possible to use our web components with any PMC-OA article or any PubAnnotation PMC hosted annotation set, we have limited our showcase to those 2,596 articles currently hosted at PubAnnotation. All of the articles are included in a sitemap, making it easier for search engines to index the content. In Fig. 2 we show the markup corresponding to PMC-OA article identified as PMC2628047, the corresponding URL is http://biotea.github.io/bioschemas?pmc=2628047.

Conclusions and Future Work

By mapping Biotea to schema.org following the standards proposed by Bioschemas, we are preparing the basis towards literature-based knowledge graphs based on Bioschemas. At this point, it remains open what types of inferences based on such a graph could be drawn, under the consideration that the semantics given by schema.org is less strict in comparison to ontology-based approaches such as https://scigraph.springernature.com. Schema.org and Bioschemas are, however, easier to adopt as adding markup to HTML does not require such a robust infrastructure as ontology-based approaches do. As a future work, our initial step on realizing a Bioschemas knowledge graph potential would be extracting author-expertise networks. From there, we hope more developments will come, exploring more complex associations such as target-disease.

ORCID

Leyla Garcia: https://orcid.org/0000-0003-3986-0510
Olga Giraldo: https://orcid.org/0000-0003-2978-8922
Alexander Garcia: https://orcid.org/0000-0003-1238-2539
Dietrich Rebholz-Schuhmann: https://orcid.org/0000-0002-1018-0370

Authors' Contribution

Conceptualization: LG, OG, AG. Formal analysis: LG. Funding acquisition: LG, AG. Methodology: LG. Writing – original draft: LG. Writing – review & editing: DRS.

Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

We acknowledge the support of the Elixir BioHackathon 2018 and the Biomedical Literature Annotation Hackathon 2019 as part of the Biotea mapping to Bioschemas and corresponding web components were developed at those events. We also acknowledge approach, efforts, contributions and tools provided by the Bioschemas community.

References

1. Lapeyre D. Introduction to JATS (Journal Article Tag Suite). Rockville: Mulberry Technologies, Inc., 2018. Accessed 2019 Mar 11. Available from: https://www.xml.com/articles/2018/10/12/introduction-jats/.
2. Gray AJ, Goble C, Jiménez RC; The Bioschemas Community. Bioschemas: from potato salad to protein annotation. In: International Semantic Web Conference. Berlin: Springer-Verlag, 2017. Accessed 2019 Mar 11. Available from: https://pdfs.semanticscholar.org/74ec/a9c89622bff731b21b03acb4f2400a0f00fa.pdf.
3. Garcia Castro LJ, McLaughlin C, Garcia A. Biotea: RDFizing PubMed Central in support for the paper as an interface to the Web of Data. J Biomed Semantics 2013;4 Suppl 1:S5.
4. Garcia A, Lopez F, Garcia L, Giraldo O, Bucheli V, Dumontier M. Biotea: semantics for Pubmed Central. PeerJ 2018;6:e4201.
5. Jonquet C, Shah NH, Youn CH, Musen MA, Callendar C, Storey MA. NCBO Annotator: semantic annotation of biomedical data. In: International Semantic Web Conference. Berlin: Springer-Verlag, 2009. Accessed 2019 Mar 11. Available from: http://www.lirmm.fr/~jonquet/publications/documents/Demo-ISWC09-Jonquet.pdf.
6. Ciccarese P, Ocana M, Garcia Castro LJ, Das S, Clark T. An open annotation ontology for science on web 3.0. J Biomed Semantics 2011;2 Suppl 2:S4.
7. Sanderson R, Ciccarese P, Van de Sompel H. Open Annotation Data Model. Open Annotation Community Group, 2013. Accessed 2019 Mar 14. Available from: http://www.openannota-

tion.org/spec/core/.
8. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3:160018.
9. Kim JD, Wang Y. PubAnnotation: a persistent and shareable corpus and annotation repository. In: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing, 2012 Jun 8, Montreal, Canada. Stroudsburg: Association for Computational Linguistics, 2012. pp. 202-205.

https://doi.org/10.5808/GI.2019.17.2.e14

G&I
Genomics & Informatics

Application note

eISSN 2234-0742
Genomics Inform 2019;17(2):e15
https://doi.org/10.5808/GI.2019.17.2.e15

Received: March 15, 2019
Revised: May 27, 2019
Accepted: May 27, 2019

*Corresponding author:
E-mail: felipe.soares@bsc.es

© 2019, Korea Genome Organization
This is an open-access article distributed under the terms of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

PharmacoNER Tagger: a deep learning-based tool for automatically finding chemicals and drugs in Spanish medical texts

Jordi Armengol-Estapé[1], Felipe Soares[2*], Montserrat Marimon[2], Martin Krallinger[2,3]

[1]Universitat Politècnica de Catalunya (UPC), 08034 Barcelona, Spain
[2]Barcelona Supercomputing Center (BSC), 08034 Barcelona, Spain
[3]Centro Nacional de Investigaciones Oncológicas (CNIO), 28029 Madrid, Spain

Automatically detecting mentions of pharmaceutical drugs and chemical substances is key for the subsequent extraction of relations of chemicals with other biomedical entities such as genes, proteins, diseases, adverse reactions or symptoms. The identification of drug mentions is also a prior step for complex event types such as drug dosage recognition, duration of medical treatments or drug repurposing. Formally, this task is known as named entity recognition (NER), meaning automatically identifying mentions of predefined entities of interest in running text. In the domain of medical texts, for chemical entity recognition (CER), techniques based on hand-crafted rules and graph-based models can provide adequate performance. In the recent years, the field of natural language processing has mainly pivoted to deep learning and state-of-the-art results for most tasks involving natural language are usually obtained with artificial neural networks. Competitive resources for drug name recognition in English medical texts are already available and heavily used, while for other languages such as Spanish these tools, although clearly needed were missing. In this work, we adapt an existing neural NER system, NeuroNER, to the particular domain of Spanish clinical case texts, and extend the neural network to be able to take into account additional features apart from the plain text. NeuroNER can be considered a competitive baseline system for Spanish drug and CER promoted by the Spanish national plan for the advancement of language technologies (Plan TL).

Keywords: machine learning, natural language processing, neural networks (computer)
Availability: PharmacoNER Tagger can be accessed at https://github.com/PlanTL-SANIDAD/PharmacoNER.

Introduction

Chemical compounds and drugs represent a key biomedical entity of common interest for a range of scientific disciplines, including medicine and clinical research, pharmacology as well as basic biomedical research. Due to the growing amount of clinical texts, medical and biomedical literature and medicinal chemistry patents, a systematic approach to recognize such entities in order to semantically enrich these documents and enable further relation extraction task is needed. Text mining and information extraction efforts are gradually being adopted to empower the transformation of unstructured running texts to more structured data representations that can be directly consumed by content analytics and information retrieval infrastructures [1].

Semantically labeling drug mentions in medical documents is a necessary step to allow a range of downstream text mining applications, like drug-resistance, drug dosage recognition, duration of medical treatments, drug repurposing, detection of medication-related allergies, drug-drug interactions or disease-drug relations. Drug names show some inherent characteristics such as the constraints imposed by the International Nonproprietary Names (INN) system by exploiting certain suffixes and prefixes to group them according to certain attributes. For instance, the suffix '-caine' is usually used for local anesthetics. These properties are not only useful for healthcare professionals and pharmacologists to memorize drug names, their attributes and fining them in the clinical literature, they are also constituting valuable attributes for automatically labeling them through language technology tools.

Named entity recognition (NER) [2] is the task of automatically extracting and identifying mentions of entities of interest in running text, typically through their mention offsets or by classifying individual tokens whether they belong to entity mentions of not. Among many other uses, such as entity-aware machine translation, NER can also be used for the automated information extraction and anonymization [3] of medical texts. Early techniques involved the use of hand-crafted rules. The most successful approach consisted in using Conditional Random Fields (CRFs), a graph-based model for modeling sequences that still relied on features defined by humans. In the last years, Natural Language Processing (NLP) techniques have been improved with the use of word embeddings, that is to say, projections of words to vector spaces such that among other findings, the distance can be used as a semantic similarity measure [4]. Deep learning (e.g., artificial neural networks [ANNs] with a high number of hidden layers) has proved to be extremely useful in solving NER problems. In this article, we introduce PharmacoNER tagger, NER software based on NeuroNER [5] adapted to the domain of Spanish medical texts and improved with non-domain-specific features.

The problem

Medical documents mentioning drugs and chemical compounds offer a relevant source of information regarding the patients' treatment. For instance, doctors and researchers may be interested in finding health records of patients treated with a particular kind of compound in order to manually analyze them or use them in biomedical NLP pipelines.

Nevertheless, the task of manually detecting these entities would be expensive and time-consuming, which prompts for the development of NER systems for automating this task. In the context of chemical compounds, NER is usually referred to as chemical entity recognition (CER).

Challenges in NER for medical texts and chemical compounds

NER itself has proven to be a challenging task because of the ambiguity and complexity of human languages. In the particular case of CER, additional difficulties arise.

First of all, not all chemical names present distinctive patterns as far as name segments or chemical word morphology are concerned. CER systems are particularly sensitive with regard to both spelling errors and the tokenization strategy of choice since chemical documents usually exhibit hyphenated text segments or variable use of parentheses, brackets, dots, and other punctuation signs. In addition, chemical documents tend to be loaded with acronyms and abbreviations, which are one of the main sources of false positives.

Another characteristic that makes CER difficult is the fact that the detection of mention boundaries is especially cumbersome when long chemicals or modifiers are present. A more in-depth description of difficulties in tagging chemicals can be found at [6].

Other approaches

In the past, most NER systems were rule-based or leveraged graph-based models such as CRFs, while nowadays the task is usually approached with deep learning techniques. In the case of CER systems, we can highlight some relevant works.

ChemicalTagger [7] parsed the text with a formal grammar and domain-specific regular expressions (regex) and used the parse tree in combination with the Part-of-Speech (POS) tags obtained with an English tagger in order to extract chemical entities. They reported achieving machine-annotator agreements of 88.9% for phrase recognition and 91.9% for phrase-type identification. The test corpus was assembled by compiling 50 paragraphs from the experimental sections of polymer synthesis related papers.

CheNER [8], a tool for the identification of chemical entities and their classes in the biomedical literature, is based on a combination of CRFs, regex and dictionary matching. An F-Score of about 73% was reported, with minor differences depending on the particular experiment. For evaluation, two corpora containing 3,500 documents with approximately 29,500 annotated chemical entities divided into several classes were used.

On the other hand, MetaMap [8], a tool for recognizing Unified Medical Language System (UMLS) concepts in the text, mapped entities to UMLS concepts by the means of rules and a parser.

In CHEMDNER [6], a drug and chemical names extraction competition, the top scoring teams obtained F-scores of 87.39% and 88.2% depending on the particular task, employing CRFs and domain-specific rules. These scores were obtained in the CHEMDNER corpus, a dataset with chemical compounds manually annotated by domain experts.

With regard to drug recognition, a vast survey on approaches and

https://doi.org/10.5808/GI.2019.17.2.e15

resources was presented in Liu et al.'s study [9], including classical machine learning techniques and hybrid approaches. The authors suggested using deep learning techniques for drug recognition as future work.

A system for drug name recognition and classification in biomedical texts was introduced in Segura-Bedmar et al.'s study [10]. By combining information obtained by the UMLS MetaMap Transfer (MMTx) program and nomenclature rules recommended by the World Health Organization (WHO) INNs Program, a broader coverage than previous approaches based on standalone MMTx was achieved.

For an extensive review of text mining techniques for drugs and chemical compounds, readers can refer to Vazquez et al.'s study [11].

Nevertheless, the aforementioned works used techniques that have some important limitations [1]. Dictionary-based methods have problems dealing with name variability since chemical naming exhibits high variation. Rule construction, apart from requiring domain knowledge and an extensive manual workload, has proven to be difficult to scale to new rules and to transfer to slightly different domains. On the other hand, statistical models such as CRFs and classic machine learning algorithms are not powerful enough for detecting some patterns and require a certain amount of feature engineering. All these previous approaches focused on data in English, despite the fact that there is a considerable amount of textual data and repositories (e.g., MEDES, Scielo, Ibecs, or Cuiden), as well as a large medical end users community interested in tools for processing data in Spanish. Regarding other languages, this work [12] presented a system based on CRFs that was capable of recognizing entities in French biomedical documents.

Our proposal

Our proposal involves the adaption of a state-of-the-art NER (i.e., NeuroNER) based on deep learning to the particular domain of the Spanish medical texts in order to identify identities as proteins or other components. Remarkably, the system we base our work on has a CRF layer apart from deep learning components. The original system takes tokens as its input, while our proposal involves using additional features in the neural network.

Key innovations

Our contribution consists in extending an existing neural NER system with additional features that had worked with classic CERs. In particular, POS tags, gazetteer features, and affixes features have been added to the network and were as a pre-preprocessing step. Different experiments have been conducted in order to determine whether these features improved the performance of the neural network. The extension of the neural network is domain and language-agnostic, while the pre-processing is domain-dependent and specific tools for Spanish medical texts have been used. By the neural network itself being language-agnostic we mean that our system could be adapted to other languages by building gazetteer and affix dictionaries and using a POS tagger trained for the required language.

NeuroNER

NeuroNER, an open-source program for named-entity recognition, achieved state-of-the-art performance by having a neural architecture containing three layers: (1) character-enhanced token-embedding layer, (2) a label prediction layer, and (3) a label sequence optimization layer.

Recurrent neural networks (RNNs) are ANNs such that the computational graphs have cycles, which are used for dealing with sequences [13]. Long short term memory (LSTM) [14] are gated units widely used in RNN implementations. A Bidirectional LSTM (Bi-LSTM) is an LSTM unit such that two inputs layers with opposite directions are present, which allows the network to get information from both the past and future states simultaneously. NeuroNER makes use of Bi-LSTM for the character-enhanced token-embedding layer, while character embeddings are previously passed through their own LSTM.

In Fig. 1, we show a representation of the ANN in the original NeuroNER implementation. We must notice that this is just a snippet of the structure of one token. The full ANN has connections for the RNN between tokens from the character-enhanced token embedding upwards. For details about this model, readers can refer to Dernoncourt et al. [15].

In addition, NeuroNER has the ability to load pre-trained token embeddings, which may increase final performance and decrease training time. For more details about NeuroNER, please refer to the following related works: Dernoncourt et al. [5,15].

Proposed Method

The following features have been added to both the neural network and the dataset parser of NeuroNER:

- POS features: with our modification, PharmacoNER tagger can parse annotations (both in BRAT or CONLL formats) of POS tags, encode them with a one-hot encoding for each token and include this in the token embedding layer of the network.
- Gazetteer features: our modified version of NeuroNER is able to parse a text file with terms related to the target entities, build a dictionary and assign a positive value for the words that are found in this dictionary. This feature is then concatenated with the character-enhanced token embedding layer. A file with the dictionary

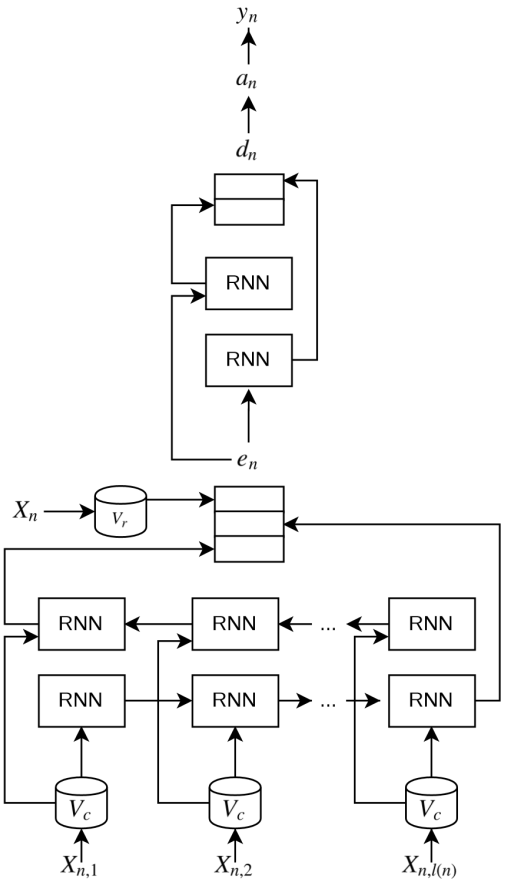

Fig. 1. Schematic representation of the artificial neural networks for a single token in NeuroNER: Snippet of the architecture used in NeuroNER. The type of recurrent neural network (RNN) is long short term memory (LSTM). n is the number of tokens X_i is the i^{th} token. V_r is the mapping from tokens to token embeddings. *l(i)* is the number of characters, and $X_{i,j}$ is the j^{th} character of the i^{th} token. V_c is the mapping from characters to character embeddings. ei is the character-enhanced token embeddings of the i^{th} token. di is the output of the LSTM of label prediction layer, a_i is the probability vector over labels, and y_i is the predicted label of the i^{th} token. Adapted from Dernoncourt et al. J Am Med Inform Assoc 2017;24:596-606 [15].

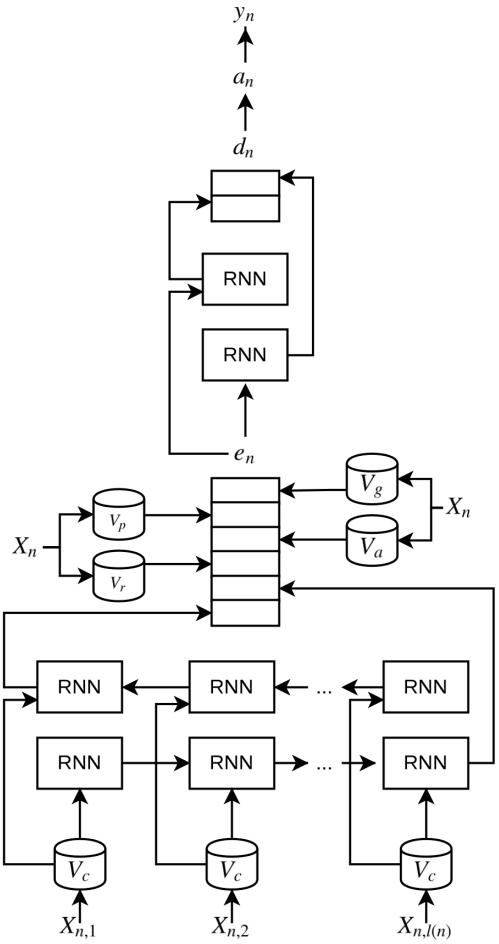

Fig. 2. Schematic representation of the the artificial neural networks for a single token in PharmacoNER tagger: Snippet of the architecture used in PharmacoNER tagger. All notations are the same as in Fig. 1. We added the following features in the character-enhanced token embeddings: V_p is the mapping from the specific token to its Part-of-Speech, V_a is boolean mapping to identify if the given token contains any of the affixes in the database, and V_g is a boolean mapping to the gazetteer database. RNN, recurrent neural network.

has to be supplied by the user.

- Affixes features: PharmacoNER tagger can use information regarding affixes in a similar manner as the gazetteer features, but using regular expressions to detect whether a particular word has an affix related to the target entities. A file of affixes shall be supplied by the user.

Notice that the aforementioned features are domain and language independent. The final software can be found in Github (https://github.com/PlanTL-SANIDAD/PharmacoNER), with information about installation and samples. In Fig. 2, the adaptation of the neural network is also depicted, showing the differences from Fig. 1.

Evaluation

Data

For evaluating PharmacoNER tagger, we used the SPACCC dataset, a manually annotated corpus of 1,000 clinical cases written in Spanish and annotated with mentions of chemical compounds, drugs, genes, and proteins. In particular, the set of labels consists of Normalizables (4,398 labels), No Normalizables (50 labels), Proteins (3,009 labels), and Unclear (167 labels). Note that Normalizables and No Normalizables refer to chemical entities. For instance, the word "triglicéridos" (triglycerides) is annotated as Normalizable, while "receptores de progesterona" (progesterone receptors)

https://doi.org/10.5808/GI.2019.17.2.e15

is labelled as Proteins. One can notice that the dataset is heavily imbalanced. The dataset has not been released yet due to an ongoing shared task (http://temu.bsc.es/pharmaconer).

Embeddings

Eight different pre-trained word embeddings have been tested, from the Spanish Billion Word Corpus (SBWC) and data crawled from SciELO (http://www.scielo.org) and health-related categories of Wikipedia. SBWC data is general-domain, while SciELO and Wikipedia are in-domain.

- General domain (Universidad de Chile Spanish Word Embeddings, https://github.com/uchile-nlp/spanish-word-embeddings):
  FastText embeddings from SBWC
  GloVe embeddings from SBWC
- Domain-specific (PlanTL Embeddings, https://github.com/PlanTL-SANIDAD/Embeddings):
  FastText Scielo
  FastText medical Wikipedia
  FastText Scielo + Health Wikipedia
  Word2Vec Scielo
  Word2Vec medical Wikipedia
  Word2Vec Scielo + Health Wikipedia

Sampling

The SPACCC dataset was split with the following proportions: 80% for training, 10% for validation, and 10% for test.

Since the dataset is imbalanced, a stratified splitting has been applied, in order to approximately maintain the same proportions of labels in the three sets.

SPACCC POS tagger and freeling

FreeLing [16] is an open source language analysis tool suite, released under the Affero GNU General Public License of the Free Software Foundation. FreeLing was used for building the SPACCC POS Tagger (https://github.com/PlanTL-SANIDAD), which is a POS tagger trained with Spanish medical texts. SPACCC POS Tagger was successfully applied to the provided data as a pre-processing step.

Gazetteer

The gazetteer dictionary was built with a pharmaceutical "nomenclator" maintained by the Spanish government (https://www.mscbs.gob.es/profesionales/nomenclator.do), which consists of drug names and active components of medicines. For instance, indomethacin is listed as an active ingredient, and therefore the appearances of this particular word are marked as belonging to the dictionary.

List of affixes

The list of affixes was retrieved from a biomedical website (http://blog.nclexmastery.com/drug-stems-prefixes-roots-suffixes/) and translated into Spanish. The final format is a tab-separated file, with the type of affix (e.g., suffix, prefix, or root), the affix in English, the affix in Spanish, an example in English and the Drug class. For PharmacoNER tagger, only the columns type, and affix in Spanish are used.

Metrics

The CONLL evaluation script is used for evaluating the results. The following overall metrics are computed: accuracy, precision, recall, F1 score. For details about the metrics, please refer to Krallinger et al.'s study [1]. For the sake of comparison, we will utilize the F1 score as the primary performance metric.

Results with the best configuration

To train our system, we tested several configurations with various levels, which are now described:

- Pre-trained embeddings: Total of eight models, from general domain and in-domain.
- POS: Yes (using POS) and No (without POS information)
- Gazetteer: Yes (with gazetteer) and No (without gazetteer)
- Affixes: Yes (with affix information) and No (without affix information).

We trained our systems until convergence, that is until no improvement was identified in the development set for at least 10 epochs. In addition, given that the number of samples from the classes No Normalizables and Unclear are too small, we decided to discard those classes. This led our NER system to have the goal of identifying Proteins and Normalizable chemicals. First of all, to narrow down the search space, we turned off all additional features and just tested the word embeddings. Once the best embedding was found, we fixed it and tested the other configurations.

In Table 1, we present the result of our system regarding the metrics mentioned in Section 4.7. The final setting is: POS with Gazetteer turned one. One can notice that there is not a great difference between the validation and the test set in all metrics, thus meaning that no overfitting occurred.

Conclusion

In this article we have introduced PharmacoNER tagger, a neural NER based on an existing state-of-the-art system, NeuroNER. We extended and especially adapted to the particular domain of chemi-

Table 1. Results for the best combination of embedding and features

	Accuracy		Precision		Recall		F1	
	Val	Test	Val	Test	Val	Test	Val	Test
Overall	99.59	99.66	92.37	91.35	89.7	86.89	91.01	89.06
Normalizables	-	-	94.99	93.97	88.97	85.54	91.88	89.56
Proteins	-	-	89.57	87.64	90.54	89.02	90.05	88.33

The features turned on were Part-of-Speech and Gazetteer.

cal compounds in Spanish medical texts. The results in the validation set of the experiments with the SPACCC dataset have showed that the best configuration consisted of the FastText Scielo + Health Wikipedia embeddings, the POS, and the gazetteer, which have proven to be an additional source of information that can be leveraged by neural networks. In the test set, the aforementioned best configuration obtained an F1 score of 89.06. The extension to the network is domain-agnostic and could be used in other fields, but the pre-processing steps have been specifically designed for our domain.

CER involves additional challenges to the ones already present in generic NER. Our work shows that some of the sources of additional information typically used in previous CER systems based on non-neural techniques, such as affixes, can be leveraged as well by state-of-the-art neural NER systems.

We foresee that this resource will be a valuable contribution not only to semantically enhance medical texts in Spanish for pharmacological and drug-related information, but it also highlights a useful approach on how to implement medical NER taggers for languages other than English. Moreover, PharmacoNER tagger will be a useful competitive baseline system and design principle for the upcoming PharmacoNER BioNLP 2019 shared task.

ORCID

Jordi Armengol-Estapé: https://orcid.org/0000-0001-8893-6185
Felipe Soares: https://orcid.org/0000-0002-2837-1853
Montserrat Marimon: https://orcid.org/0000-0002-0211-8681
Martin Krallinger: https://orcid.org/0000-0002-2646-8782

Authors' Contribution

Conceptualization: FS, JAE. Data curation: MM. Funding acquisition: MK. Methodology: FS, JAE. Writing – original draft: JAE, FS. Writing – review & editing: FS, MK.

Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

This work was supported by the Encargo de Gestion SEAD-CNIO of the Spanish National Plan for the Advancement of Language Technologies, the ICTUSnet INTERREG Sudoe programme, the European Union Horizon2020 eTransafe project, grant agreement ID 777365.

References

1. Krallinger M, Rabal O, Lourenco A, Oyarzabal J, Valencia A. Information retrieval and text mining technologies for chemistry. Chem Rev 2017;117:7673-7761.
2. Yadav V, Bethard S. A survey on recent advances in named entity recognition from deep learning models. In: Proceedings of the 27th International Conference on Computational Linguistics (Bender EM, Derczynski L, Isabelle P, eds.), 2018 Aug 20-26, Santa Fe, New Mexico, USA. Stroudsburg: Association for Computational Linguistics, 2018. pp. 2145-2158.
3. Kleinberg B, Mozes M, van der Toolen Y, Verschuere B. NEMANOS: named entity-based text anonymization for open science. OSF Preprints 2017 Jun 4 [Epub]. https://doi.org/10.31219/osf.io/w9nhb.
4. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. Ithaca: arXiv, Cornell University, 2013. Accessed 2019 May 1. Available from: http://arxiv.org/abs/1301.3781.
5. Dernoncourt F, Lee JY, Szolovits P. NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. Ithaca: arXiv, Cornell University, 2017. Accessed 2019 May 1. Available from: http://arxiv.org/abs/1705.05487.
6. Krallinger M, Leitner F, Rabal O, Vazquez M, Oyarzabal J, Valencia A. CHEMDNER: the drugs and chemical names extraction challenge. J Cheminform 2015;7:S1.
7. Hawizy L, Jessop DM, Adams N, Murray-Rust P. ChemicalTagger: a tool for semantic text-mining in chemistry. J Cheminform 2011;3:17.
8. Usie A, Cruz J, Comas J, Solsona F, Alves R. CheNER: a tool for

https://doi.org/10.5808/GI.2019.17.2.e15

the identification of chemical entities and their classes in biomedical literature. J Cheminform 2015;7:S15.
9. Liu S, Tang B, Chen Q, Wang X. Drug name recognition: approaches and resources. Information 2015;6:790-810.
10. Segura-Bedmar I, Martinez P, Segura-Bedmar M. Drug name recognition and classification in biomedical texts. A case study outlining approaches underpinning automated systems. Drug Discov Today 2008;13:816-823.
11. Vazquez M, Krallinger M, Leitner F, Valencia A. Text mining for drugs and chemical compounds: methods, tools and applications. Mol Inform 2011;30:506-519.
12. Ho-Dac LM, Tanguy L, Grauby C, Mby AH, Malosse J, Riviere L, et al. LITL at CLEF eHealth2016: recognizing entities in French biomedical documents. In: CLEF eHealth 2016 (Balog K, Cappellato L, Ferro N, Macdonald C, eds.), 2016 Sep 5-8, Evora, Portugal. hal-0136592. pp. 81-93.
13. Lipton ZC, Berkowitz J, Elkan C. A critical review of recurrent neural networks for sequence learning. Ithaca: arXiv, Cornell University, 2015. Accessed 2019 May 1. Available from: http://arxiv.org/abs/1506.00019.
14. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9:1735-1780.
15. Dernoncourt F, Lee JY, Uzuner O, Szolovits P. De-identification of patient notes with recurrent neural networks. J Am Med Inform Assoc 2017;24:596-606.
16. Padro L, Stanilovsky E. FreeLing 3.0: towards wider multilinguality. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012) (Calzolari N, Choukri K, Declerck T, Dogan MU, Maegaard B, Mariani J, et al., eds.), 2012 May 21-27, Istanbul, Turkey. Paris: European Language Resources Association, 2012. pp. 2473-2479.

G&I
Genomics & Informatics

Application note

eISSN 2234-0742
Genomics Inform 2019;17(2):e19
https://doi.org/10.5808/GI.2019.17.2.e19

Received: February 26, 2019
Revised: May 31, 2019
Accepted: May 31, 2019

*Corresponding author:
E-mail: eckart@ukp.informatik.tu-darmstadt.de

Richard Eckart de Castilho, Nancy Ide, and Jin-Dong Kim contributed equally to this work.

© 2019, Korea Genome Organization
This is an open-access article distributed under the terms of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Towards cross-platform interoperability for machine-assisted text annotation

Richard Eckart de Castilho[1*], Nancy Ide[2], Jin-Dong Kim[3], Jan-Christoph Klie[1], Keith Suderman[2]

[1]UKP Lab, Technical University Darmstadt, 64289 Darmstadt, Germany
[2]Vassar College, Poughkeepsie, NY 12604-0520, USA
[3]Database Center for Life Science, Research Organization of Information and Systems, Kashiwa 277-0871, Japan

In this paper, we investigate cross-platform interoperability for natural language processing (NLP) and, in particular, annotation of textual resources, with an eye toward identifying the design elements of annotation models and processes that are particularly problematic for, or amenable to, enabling seamless communication across different platforms. The study is conducted in the context of a specific annotation methodology, namely machine-assisted interactive annotation (also known as human-in-the-loop annotation). This methodology requires the ability to freely combine resources from different document repositories, access a wide array of NLP tools that automatically annotate corpora for various linguistic phenomena, and use a sophisticated annotation editor that enables interactive manual annotation coupled with on-the-fly machine learning. We consider three independently developed platforms, each of which utilizes a different model for representing annotations over text, and each of which performs a different role in the process.

Keywords: annotation software, biomedical text mining, interoperability
Availability: INCEpTION (https://inception-project.github.io) is available as Open Source Software published under the Apache License 2.0. The LAPPS Grid (https://github.com/lapps) is available as Open Source Software published under the Apache License 2.0. PubAnnotation (https://github.com/pubannotation/pubannotation) is available as Open Source Software published under the MIT License.

Introduction

Natural language processing (NLP) text mining strategies are a recognized means to approach the increasingly urgent need for usable and effective text mining facilities for scientific publications. Numerous platforms and frameworks that support text mining activity have been developed, including the General Architecture for Text Engineering (GATE [1]), CLARIN WebLicht [2], the Language Applications (LAPPS) Grid [3], OpenMinTeD [4], and several systems based on the Unstructured Information Management Architecture (UIMA [5]), e.g., ARGO [6], Apache cTAKES [7], DKPro Core [8]. However, in many cases the full suite of tools and resources required for a given task is not available within any single platform. Attempting to access different functionalities by combining tools and services from different platforms inevitably leads to roadblocks due to a lack of "interoperability" among them, which can demand substantial computational expertise to overcome.

In this paper, we investigate cross-platform interoperability, with an eye toward identifying the design elements of annotation models and processes that are particularly problemat-

ic for, or amenable to, enabling seamless communication among different platforms providing different functionalities. As a case study, we focus on a specific methodology, namely machine-assisted interactive annotation (also known as human-in-the-loop annotation), which requires the ability to freely combine resources from different document repositories, access to a wide array of NLP tools to automatically annotate corpora for various linguistic phenomena, and a sophisticated annotation editor that enables interactive manual annotation coupled with on-the-fly machine learning. We consider three independently-developed platforms, which together provide the required functionalities: a document repository, an NLP service provider, and an interactive annotation tool. Our goal is to shed light on the issues that arise when attempting to make these platforms pairwise interoperable, and determine the extent to which pairwise interoperability entails interoperability across a proxy, e.g., if the text annotation editor and the NLP services are automatically interoperable when communicating via the document repository. Our analysis is the result of a collaboration at the 5th Biomedical Linked Annotation Hackathon (BLAH 5, http://blah5.linkedannotation.org) and takes into account both implemented modifications to the three platforms and proposed changes that are not fully implemented at the time of this writing.

Background and Motivation

Consider the scenario where a researcher wants to investigate recent advances in gene interaction research documented in publications from a "document repository" such as PubMed Central. The researcher will select a set of appropriate texts from the repository and apply a named entity recognition (NER) "text analysis service" to identify potential gene mentions in the data. However, even specialized NER tools [9] for the biomedical domain perform at rates of about 0.56 F1-score, at best. So at this point, human intervention is required to correct mis-identified occurrences of gene names as well as annotate unrecognized gene names. A sophisticated "annotation editor" that learns from the user's activity and can thereby propose new annotations or modifications can significantly increase the speed of the correction process. The revised annotations can then be used to train a machine learning algorithm and applied to other, unannotated texts; results are evaluated, and the training texts are corrected anew, where necessary, by the human user. This overall cycle involving the human-in-the-loop is repeated as many times as necessary until a satisfactory result is obtained. We consider here three platforms, each of which supports some aspect(s) of the process described above, but none of which provides the entire suite of required tools and resources:

PubAnnotation

PubAnnotation [10] is a repository of annotation data sets. It aims at (1) linking annotations contributed by various groups through canonical texts, (2) providing an easy and fine-grained access to the linked annotations through dereferenceable URIs, and (3) enabling search across multiple annotation data sets. It is designed to be an open platform so that it can interact with other systems through a REST API.

The LAPPS Grid

The LAPPS Grid [3] provides a large collection of NLP tools exposed as SOAP (Simple Object Access Protocol) web services, together with a variety of resources commonly used in the domain. The services are made available to users via a web-based workflow development engine (https://galaxy.lappsgrid.org), directly via SOAP calls, and programmatically through Java and Python interfaces. All tools and resources in the LAPPS Grid are rendered mutually interoperable via transduction to the JSON-LD LAPPS Grid Interchange Format (LIF [11]) and the Web Service Exchange Vocabulary (WSEV [12]), both designed to capture fundamental properties of existing annotation models in order to serve as a common *pivot* among them. The basic annotation model underlying the LIF format includes document-level metadata, text, and a set of views, where a view consists of an ID, a list of annotations, and view-specific metadata. LIF documents are meant to be passed along a pipeline of NLP components, where each component creates a new view and adds its annotations to it. Existing views cannot be modified, but their annotations may be copied to a new view if necessary to add or modify names and/or attribute values.

INCEpTION

INCEpTION [13] is a text annotation platform that integrates interactive annotation, knowledge management and corpus creation into a single platform. The system provides "recommenders" that learn from user annotations and provide annotation suggestions. External document repositories can be accessed to search and load documents into INCEpTION for later annotation. The platform aims at a high level of interoperability by supporting common formats and standards for annotation representation and knowledge representation and it offers a remote API allowing it to be integrated into external workflows. It is based on the UIMA CAS [14] data model. In addition to supporting the definition of a custom annotation schema, several of the annotation types defined by DKPro Core [8] come pre-configured (e.g., Part-of-Speech [POS], Named Entity, etc.).

We envision a scenario where, for example, documents can be retrieved from PubMed Central via PubAnnotation, automatically an-

https://doi.org/10.5808/GI.2019.17.2.e19

notated using LAPPS Grid services, and manually annotated/corrected using INCEpTION (where INCEpTION can use LAPPS Grid services to automatically generate annotation suggestions). However, at present combining the relevant functionalities of each of the three platforms is not fully achievable, due to a lack of cross-platform interoperability. Inter-platform interoperability among the platforms is fundamentally a function of their ability to exchange data consisting of text and associated annotations. This means that the data must be mutually "understandable," either directly or via trivial conversion, and that it must further be possible to appropriately utilize data from the other platforms within the constraints of their respective architectures. To address both of these considerations, in the following sections we consider two levels of interoperability for each pair of platforms: the "data level" (model and schema) interoperability, and "process level" (triggering of and reacting to events) interoperability.

Data-Level Interoperability

At the data level, we investigate to which degree information is preserved or lost when converting data from one format to another or when mapping data from one schema to another. By "annotation model" (short: model), we refer to the basic building blocks (e.g., spans, relations, attributes) which are largely independent of the domain in which the annotation takes place. By "annotation schema," we refer to domain-specific categories ranging from linguistic categories such as part-of-speech, named entities, and dependency relations, to domain-specific categories such as proteins or habitats.

Table 1 shows a point-by-point comparison of the annotation formats of the three platforms. Fig. 1 shows a coreference annotation example represented in the original formats of the three platforms, illustrating their difference.

Table 1. Comparison of annotation model features between LAPPS/LIF, INCEpTION, and PubAnnotation

Feature	LAPPS/LIF	INCEpTION	PubAnnotation
Annotations	LIF Annotations are JSON-LD objects that have the following properties: ID, type, label, start, end, features, and metadata. Metadata and features are both key-value maps. References between annotations are encoded as ID references.	UIMA annotations are feature structures which have the built-in properties: "sofa" (subject of analysis), "begin", "end". References between annotations (feature structures) are object references, so IDs are not required.	Triple representation serialized in JSON. The format is motivated by Resource Description Framework (RDF).
Spans	Subtypes of "Region" (can refer to multiple other regions [e.g., "Markable"] to represent discontinuous spans)	Subtypes of "Annotation". INCEpTION has no provisions for discontinuous annotations.	A denotation is a JSON object which connects a span (or a set of spans for discontinuous spans) to an object.
Relations	Subtypes of "Relation". The individual subtypes define the endpoints of the relation, e.g., dependency defines a "governor" and "dependent". Relations are not necessarily binary. For example, constituent defines an optional parent as well as a list children.	Relations are annotations which have exactly two attributes that refer to other span annotations. For example, the dependency type defines the attributes "Governor" and "Dependent" which both point to "Token" annotations. Relations may have additional primitive attributes. There is no common supertype for all relation types.	A relation is a JSON object, which represents a typed, directed, binary relation, to connects two denotation objects.
Chains	The "Coreference" type. Links between the chain elements are not explicitly modelled and cannot be labeled.	Linked lists of spans where span and link can both have a label.	No dedicated annotation type for chains. However, a chain can be represented by a combination of denotations and relations.
Attributes of annotation instances	Attributes are stored in the "features" map of the LIF JSON-LD object.	Attributes are fields in UIMA feature structures which are used to represent annotations	An attribute is a JSON object which resembles a relation, but it is meant to add further information to denotations and relations.
Complex attributes	Attribute values are expected to be primitive, references to other annotations, or consist of nested feature sets. Sets and lists of references are supported.	Complex attribute values can be encoded as subtypes of "TOP". However, INCEpTION uses such complex attributes, e.g., to model argument slots on semantic predicates.	Complex attributes can be encoded using a naming convention.
Multi-valued attributes	Unordered sets and ordered lists/arrays are supported.	UIMA supports multi-valued features (e.g., via arrays) and INCEpTION uses this internally in some cases. However, user-created features can presently not be multi-valued.	Instead of having multi-valued attributes, in PubAnnotation an attribute can be added multiple times with the same subject/predicate but different objects. This resembles a set behavior.
Document level annotation	Features of "Document", plus those inherited from Thing.	Subtypes of "AnnotationBase", e.g., "DocumentMetaData"	Attributes with the document itself at the subject position.

LAPPS, Language Applications; LIF, LAPPS Grid Interchange Format; UIMA, Unstructured Information Management Architecture.

A

```
<xmi:XMI xmi:version="2.0"
    xmlns:cas="http:///uima/cas.ecore"
    xmlns:coref="http:///de/tudarmstadt/ukp/dkpro/core/api/coref/type.ecore">
  <coref:CoreferenceChain xmi:id="1" sofa="4" first="2"/>
  <coref:CoreferenceLink xmi:id="2" sofa="4" begin="0" end="4"
    next="3"
    referenceType="PER"
    referenceRelation="sameAs"/>
  <coref:CoreferenceLink xmi:id="3" sofa="4" begin="17" end="19"
    referenceType="PER"/>
  <cas:Sofa xmi:id="4" sofaString="John is thirsty. He drinks." members="1 2 3"/>
</xmi:XMI>
```

B

```
{
  "text": "John is thirsty. He drinks.",
  "tracks": [
    {
      "project": "named-entity-annotation-example",
      "denotations": [
        {"id": "T1", "span": {"begin": 0, "end": 4}, "obj": "Person"},
        {"id": "T2", "span": {"begin": 17, "end": 19}, "obj": "Person"}
       ],
       "namespaces": [
         {
           "prefix": "_base",
           "uri": "https://schema.org/"
         }
       ]
    },
    {
      "project": "coreference-annotation-example",
      "denotations": [
        {"id": "T1", "span": {"begin": 0, "end": 4}, "obj": "Antecedent"},
        {"id": "T2", "span": {"begin": 17, "end": 19}, "obj": "Anaphor"}
       ]
       "relations": [
         {"id": "R1", "subj": "T1", "pred": "boundBy", "obj": "T2"}
       ]
    }
  }
}
```

C

```
{
  "text": "John is thirsty. He drinks.",
  "views": [
    { "id": "v1",
       "metadata": {
        "contains": {
          "Token": {
            "producer": "edu.brandeis.cs.lappsgrid.opennlp.Tokenizer:n.n.n",
            "type": "tokenizer:opennlp" }}},
      "annotations": [
        { "@type": "Token", "id": "tok0", "start": 0, "end": 4 },
        { "@type": "Token", "id": "tok1", "start": 17, "end": 19 } ]},
    { "id": "v2",
      "metadata": {
        "contains": {
          "Markable": {
            "producer": "edu.brandeis.cs.lappsgrid.xxxx.coref:n.n.n" },
          "Coreference": {
            "producer": "edu.brandeis.cs.lappsgrid.xxxx.coref:n.n.n" }}},
      "annotations": [
        { "@type": "Markable", "id": "m0", "targets": [ "v1:tok0" ] },
        { "@type": "Markable", "id": "m1", "targets": [ "v1:tok2" ] },
        { "@type": "Coreference",
          "id": "coref0",
          "features": {
            "mentions": [ "m0", "m1" ],
            "representative": "m0" }}]}]
}
```

Fig. 1. Coreference annotation in INCEpTION (A, UIMA XMI), PubAnnotation (B), and LAPPS (C, LIF). UIMA, Unstructured Information Management Architecture; LAPPS, Language Applications; LIF, LAPPS Grid Interchange Format.

https://doi.org/10.5808/GI.2019.17.2.e19

PubAnnotaton ↔ INCEpTION

The PubAnnotation model defines two primary annotation types: "denotations" (to connect text spans with annotations) and "relations" (to connect annotations). INCEpTION defines three types of annotations ("spans," "relations," and "chains"), each of which can be associated with any number of "features." The two data models largely share important annotation concepts, e.g., denotations/spans, and relations. However, PubAnnotation does not currently allow features to be associated with denotations and relations; therefore, to accommodate INCEpTION's (and LAPPS') features, it has been proposed to add "attributes" to the PubAnnotation model (cf. Fig. 2). Other elements of INCEpTION's annotation model require a more non-trivial solution: for example, INCEpTION defines "chain"-type annotation to represent sequentially connected linguistic elements (e.g., coreferences). While PubAnnotation model does not define an element corresponding to INCEpTION's chains, they can be represented by a set of denotations which are sequentially connected by relations (Fig. 1) .

INCEpTION requires a schema defining all annotation types and their features. Since the PubAnnotation model does not explicitly define where information about the type of an annotation is stored, a convention for retaining INCEpTION's schema information must be introduced. We considered two options: (1) store the annotation type in as the "obj" of a denotation or relation; or (2) introduce an attribute with a special name (e.g., "_type," leaving the "obj" available to store a human-readable label that the annotation viewer can use to render an annotation. Storing the type information in "obj" best conforms to PubAnnotation's "semantic-web-spirit" and defers the problem of choosing a suitable label for rendering an annotation to the annotation viewer (cf. Fig. 1B, the "namespaces" definition makes the "Person" tag resolvable to "https://schema.org/Person"). Therefore, it was decided to provide type information in the form of a resolvable URI enabling access to the schema including the type, its possible attributes, their ranges, and the hierarchy of annotation types in which the type resides (e.g., "ex:NamedEntity rdf:type ex:Region"). This approach imposes additional overhead when defining custom annotation types in INCEpTION; to publish annotations using that type to PubAnnotation, the relevant schema must be made available at a URL within a domain they control or in a schema repository (http://schema.org). Ideally, this task would be performed by INCEpTION as a part of the process of publishing annotations to an external repository.

In PubAnnotation, document-level information is represented as attributes of the document itself. INCEpTION models document-level metadata using subtypes of "AnnotationBase" (a built-in UIMA type representing an annotation anchored on an object without specifying what type of object [text, video, audio, etc.] it is. The type "Annotation" used for annotations on text is a subtype of "AnnotationBase").

PubAnnotation does not support complex attributes with nested features (e.g., "first name" and "last name" components in an "author" attribute). Structured attributes are modeled in UIMA as subtypes of "TOP" (The root of the built-in UIMA type hierarchy, custom subtypes of which are typically used to model data structures which are not annotations) and can be converted to PubAnnotation format using an attribute naming conventions. For example, if an entity is associated with a structured attribute called "Author" in UIMA with the fields "first" and "last", this can be represented in PubAnnotation as two simple attributes, "name.first" and "name.last".

A

```
<xmi:XMI xmi:version="2.0"
    xmlns:cas="http:///uima/cas.ecore"
    xmlns:ner="http:///de/tudarmstadt/ukp/dkpro/core/api/ner/type.ecore">
  <ner:NamedEntity xmi:id="1" sofa="2" begin="0" end="4" value="PER"/>
  <cas:Sofa xmi:id="2" sofaString="John is thirsty." members="1"/>
</xmi:XMI>
```

B

```
{
  "text": "John is thirsty.",
  "denotations": [
     {"id": "T1", "span": {"begin": 0, "end": 4},
       "obj": "dkpro-core:de.tudarmstadt.ukp.dkpro.core.api.ner.NamedEntity"}
   ],
   "attributes": [
     {"subj": "T1", "pred": "value", "obj": "PER"},
   ]
}
```

Fig. 2. Named entity annotation in INCEpTION (A), and its conversion into PubAnnotation (B).

PubAnnotaton ↔ LAPPS Grid

Data interoperability between PubAnnotation and the LAPPS Grid faces many of the same challenges as interoperability between PubAnnotation and INCEpTION; therefore, many of the solutions for INCEpTION ↔ PubAnnotation conversion can be used for the LAPPS Grid as well.

As noted before, PubAnnotation's annotation model is schemaless or may optionally rely on an external resource, e.g., a RDF schema or OWL ontology. While the LAPPS Grid recommends the use of the annotation types defined in the WSEV, this is not a strict requirement, and any string or URI can be used to specify the annotation type. This flexibility allows for a trivial "syntactic" mapping between the PubAnnotation JSON format and LIF (This is also true for mapping to a UIMA schema, assuming that the type names follow Java naming conventions). However, the *semantics* of PubAnnotation annotations (i.e., are they POS tags, named entities, etc.) are usually not specified; therefore, it is in many cases necessary to apply LAPPS Grid tools (annotation/feature renamers) to annotations imported from PubAnnotation in order to massage data so that it is accessible by other tools.

To import annotations from the LAPPS Grid to PubAnnotation, type information must be retained, as discussed above in "PubAnnotaton ↔ INCEpTION" for the case of importing annotations from INCEpTION. For LAPPS Grid annotations that reference types in the WSEV via a URI, the URI can be used as the value of *obj* in order to preserve the link between the LAPPS Grid annotation and its definition in the schema. For LAPPS Grid annotations that do not reference types in the WSEV, we can apply the same solution as previously discussed for INCEpTION, that is, LAPPS Grid type information can be preserved in PubAnnotation using the "obj" element.

The LAPPS Grid separates annotations produced by different tools into different "layers," or "views," and supports inter-layer dependency via ID linking. In contrast, PubAnnotation groups annotations over the same text that have been produced by different projects into a single "collection." PubAnnotation provides no support for inter-project dependency (see Fig. 1B); instead, PubAnnotation allows a project to be included in multiple collections. The differences in structuring multiple annotations over the same text reflects the difference in focus between the two frameworks: as a workflow development system, LAPPS Grid is primarily concerned with the way that annotations are created across multiple processes, while as a repository of annotations, PubAnnotation focuses on making it easy to use individual sets of annotations with others that have been applied to the same text. To accommodate this difference, annotations from all LIF views in an annotation document are collapsed into a single collection of annotations (denotations/relations/attributes) in PubAnnotation, retaining information about dependencies between LIF views in atrributes. In the reverse direction, annotations from each project in a PubAnnotation collection can be safely placed into its own view in LIF.

Like INCEpTION, LIF allows complex types including nested features; here we apply the solution implemented for PubAnnotation to INCEpTION conversion, that is, using a set naming convention when generating PubAnnotation attribute names from LIF features.

In summary, there are two main issues that must be addressed to achieve PubAnnotation/LAPPS Grid interoperability. The first is a requirement that PubAnnotation allow for retention of LIF metadata so that this information is not lost in a round trip transaction between the two platforms. The second concerns semantic interoperability, that is, the mapping of annotation types between annotation schemes, which is a problem for cross-platform interoperability in general and remains an open problem for the most part. Semantic interoperability can often be achieved by mapping names defined in one scheme to names denoting the same linguistic phenomenon in the other; however, in the case of PubAnnotation, use of definition-anchored names (URIs) is not mandatory, and therefore most annotation sets use their own arbitrary names. In such cases, manual intervention is required to determine inter-platform correspondences.

LAPPS Grid ↔ INCEpTION

The WSEV defines an ontology of annotation types and their attributes that may be referenced from within LIF annotation documents via unique URIs. Objects produced and consumed by NLP components in LAPPS Grid pipelines are mapped to corresponding terms in the WSEV, which enables ensuring that a given component can use the annotations produced by a previous component. To maximize generality across NLP components, the WSEV is designed so that only higher-level linguistic annotation objects upon which there is considerable consensus in practice (e.g., "Token", "Sentence", "Constituent", "NamedEntity", etc.) are explicitly specified, thereby accommodating the annotation objects produced and consumed by most NLP tools. More specific identifiers, such as part of speech tags, constituent labels, and entity types, are not prescribed, but reference to a defined list used in a given annotation can (should) be given in the metadata for the view in which the annotation lives. The LIF format also allows adding any number of user-specified feature-value pairs to an annotation, as required for a given application; these may be used or ignored by subsequent tools in the pipeline, as appropriate.

While in principle, a user can define custom annotation types in INCEpTION, there is a set of built-in types that conform to the DKPro

https://doi.org/10.5808/GI.2019.17.2.e19

Core type system. While DKPro Core and the WSEV differ in details, there is generally significant overlap between the annotation types they cover. Therefore, in order to make productive use of LAPPS Grid services, it is necessary to perform a format conversion as well as a mapping of INCEpTION annotation types to types defined in the WSEV. This means that custom annotation types defined in INCEpTION may not be readily usable by LAPPS Grid services. On the INCEpTION side, there is presently no concept of grouping annotations by provenance. Thus, when ingesting LAPPS Grid data to INCEpTION, the information encoded in the LIF views identifying the producer (software that produced the annotations) is currently dropped; to enable round-tripping between the LAPPS Grid and INCEpTION, some means to preserve this information in the INCEpTION representation must be implemented. Similarly, INCEpTION does not allow multiple annotations of the same type, and so only the latest LIF view containing a particular annotation type is included. Again, means to represent the information from multiple LIF views involving the same annotation type in INCEpTION would demand creating a special mechanism to accommodate this information, if a round trip between the two platforms is desired.

Process-Level Interoperability

In addition to being able to transport data between the platforms, it is also important that each platform can fulfill its specific role in the cross-platform process. To this end, we investigate here the interactions among the three platforms.

PubAnnotation ↔ INCEpTION

The process-level integration of PubAnnotation and INCEpTION is currently controlled by the INCEpTION side which acts as a client to the PubAnnotation API. PubAnnotation can be connected as an external document repository to the INCEpTION platform, meaning that its contents can be searched from within the INCEpTION UI and documents of interest to the user can be imported for annotation. Currently, the import is limited to the document text (retaining section information). The user can then manually annotate these texts within INCEpTION. Upon finishing the annotation, the user can choose to publish the annotations to the PubAnnotation repository. While it is not necessary to have a PubAnnotation account for searching and importing, publishing annotations requires a working login and the ability to create a PubAnnotation project to which the annotations are published. PubAnnotation recommends that a project should contain only one type of annotation (e.g., "NamedEntity") or a set of closely related annotation types (e.g., "POS" tags and "Dependency" relations), e.g., to avoid visual clutter when viewing the annotations. Depending on the needs of the user, INCEpTION projects may contain many annotation types. When viewing annotations, the user may choose in INCEpTION which types to display and which to hide. For the sake of simplicity, we consider that one INCEpTION project is exported as one PubAnnotation project and do not attempt to split INCEpTION projects up into multiple PubAnnotation projects, even if that means that viewing the annotated data using PubAnnotation's native TextAE visualization may be inconvenient.

PubAnnotation ↔ LAPPS Grid

Documents from PubAnnotation are made available in the LAPPS Grid as data sources, that is, services that provide documents for processing by other services. Documents are retrieved from PubAnnotation using their PubMed ID number. Currently, the LAPPS Grid does not provide any search or query mechanisms for PubAnnotation, and therefore users must determine document ID numbers via other means–for example, by using the PubAnnotation search facility (http://pubannotation.org/search) or the LAPPS Grid AskMe service (https://services.lappsgrid.org/eager/ask). The documents may be downloaded from PubAnnotation as text documents with no annotations or with annotations from PubAnnotation projects included (if available). Annotated documents can also be sent back to the PubAnnotation platform for publication and alignment with the canonical text, so that the annotations are available along with all other published annotations of the same text for others to use.

Several LAPPS Grid services are also made available as annotators on the PubAnnotation platform, so that documents may be annotated directly on the PubAnnotation web site. The tools specifically tuned to biomedical data that are currently available from the LAPPS Grid include:

- Abner biomedical named entity recognizer
- PennBio gene tagger
- TimeML time and event annotator
- Tokenizer and part of speech tagger from Stanford CoreNLP

LAPPS Grid ↔ INCEpTION

INCEpTION integrates interactive annotation support for users to speed up annotating and improve the annotation quality. One kind of support is implemented in form of so-called recommenders. A recommender is a machine learning algorithm that provides annotators with suggestions for potential annotations. Suggestions are shown next to the annotations, and they can be accepted or rejected. From this, recommenders can (if the algorithm supports it) learn and improve the quality of the suggestions, thus implementing a cycle of manual and machine correction ("human-in-the-loop"). Cur-

rently, INCEpTION ships with several different recommenders and allows developers to add their own. Given that the LAPPS Grid provides a wide range of different NLP tools as services, we have implemented a recommender that calls services available from the LAPPS Grid. Because INCEpTION allows users to create their own annotation layers and features, we need to ensure that recommenders called from the LAPPS Grid can only be configured for layers whose types are compatible. For this, we leverage the metadata information given by the LAPPS Grid service API to prune recommenders that are not compatible.

Note that while INCEpTION can use LAPPS Grid services to generate annotation suggestions for the human to correct, but there is currently no way of feeding the corrections directly back to the services for re-training. However, we intend to configure another trainable recommender within INCEpTION, which learns based on the suggestions from the LAPPS Grid services that the user has explicitly accepted or corrected.

INCEpTION internally uses UIMA CAS as a syntactic format, whereas the LAPPS Grid uses its LIF (JSON-LD) format. When recommending, we therefore use the DKPro Core CAS ↔ LIF converter, which was developed in the recent past to enable incorporation of DKPro Core modules into the LAPPS Grid (and vice versa). The basic type system of DKPro Core and the LAPPS Grid (tokens, sentences, POS, and NER) corresponds one-to-one, allowing a simple conversion.

Discussion

Our analysis reveals that interoperability across platforms intended to facilitate the creation and use of resources and NLP applications for accessing and mining scientific publications is feasible, due to the many commonalities in the representation of linguistically annotated data that have been adopted by frameworks developed over the past decade.

Pairwise interoperability among the three platforms is sufficient to enable the use of specific features by making use of underspecification, limited manual configuration, and/or conventions (best practices). For example, the fact that PubAnnotation is schemaless allows LAPPS and INCEpTION to set up conventions to encode platform-specific metadata using the PubAnnotation model. An example of minimal manual configuration is the interoperability between INCEpTION and LAPPS where the user needs to specify the annotation type and feature carrying the predicted labels. To meet the user's needs, it is sufficient to access these labels, and full mapping of all annotations produced by the LAPPS Grid is not required. Additionally, we have identified several places where, by including more metadata, interoperability among the three platforms can be further improved. For example, PubAnnotation could provide support for storing schema information, thereby removing the need for INCEpTION and the LAPPS Grid to use their own conventions, and thus facilitating conversion of annotations between the platforms.

At the same time, we have identified several area(s) where obstacles to interoperability among the three platforms remain, which are attributable to two major sources:

(1) Differences in the "overall organization" of annotations (as opposed to the structure of the annotation content itself), i.e., the PubAnnotation organization of annotation sets in projects, the LIF organization into views with accompanying metadata, and the INCEpTION organization as a UIMA hierarchy of objects.

(2) Difficulty of mapping annotation names that do not necessarily have a corresponding object or feature in the other scheme(s).

We have suggested a number of ways to address the first issue, primarily by finding ways to retain information specific to one representation in the others so that it can be restored after a round trip between platforms. However, this is only required when the platforms exchanging annotations use the same data–for example, one platform might use annotation identifiers while the other does not. Additional work will be necessary to identify to which extent this is critical in real-world use-cases and how to address it.

The second issue concerns semantic interoperability, which is a pervasive problem for harmonizing linguistic annotations, and one for which no obvious universal solution exists at this time. It is interesting to not that the set of basic annotation objects that are commonly produced and consumed by NLP modules (token, sentence, named entity, dependency, etc.) is consistent among the three platforms, even though there are differences in naming conventions; annotation names can therefore be automatically converted if the correspondences can be pre-determined for these objects. However, a problem arises when names are arbitrary (as in PubAnnotation, where no naming conventions are specified) or where users can add arbitrary names to the provided inventory (as allowed in both LIF and INCEpTION) and no mapping can be pre-determined (Note that it is possible to represent annotations from the other platforms in LIF, regardless of naming conventions; however, many tools in the LAPPS Grid will fail if the WSEV names are not used.). Nonetheless, the overlap in basic annotation names and the structural correspondence of associated information as generic feature-value pairs enables conversion at both the syntactic and semantic levels for a good portion of the kinds of annotations likely to be ported from platform to platform.

We also observe that it is presently not possible to build true cross-platform human-in-the-loop processes. INCEpTION pro-

https://doi.org/10.5808/GI.2019.17.2.e19

poses a REST-like protocol to communicate with re-trainable external recommenders. However, NLP services providers (the LAPPS Grid as well as other NLP service providers) do not presently support the re-training of models or the use of custom models in conjunction with their services. Thus, these aspects need to be explicitly considered by future work on NLP service platforms in order to enable human-in-the-loop scenarios.

The bottom line here is that it should ultimately be possible to use the three platforms together to provide a "meta-platform" that can accommodate sophisticated creation, validation, and sharing of annotations over biomedical publications. For example, documents can be retrieved from PubMed Central via PubAnnotation, automatically annotated using LAPPS Grid services and/or manually annotated/corrected using INCEpTION, and subsequently sent back for publication in PubAnnotation's repository; additionally, INCEpTION can use LAPPS Grid services to generate automatically-generated annotation suggestions, thus facilitating the manual annotation process. The synergy among the three platforms may eventually enable exploiting the strengths of each, which would obviate the need to start from scratch to create a single, monolithic application that can provide their combined functionalities.

Conclusion

The purpose of our analysis is, first, to explore the potential to combine the functionalities of PubAnnotation, the LAPPS Grid, and INCEpTION in order to provide a state-of-the-art means for researchers to access, annotate, and eventually mine scientific publications. The scientific community desperately needs an easy-to-use, powerful platform that enables not only access to publication texts, but also rapid development of annotated corpora to support machine learning. This is especially needed so that text mining systems across disciplines and/or tuned to specific domains and sub-areas can be developed. A second motivation for our analysis is to identify obstacles to cross-platform interoperability for natural language processing in general, which can potentially inform design and implementation choices for future systems. We see the ability to combine the functionalities of existing platforms as an important element of progress in the field, to avoid reinventing the wheel as well as modularize component capabilities, thereby reducing the overhead of maintaining monolithic systems and distributing effort as well as cost.

ORCID

Richard Eckart de Castilho: https://orcid.org/0000-0003-0991-7045
Nancy Ide: https://orcid.org/0000-0001-9473-3544
Jin-Dong Kim: https://orcid.org/0000-0002-8877-3248
Jan-Christoph Klie: https://orcid.org/0000-0003-0181-6450
Keith Suderman: https://orcid.org/0000-0003-0502-4557

Authors' Contribution

Conceptualization: REC, NI, JDK, JCK, KS. Data curation: REC, NI, JDK, JCK, KS. Formal analysis: REC, NI, JDK, JCK, KS. Funding acquisition: REC, NI, JDK. Methodology: REC, NI, JDK, JCK, KS. Writing – original draft: REC, NI, JDK, JCK, KS. Writing – review & editing: REC, NI, JDK, JCK, KS.

Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

The LAPPS Grid is supported by US National Science Foundation grants NSF-ACI 1147944, NSF-ACI 1147912, NSF-EAGER 181123, and the Andrew K. Mellon Foundation; the LAPPS Grid is also supported by the NSF-funded Jetstream cloud computing environment.

INCEpTION is supported by the German Research Foundation under grant No. EC503/1-1 and GU 798/21-1. PubAnnotation is supported by the Database Integration Coordination Program funded by National Bioscience Database Center (NBDC) of Japan Science and Technology Agency (JST).

References

1. Cunningham H, Tablan V, Roberts A, Bontcheva K. Getting more out of biomedical documents with GATE's full lifecycle open source text analytics. PLoS Comput Biol 2013;9:e1002854.
2. Hinrichs M, Zastrow T, Hinrichs EW. WebLicht: web-based LRT services in a distributed eScience infrastructure. In: Proceedings of the 7th International Conference on Language Resources and Evaluation (LERC 2010) (Calzolari N, Choukri K, Maegaard B, Mariani J, Odijk J, Piperidis S, et al., eds.), 2010 May 19-21, Valletta, Malta. Paris: European Language Resources Association, 2010. pp. 489-493.
3. Ide N, Pustejovsky J, Cieri C, Nyberg E, DiPersio D, Shi C, et al. The Language Application Grid. In: Proceedings of the 9th International Conference on Language Resources and Evaluation (LERC 2014) (Calzolari N, Choukri K, Declerck T, Loftsson H, Maegaaard B, Mariani J, et al., eds.), 2014 May, Reykjavik, Iceland. Paris: European Language Resources Association, 2014. pp.

22-30.
4. Labropoulou P, Galanis D, Lempesis A, Greenwood MA, Knoth P, Eckart de Castilho R, et al. OpenMinTeD: a platform facilitating text mining of scholarly content. In: Proceedings of the 11th International Conference on Language Resources and Evaluation (LERC 2018) (Calzolari N, Choukri K, Cieri C, Declerck T, Goggi S, Hasida K, et al., eds.), 2018 May 7-12, Miyazaki, Japan. Paris: European Language Resources Association, 2018.
5. Ferrucci D, Lally A, Verspoor K, Nyberg E. Unstructured information management architecture (UIMA) version 1.0. OASIS Standard, 2009 Mar 2. Burlington: OASIS, 2009.
6. Rak R, Rowley A, Carter J, Ananiadou S. Development and analysis of NLP pipelines in Argo. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations (Butt M, Hussain S, eds.), 2013 Aug 4-9, Sofia, Bulgaria. Stroudsburg: Association for Computational Linguistics, 2013. pp. 115-120.
7. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010;17:507-513.
8. Eckart de Castilho R, Gurevych I. A broad-coverage collection of portable NLP components for building shareable analysis pipelines. In: Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT (Ide N, Grivolla J, eds.), 2014 Aug 23, Dublin, Ireland. Stroudsburg: Association for Computational Linguistics, 2014. pp. 1-11.
9. Furrer L, Jancso A, Colic N, Rinaldi F. OGER++: hybrid multi-type entity recognition. J Cheminform 2019;11:7.
10. Kim JD, Wang Y. PubAnnotation: a persistent and sharable corpus and annotation repository. In: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (BioNLP 2012) (Cohen KB, Demner-Fushman D, Ananiadou S, Webber B, Tsujii J, Pestian J, eds.), 2012 Jun 8, Montreal, Canada. Stroudsburg: Association for Computational Linguistics, 2012. pp. 202-205.
11. Verhagen M, Suderman K, Wang D, Ide N, Shi C, Wright J, et al. The LAPPS interchange format. In: WLSI 2015 Revised Selected Papers of the Second International Workshop on Worldwide Language Service Infrastructure (Murakami Y, Lin D, eds.), 2015 Jan 22-23, Kyoto, Japan. Lecture Notes in Computer Science, Vol. 9442. New York: Springer, 2016. pp. 33-47.
12. Ide N, Suderman K, Verhagen M, Pustejovsky J. The Language Application Grid Web Service Exchange Vocabulary. In: WLSI 2015 Revised Selected Papers of the Second International Workshop on Worldwide Language Service Infrastructure (Murakami Y, Lin D, eds.), 2015 Jan 22-23, Kyoto, Japan. Lecture Notes in Computer Science, Vol. 9442. New York: Springer, 2016. pp. 18-32.
13. Klie JC, Bugert M, Boullosa B, Eckart de Castilho R, Gurevych I. The INCEpTION platform: machine-assisted and knowledge-oriented interactive annotation. In: Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations (Zhao D, ed.), 2018 Aug 23-29, Santa Fe, New Mexico. Stroudsburg: Association for Computational Linguistics, 2018. pp. 5-9.
14. Gotz T, Suhre O. Design and implementation of the UIMA Common Analysis System. IBM Syst J 2004;43:476-489.

https://doi.org/10.5808/GI.2019.17.2.e19

G&I
Genomics & Informatics

Application note

eISSN 2234-0742
Genomics Inform 2019;17(2):e20
https://doi.org/10.5808/GI.2019.17.2.e20

Received: March 14, 2019
Revised: May 31, 2019
Accepted: May 31, 2019

*Corresponding author:
E-mail: arnaud.ferre@inra.fr

© 2019, Korea Genome Organization
This is an open-access article distributed under the terms of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Improving the CONTES method for normalizing biomedical text entities with concepts from an ontology with (almost) no training data

Arnaud Ferré[1,2*], Mouhamadou Ba[1], Robert Bossy[1]

[1]MaIAGE, INRA, Paris-Saclay University, 78350 Jouy-en-Josas, France
[2]LIMSI, CNRS, Paris-Saclay University, 91405 Orsay, France

Entity normalization, or entity linking in the general domain, is an information extraction task that aims to annotate/bind multiple words/expressions in raw text with semantic references, such as concepts of an ontology. An ontology consists minimally of a formally organized vocabulary or hierarchy of terms, which captures knowledge of a domain. Presently, machine-learning methods, often coupled with distributional representations, achieve good performance. However, these require large training datasets, which are not always available, especially for tasks in specialized domains. CONTES (CONcept-TErm System) is a supervised method that addresses entity normalization with ontology concepts using small training datasets. CONTES has some limitations, such as it does not scale well with very large ontologies, it tends to overgeneralize predictions, and it lacks valid representations for the out-of-vocabulary words. Here, we propose to assess different methods to reduce the dimensionality in the representation of the ontology. We also propose to calibrate parameters in order to make the predictions more accurate, and to address the problem of out-of-vocabulary words, with a specific method.

Keywords: biomedical text mining, entity normalization, ontology, word embedding
Availability: The tool is on GitHub, https://github.com/ArnaudFerre/CONTES; and the corpus and the ontology used are available at BioNLP Shared-Task 2016, https://sites.google.com/site/bionlpst2016/tasks/bb2; pre-calculated embeddings are available on demand.

Introduction

The world of science produces a wealth of knowledge in the form of scientific literature, a volume is so large that it is difficult or even impossible for scientists to keep abreast of new developments [1]. One solution to deal with this problem is a natural language processing approach known as information extraction [2]. Information extraction is the process of finding statements on a specific subject in texts [3]; for example, finding statements about a protein binding to another protein, or medication to treat a disease.

Normalization is the grounding of an entity mentioned in the text to an identifier in a data- or knowledge-base [4]. This is a core task of information extraction, which can be understood as a classification problem: classify entity mentions into correct class(es) from a reference vocabulary. Current normalization methods try to address two main problems: (1) the weak generalizability to different domains/tasks and (2) the high morphological variability in entity mentions and labels of classes.

Supervised approaches, combined with distributional semantics, can bring some answers.

In the general domain, where the amount of training data is high, such methods already yield good results [5]. But in specialized domains, such as biomedicine, where there is often a lack of training data and/or a high number of classes, these state of the art methods seem to have difficulties to keep a high performance [6].

Recently, the CONTES (CONcept-TErm System) method [7], a supervised method, tries to address this problem by integrating more knowledge from domain ontologies. Ontologies being sometimes used as reference for normalization tasks, their hypothesis is that deeper integration could improve the quality of the predictions. The method outperforms existing approaches on the Bacteria Biotope (BB) normalization task of the BioNLP Shared Task 2016 [8]. The CONTES method, however, has two main limitations. The first is the size of the target ontology, i.e., large sets of concepts prove difficult to handle because of their high dimensionality. The second is overgeneralization: CONTES tends to "play it safe," and predict more general concepts. Another possible limitation comes from the impossibility of the method to represent words unseen during the training step; that is, an out-of-vocabulary word problem.

Here, we propose to experiment different methods to reduce the dimensionality of the representation of the ontology structure, and to calibrate parameters to make less general predictions. We also propose to integrate another method to produce word embeddings, which addresses the problem of out-of-vocabulary words.

Description of the CONTES Method

CONTES [7] implements a supervised machine learning approach which aims to improve the generalization limitations of previous machine learning methods, especially when few training datasets are available, with respect to the number of classes. CONTES is an open-source tool written in Python (https://github.com/Arnaud-Ferre/CONTES). The core of the method relies on two different semantic vector spaces: (1) semantic vector representations of expressions from a corpus (SSC for Semantic Space of Corpus), (2) semantic vector representations of concepts from an ontology (SSO for Semantic Space of Ontology).

CONTES can use all kinds of semantic vector representations for SSC computations, such as distributional representations. It particularly uses Word2Vec word embeddings [9]. For SSO computation, it automatically builds concept vectors by analyzing subsumption relationships between all the concepts of an ontology. From the vector representations, it learns a projection of the SSC into the SSO by globally maximizing the cosine similarity between the expression vectors and the associated concept vectors, in the training set. In an unsupervised setting, instead of using (expression, concept) associations, obtained from an annotated corpus for training, CONTES can use (label of concept, concept) intrinsic associations. CONTES showed great performance on the BB normalization task of BioNLP Shared Task 2016 [7]. Analysis of the results suggested that even mispredicted concepts were not far from the ontology of the expected concepts.

Limitations

Dimension reduction and concept embeddings

The CONTES method has been evaluated as a normalization task of BB from the BioNLP Shared Task of 2016 (http://2016.bionlp-st.org/tasks/bb2). This task consists of normalizing entity mentions, in the raw text, with concepts of the OntoBiotope ontology. OntoBiotope is an ontology of microorganism habitats, and captures classifications used by biologists to describe microorganism isolation sites (e.g., GenBank, GOLD, ATCC), and contains around 2,000 concepts, which is relatively small, compared to other biomedical ontologies, but relatively high, for a machine-learning classification problem. In comparison, the Gene Ontology contains around 50,000 concepts and the Unified Medical Language System meta-thesaurus contains more than one million concepts. CONTES does not currently address normalization with these large ontologies, due to its high dimension, but rather implies the representation of the ontology concepts.

Overgeneralization

CONTES takes advantage of the ontology structure given by a set of *Is-A* relations between concepts. The representation of the ontology structure into SSO vectors allows CONTES to generalize training examples, and classify terms never seen before with more general concepts. However, error analysis on the BB development set showed that CONTES tends to overgeneralize its predictions, and often predicts too general concepts.

Out-of-vocabulary words

CONTES uses the Word2Vec-CBOW and Word2Vec-SkipGram methods to produce word embeddings, and these methods do not compute valid representations for out-of-vocabulary words. Thus, CONTES does not yet address the problem of the out-of-vocabulary words. The lack of valid representations for some words is likely to bias the performance of the method. For example, in French or Spanish, most verbs have more than forty different inflected forms, while the Finnish language has fifteen cases for nouns. These languages contain many word forms that occur rarely (or not at all) in the training corpus, making it difficult to learn good word representations.

https://doi.org/10.5808/GI.2019.17.2.e20

Experiments

Materials

For our experiments, we use the current source code of CONTES (https://github.com/ArnaudFerre/CONTES), and datasets and the OntoBiotope ontology from the BB normalization task of the BioNLP Shared Task 2016. We also downscaled word embeddings computed from a reduced corpus (40,000 words), in order to speed up the experimental cycle.

Dimensionality-reduction techniques

We aimed to reduce the dimension of the SSO to make CONTES scalable to larger ontologies. We have used three well-known techniques to reduce the dimension of the semantic space of the ontology: principal component analysis (PCA), multidimensional scaling (MDS), and t-distributed Stochastic Neighbor Embedding (t-SNE). PCA and MDS have already been tested, with similar results [7]. Here, t-SNE is a nonlinear method to reduce dimensions, with the aim to retain the similarity of vectors between the initial space and the embedded space. All these methods can be used with the machine learning Python library scikit-learn (https://scikit-learn.org/stable/).

Concept embeddings

Another solution to overcome the problem of high-dimensional vectors of ontology concepts, in CONTES, is to directly learn embeddings for each concept. With the success of the word-embedding methods, we considered graph-based embeddings, and used Node2Vec [10] to compute the concept vectors (SSO) of ontologies for CONTES.

Node2Vec is a method to learn embeddings for nodes in networks. The approach relies on a maximization of the likelihood of preserving neighborhoods of nodes in networks. It considers the hypothesis that highly connected nodes of the same community have similar embeddings, and that nodes having similar structural roles in networks, have similar embeddings [10].

For CONTES, Node2Vec has the advantage of being applicable to ontologies, and it also directly provides low-dimensional vectors. Although nodes in ontologies are not generally highly connected, as can be the social network graphs for example, we expect that Node2Vec can also be suitable for ontologies by providing low-dimensional vectors for ontology concepts, based on the neighborhoods expressed in relationships such as "*Is-A*" and "*sameAs*."

The experiments with Node2vec allowed us to obtain compact vector representations for the Ontobiotope concepts, which opens the possibility for larger ontologies.

Overgeneralization

Our hypothesis was that CONTES overgeneralization is caused by giving too much weight to the individual concept ancestors, in the vectors representing the ontology structure. CONTES uses the "Ancestry" method where each concept is represented as a vector having a number of dimensions equal to the total number of concepts in the ontology. For a given concept, the dimension corresponding to the concept itself or to one of its ancestors is set to one (1). All other dimensions are set to zero (0). The representation has the advantage of preserving the required information to rebuild the ontology structure. However, all ancestors have the same weight regardless of the number of intermediate concepts.

We experimented with a method where the dimension value is decreased using a function that depends on (1) the depth between the concept and the ancestor, and (2) a constant parameter used to set a "decay" factor. The function is w^d, where w is the decay factor whose range is $[0,1]$, and d is the number of intermediate concepts between the target concept and its ancestor (zero for itself, one for its direct parent).

Out-of-vocabulary words

New words currently required to recalculate the word embeddings in CONTES. Methods like FastText are designed to obtain embeddings for words even if they are unknown from the training data. For our experiments, we considered FastText [11] as a solution to tackle the problem of out-of-vocabulary words.

The FastText method is based on the skip-gram model [9,11], and takes into account subword information to produce word embeddings. It also considers each word as a bag of character n-grams and gives the word representation by summing the representations of the character n-grams [11].

To investigate the effects of unknown words, FastText is added to CONTES as another method, and we have conducted experiments with a subset of the datasets of the BB normalization task.

Results

Dimension reduction

As shown in [7], the PCA and MDS methods had similar behavior, and they both reduced the performance of the CONTES method (Fig. 1). The score is based on the semantic similarity proposed by Wang et al. [12]. The gap is small when the reduction is low, but increases beyond the acceptable performance threshold when it reaches 60% reduction.

The sklearn t-SNE method is not scalable: increasing the reduced space dimension number seems to prevent any execution of reduction. The sklearn description of the method mentions that "It is

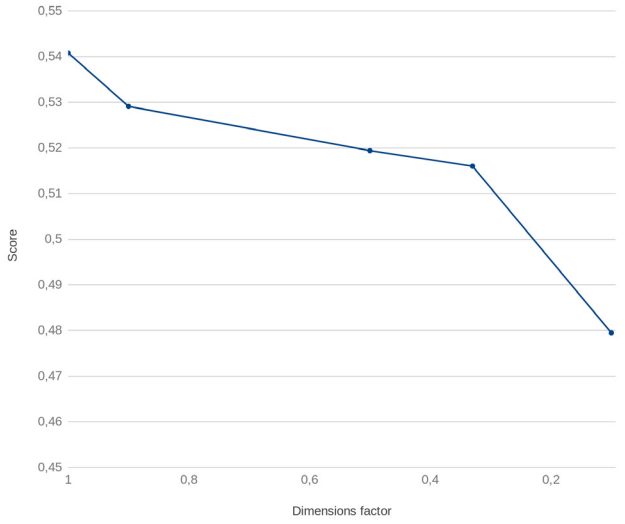

Fig. 1. Performance of the method, according to the dimension reduction.

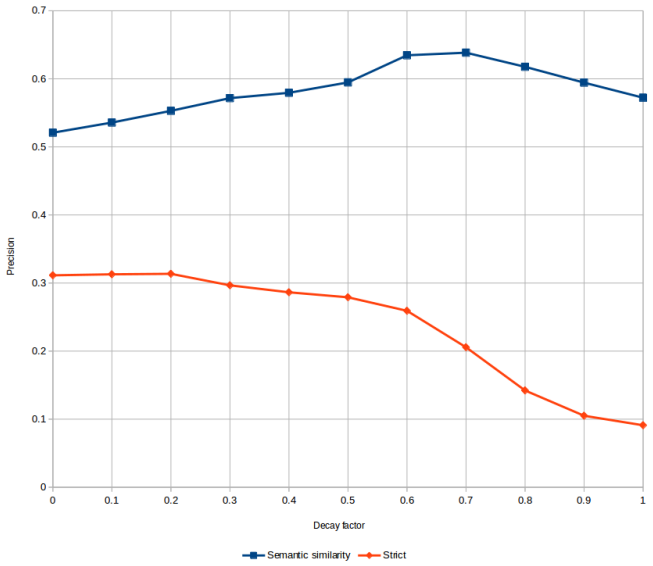

Fig. 2. Results with semantic similarity and strict measures.

highly recommended to use another dimensionality reduction method [...] to reduce the number of dimensions to a reasonable amount (e.g., 50), if the number of features is very high" (https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html). The initial goal of the method is indeed to enable a 2D/3D representation of a high-dimensional space.

Overgeneralization

We experimented the aforementioned function for computing SSO by varying the decay factor from zero to one, by steps of 0.1. With a decay factor of one, the function produces the same vectors as the original "Ancestry" method. With a decay factor of zero, the function reproduces a "One Hot" representation.

We measured the impact of the decay factor on the BBs development set, using two metrics. The first metric is the official BBs precision measure that is a mean of the semantic similarity between predicted and reference normalizations, and the semantic similarity rewards predictions that are near the reference. The second metric is a strict measure, where the prediction is only rewarded if it is exactly the same as the reference.

The results are shown in Fig. 2. The strict measure is necessarily lower, since it does not reward generalization or near-misses. With the semantic similarity measure, we reproduced the results published in Ferre et al.'s study [13] where "Ancestry" (decay = 1) outperforms "One Hot" (decay = 0). We also demonstrated an optimal decay factor between 0.6 and 0.7 that further improves CONTES performance. With the strict measure, we notice that "One Hot" is the best representation.

Out-of-vocabulary words

Concerning out-of-vocabulary words, our experiments showed no significant difference in performance between FastText and Word-2vec.

Perspectives and conclusion

In this work, we addressed some important issues to normalize biomedical text entities using concepts of ontologies, in the context of small training data. Reduction of dimensions, in the representation of the concepts, appears to be a central issue when trying to support large ontologies. We also tested PCA and MDS methods, and showed that the gap in performance between these methods, and the original CONTES, rapidly widens when the reduction gets high. We also introduced concept embeddings to tackle the dimensional reduction, and we plan to perform follow-up experiments with Node2Vec by more generally considering the other relationships between the concepts in the ontologies. New and large ontologies will be addressed in future work.

By tackling overgeneralization, our results showed a way to enhance CONTES with a function weighing the ontology concepts, based on the depths between the concepts and their ancestors. Out-of-vocabulary words is addressed with FastText, but it does not show significant differences, compared to results without FastText, but we expect that results may change with larger datasets.

The experiments were conducted under specific conditions with reduced datasets, and future work will consider that point.

https://doi.org/10.5808/GI.2019.17.2.e20

ORCID

Arnaud Ferré: https://orcid.org/0000-0001-9115-8222
Mouhamadou Ba: https://orcid.org/0000-0001-7233-7831
Robert Bossy: https://orcid.org/0000-0001-6652-9319

Authors' Contribution

Conceptualization: AF. Data curation: RB, MB. Formal analysis: RB, MB, AF. Funding acquisition: RB, AF. Methodology: RB, MB, AF. Writing – original draft: AF, MB, RB. Writing – review & editing: MB, AF, RB.

Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

This work was supported by the "IDI 2015" project funded by the IDEX Paris-Saclay, ANR-11-IDEX-0003-02. The work at BLAH5 was supported by Database Center for Life Science (DBCLS).

References

1. Cohen KB, Demner-Fushman D. Biomedical Natural Language Processing. Amsterdam: John Benjamins Publishing Company, 2014.
2. Craven M, Kumlien J. Constructing biological knowledge bases by extracting information from text sources. Proc Int Conf Intell Syst Mol Biol 1999:77-86.
3. Jurafsky D, Martin JH. Speech and Language Processing. Upper Saddle River: Prentice-Hall Inc., 2014.
4. Morgan AA, Lu Z, Wang X, Cohen AM, Fluck J, Ruch P, et al. Overview of BioCreative II gene normalization. Genome Biol 2008;9 Suppl 2:S3.
5. Sil A, Kundu G, Florian R, Hamza W. Neural cross-lingual entity linking. Edinburgh: University of Edinburgh, 2018. Accessed 2019 May 1. Available from: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/viewPaper/16501.
6. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015; 521:436-444.
7. Ferre A, Zweigenbaum P, Nedellec C. Representation of complex terms in a vector space structured by an ontology for a normalization task. In: Proceedings of the BioNLP 2017 (Cohen KB, Demner-Fushman D, Ananiadou S, Tsukii J, eds.), 2017 Aug 4, Vancouver, Canada. Stroudsburg: Association for Computational Linguistics, 2017. pp. 99-106.
8. Deleger L, Bossy R, Chaix E, Ba M, Ferre A, Bessieres P, et al. Overview of the bacteria biotope task at BioNLP Shared Task 2016. In: Proceedings of the 4th BioNLP Shared Task Workshop (Nedellec C, Bossy R, Kim JD, eds.), 2016 Aug 13, Berlin, Germany. Stroudsburg: Association for Computational Linguistics, 2016. pp. 12-22.
9. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. Ithaca: arXiv, Cornell University, 2013. Accessed 2019 May 1. Available from: https://arxiv.org/abs/1301.3781.
10. Grover A, Leskovec J. node2vec: scalable feature learning for networks. In: KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016 Aug 13-17, San Francisco, CA, USA. New York: Association for Computing Machinery, 2016. pp. 855-864.
11. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. Trans Assoc Comput Linguist 2017;5:135-146.
12. Wang JZ, Du Z, Payattakool R, Yu PS, Chen CF. A new method to measure the semantic similarity of GO terms. Bioinformatics 2007;23:1274-1281.
13. Ferre A, Deleger L, Zweigenbaum P, Nedellec C. Combining rule-based and embedding-based approaches to normalize textual entities with an ontology. Miyazaki: European Languages Resources Association, 2018. Accessed 2019 May 1. Available from: http://www.aclweb.org/anthology/L18-1543.

G&I
Genomics & Informatics

Application note

eISSN 2234-0742
Genomics Inform 2019;17(2):e21
https://doi.org/10.5808/GI.2019.17.2.e21

Received: February 23, 2019
Revised: May 31, 2019
Accepted: May 31, 2019

*Corresponding author:
E-mail: colic@ifi.uzh.ch

© 2019, Korea Genome Organization
This is an open-access article distributed under the terms of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Improving spaCy dependency annotation and PoS tagging web service using independent NER services

Nico Colic[1*], Fabio Rinaldi[1,2,3]

[1]Institute of Computational Linguistics, University of Zurich, CH-8050 Zurich, Switzerland
[2]IDSIA, CH-6928 Manno, Switzerland
[3]Swiss Institute of Bioinformatics, Quartier Sorge-Bâtiment Amphipôle, CH-1015 Lausanne, Switzerland

Dependency parsing is often used as a component in many text analysis pipelines. However, performance, especially in specialized domains, suffers from the presence of complex terminology. Our hypothesis is that including named entity annotations can improve the speed and quality of dependency parses. As part of BLAH5, we built a web service delivering improved dependency parses by taking into account named entity annotations obtained by third party services. Our evaluation shows improved results and better speed.

Keywords: dependency parsing, named entity recognition, natural language processing
Availability: The code for our web service is publicly available (https://github.com/OntoGene/blah5).

Introduction

Dependency parsing might be used as a component for tackling many text mining problems. It is often used as a feature to train machine learning algorithms or used for rule-based approaches for relation extraction [1,2]; and can be used to train better word embeddings [3]. However, as continued research and community efforts in the form of the CoNLL shared tasks, for example, show, dependency parsing in general is still not a problem solved completely [4]. Making a fast, reliable dependency parser readily available for the wider research community for further processing to build upon will help spur efforts in event and relation extraction.

As previous research has shown, providing named entity information to the dependency parser can improve the accuracy of the parses [5,6]. The reasoning is that dedicated named entity recognition (NER) tools perform much better in their specific domain, and by extracting named entities with higher accuracy will facilitate appropriate parsing tools that are not trained on biomedical data.

spaCy is a open-source natural language processing (NLP) library written in Python that performs tokenization, Part-of-Speech (PoS) tagging and dependency parsing. It is the fastest NLP parser available, and offers state-of-the-art accuracy [2,7].

Services such as PubDictionaries and OGER perform dictionary-based entity look up [8]. Other state-of-the-art taggers are the Jensen tagger [9] and TaggerOne [10].

In this work, we present our contribution to BLAH5. We proposed to build a public web service that can be used by researchers and other services, such as PubAnnotation, to obtain improved dependency annotations, based upon integration with a NER service (Fig. 1). We report on its efficacy here.

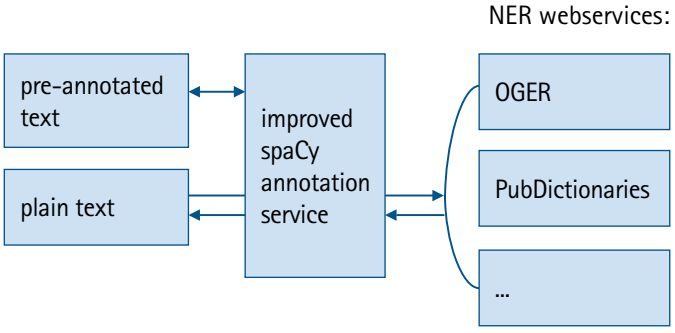

Fig. 1. Architecture of our web service.

```
{"denotations": [
    { "span": { "begin": 11, "end": 14 },
    "obj": "DO:005313",
    "id": "T1" },
    { "span": { "begin": 30, "end": 53 },
    "obj" : "DO:009456",
    "id": "T2" }],
         "text": "Impact of ADHD in adults with
neurofibromatosis type 1" }
```

Fig. 2. Example of input to our web service, in json format.

Related Work

The most recent extensive evaluation of existing dependency parsers has been performed by Choi et al. [11]. They evaluate 10 different off-the-shelf parsers for accuracy and speed; reporting labeled attachment scores (LAS) of 85% to 90%. While spaCy does not perform the most accurate in their evaluation, it performs fastest maintaining comparable accuracy. Similar scores are reported in the 2015 SemEval task [12].

Pletscher-Frankild and Jensen [9] note that in text mining literature, tools are often evaluated for accuracy only, omitting practical considerations such as speed and robustness.

Since the publication of the previously mentioned evaluation of dependency parsers [11], new machine learning-based parsers have been published, surpassing existing approaches according to the respective authors' own evaluation [13,14]. Amongs them, spaCy was updated to employ a neural network to improve performance [7]. As Yang et al. [15] show, however, domain adaptation remains an issue.

Methods

Our web service is implemented using Flask (http://flask.pocoo.org/), a Python library that facilitates the creating of web interfaces, and the aforementioned spaCy (https://spacy.io). spaCy's structure is modular, and processing text happens by calling the respective modules. For our approach, we added a further processing step between "tokenization" and "dependency parsing," during which we will recompute token segmentation taking into account named entities provided. Then we pass the adjusted token offsets to the dependency parsing module.

Input format

Input is provided either as plain text or as a JSON string. The latter may either contain just a text field, or additionally annotations in PubAnnotation's JSON format (http://www.pubannotation.org/docs/annotation-format/) (Fig. 2).

The input is then internally converted into a spaCy object. From there, spaCy is used to perform tokenization, PoS tagging and dependency parsing, taking into account existing annotations if available.

This is done by tokenizing the text provided, and iterating through the given annotations to form a new sequence of tokens. This new sequence keeps the NER annotations as individual tokens and does not break them down further. In case of conflicting and overlapping annotations, we use the leftmost longest match.

Using this new token sequence, a new spaCy representation object is created on which dependency parsing is performed.

If no annotations are provided manually, our web service will automatically call another service to automatically obtain named entity annotations. While this could be any service, it is currently set to using OGER.

Output format

The resulting object is then converted back into JSON, and token offsets are computed. The output format is PubAnnotation JSON, and can directly be uploaded to PubAnnotation. This also allows to make use of PubAnnotation's annotation visualizer, *TextAE*, as seen in Fig. 3.

Usage

The code for our web service is publicly available (https://github.com/OntoGene/blah5), and installation is straight-forward. For demonstration purposes, we are running an instance of the web service.

Improved dependency annotations can be obtained using curl or wget in the fashion shown in Fig. 4.

https://doi.org/10.5808/GI.2019.17.2.e21

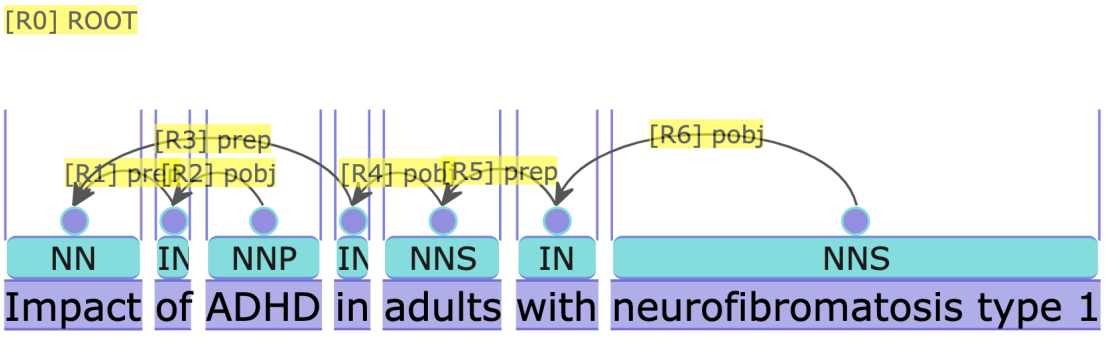

Fig. 3. Visualizing the web service's output in TextAE.

```
curl -F file='@data/sample.json' https://pub.
cl.uzh.ch/projects/ontogene/blah5/ > out.json
```

Fig. 4. Example of complete query to the dependency annotation server: sample.json is the input file, out.json the output file, both in json format.

Adaptation

The application can be run locally, provided the necessary Python libraries are available. To add other independent web services to provide NER information, refer to the documentation of the code.

Results

We evaluated our service in two regards: speed and accuracy.

Speed

In order to evaluate speed of our web service, we created a set of 10,000 randomly selected PubMed abstracts, (1) which we sent to the web service directly to obtain "default" dependency parses, and (2) which we first tagged for named entities using OGER, and then sent to the web service to obtain "improved" parses. Table 1 lists the time it took our tool to complete the dependency parsing step for both cases.

We attribute this decrease in processing time to the diminished complexity of computing the parses.

In a practical setting, however, our web service needs to recompute token offsets in order to correctly integrate NERs in the token stream.

While spaCy in the "default" case does this in a internally optimized fashion, our service currently is not optimized to the same degree, which in practice results in overall lower processing times. The integration of NERs still produces an advantage, as the generated parses avoid some errors that otherwise would have been generated (see below). In practice, when using the web service over the network, total processing time will be around 0.3 s per abstract.

Table 1. Time required to process a batch of 10,000 PubMed abstracts, in the default case (without prior named entity recognition) and with added named entity recognition

	Default	*Improved*
Total time (s)	1018	931
Per abstract (s)	0.10	0.09

Accuracy

We expect our service to perform well in situations where multiword named entities cause problems for the parser. Below we demonstrate two examples in which our service performs well in such situations.

As Figs. 5 and 6 show, complex named entities are problematic for the dependency parser and tagger, causing it to tag "kinase" as a verb; when in fact it is part of the compound entity "Testis-specific serine/threonine protein kinase 4 (Tssk4)." This results in an incorrect parse tree.

The second example of Figs. 7 and 8 shows the case where the parse is not technically wrong, but convoluted and useless to downstream tasks.

Discussion

The examples shown above point to the efficacy of our web service. However, a more thorough evaluation, while useful, is difficult. This is mainly due to the fact that manual creation of parsing annotations is prohibitively expensive, and the fact that existing annotated corpora are annotated in various incompatible formats. For example, depending on the annotation schema, conjunctions can be represented

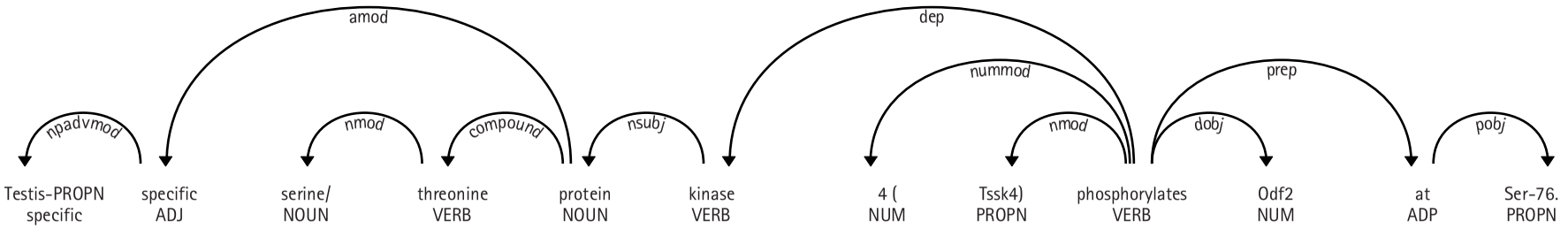

Fig. 5. Default dependency parse for sentence Testis-specific serine/threonine protein kinase 4 (Tssk4) phosphorylates Odf2 at Ser-76.

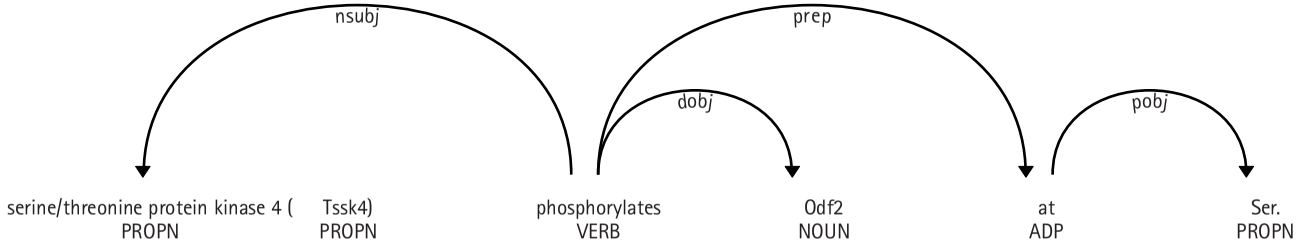

Fig. 6. Improved dependency parse.

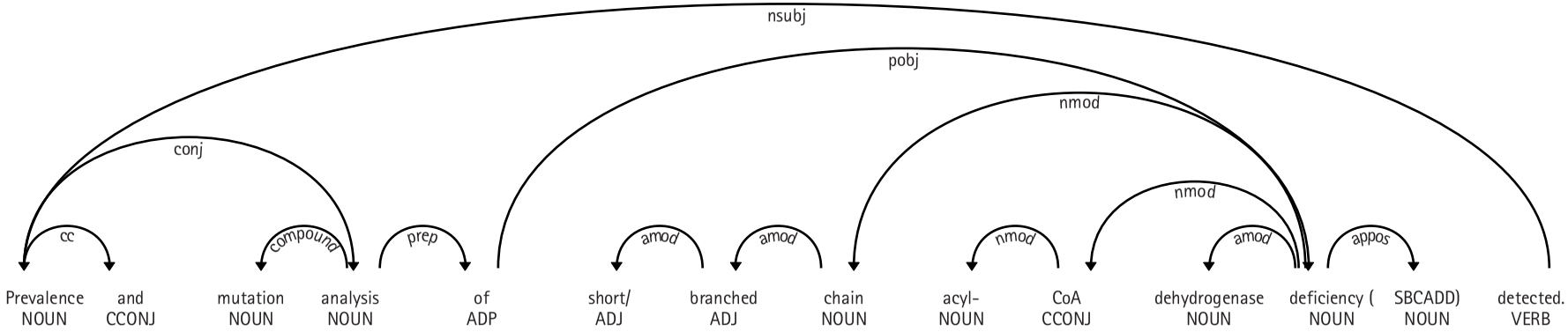

Fig. 7. Default dependency parse for the sentence "Prevalence and mutation analysis of short/branched chain acyl-CoA dehydrogenase deficiency (SBCADD) detected."

https://doi.org/10.5808/GI.2019.17.2.e21

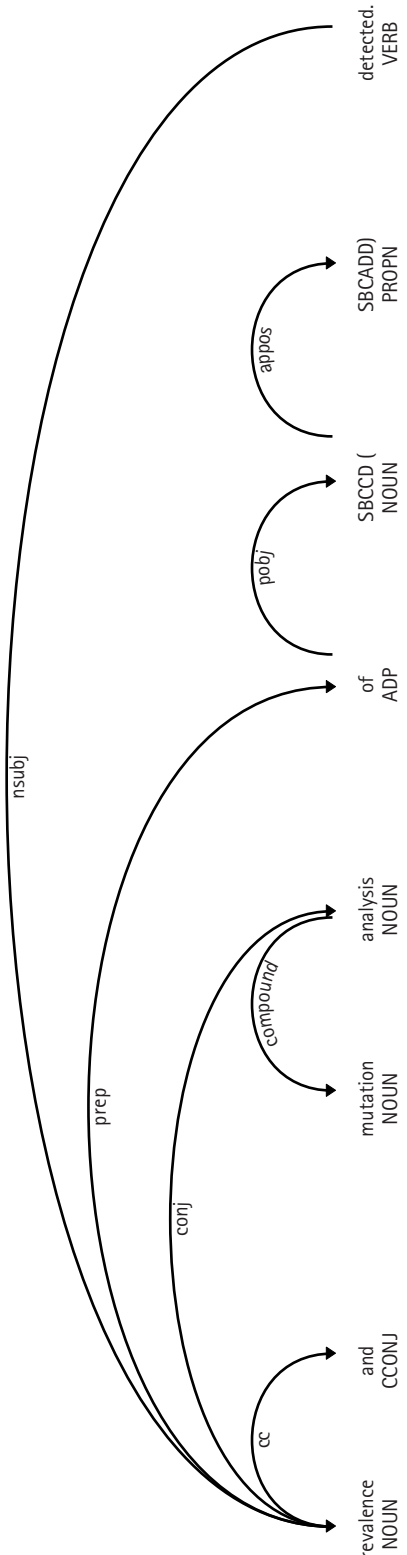

Fig. 8. Improved dependency parse ("short/branched chain acyl-CoA dehydrogenase deficiency" abbreviated for display).

in various ways, or relationships may be either undirected or have opposite directionality. Even though there are efforts to homogenize annotation schemes, such as Universal Dependencies [16], they are not yet widely adopted. This makes automatic comparison of results very difficult. However, since dependency parsing is not usually a goal in itself, but a component used for downstream tasks, an extrinsic evaluation could be carried out on such a downstream task such as relation extraction.

Conclusion

We have presented an improved service for dependency parses, and show that it can deliver better parses for phrases containing complex tokens.

ORCID

Nico Colic: https://orcid.org/0000-0003-1071-5571
Fabio Rinaldi: https://orcid.org/0000-0001-5718-5462

Authors' Contribution

Conceptualization: NC, FR. Data curation: NC, FR. Formal analysis: NC, FR. Funding acquisition: NC, FR. Methodology: NC, FR. Writing – original draft: NC, FR. Writing – review & editing: NC, FR.

Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

The authors express their gratitude to DBCLS for funding participation at BLAH5.

References

1. Nguyen TH, Grishman R. Relation extraction: perspective from convolutional neural networks. In: 2015 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT 2015), 2015 May 31-Jun 5, Denver, CO, USA. Stroudsburg: Association for Computational Linguistics, 2015. pp. 39-48.
2. Colic N. Dependency parsing for relation extraction in biomedical literature [thesis]. Zurich: University of Zurich, 2016.
3. Levy O, Goldberg Y. Dependency-based word embeddings. In:

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers) (Toutanova K, Wu H, eds.), 2014 Jun 23-25, Baltimore, MD, USA. Stroudsburg: Association for Computational Linguistics, 2014. pp. 302-308.

4. Zeman D, Hajic J, Popel M, Potthast M, Straka M, Ginter F, et al. CoNLL 2018 shared task: multilingual parsing from raw text to universal dependencies. In: Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies (Zeman D, Hajic J, Popel M, Straka M, Nivre J, Ginter F, et al., eds.), 2018 Oct 31-Nov 1, Brussels, Belgium. Stroudsburg: Association for Computational Linguistics, 2018. pp. 1-21.
5. Rinaldi F, Dowdall J, Hess M, Kaljurand K, Koitand M, Kahusk N, et al. Terminology as knowledge in answer extraction. In: Proceedings of the 6th International Conference on Terminology and Knowledge Engineering (TKE 2002) (Melby A, ed.), 2002 Aug 28-30, Nancy, France. Le Chesnay: INRIA, 2002. pp. 107-113.
6. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural architectures for named entity recognition. Ithaca: arXiv, Cornell University, 2016. Accessed 2019 Apr 2. Available from: https://arxiv.org/abs/1603.01360.
7. Honnibal M, Johnson M. An improved non-monotonic transition system for dependency parsing. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015 Sep 17-21, Lisbon, Portugal. Stroudsburg: Association for Computational Linguistics, 2015. pp. 1373-1378.
8. Furrer L, Jancso A, Colic N, Rinaldi F. OGER++: hybrid multi-type entity recognition. J Cheminform 2019;11:7.
9. Pletscher-Frankild S, Jensen LJ. Design, implementation, and operation of a rapid, robust named entity recognition web service. J Cheminform 2019;11:19.
10. Leaman R, Lu Z. TaggerOne: joint named entity recognition and normalization with semi-Markov Models. Bioinformatics 2016;32:2839-2846.
11. Choi JD, Tetreault J, Stent A. It depends: dependency parser comparison using a web-based evaluation tool. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Zong C, Strube M, eds.), 2015 Jul 26-31, Beijing, China. Stroudsburg: Association for Computational Linguistics, 2015. pp. 387-396.
12. Oepen S, Kuhlmann M, Miyao Y, Zeman D, Cinkova S, Flickinger D, et al. SemEval 2015 task 18: broad-coverage semantic dependency parsing. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015) (Nakov P, Zesch T, Cer D, Jurgens D, eds.), 2015 Jun, Denver, CO, USA. Stroudsburg: Association for Computational Linguistics, 2015. pp. 915-926.
13. Dozat T, Manning CD. Deep biaffine attention for neural dependency parsing. Ithaca: arXiv, Cornell University, 2016. Accessed 2019 Apr 2. Available from: https://arxiv.org/abs/1611.01734.
14. Kiperwasser E, Goldberg Y. Simple and accurate dependency parsing using bidirectional LSTM feature representations. Trans Assoc Comput Linguist 2016;4:313-327.
15. Yang H, Zhuang T, Zong C. Domain adaptation for syntactic and semantic dependency parsing using deep belief networks. Trans Assoc Comput Linguist 2015;3:271-282.
16. Nivre J, de Marneffe MC, Ginter F, Goldberg Y, Hajic J, Manning CD, et al. Universal dependencies v1: a multilingual treebank collection. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016) (Calzolari N, Choukri K, Declerck T, Goggi S, Grobelnik M, Maegaard B, et al., eds.), 2016 May 23-28, Portoroz, Slovenia. Paris: European Language Resources Association, 2016. pp. 1659-1666.

https://doi.org/10.5808/GI.2019.17.2.e21

G&I
Genomics & Informatics

Application note

eISSN 2234-0742
Genomics Inform 2019;17(2):e17
https://doi.org/10.5808/GI.2019.17.2.e17

Received: December 14, 2018
Revised: May 30, 2019
Accepted: May 30, 2019

*Corresponding author:
E-mail: pierre.larmande@ird.fr

© 2019, Korea Genome Organization
This is an open-access article distributed under the terms of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

OryzaGP: rice gene and protein dataset for named-entity recognition

Pierre Larmande[1,2*], Huy Do[2], Yue Wang[3]

[1]UMR DIADE, Institute of Research for Sustainable Development (IRD), F-34394 Montpellier, France
[2]ICT Lab, University of Science and Technology of Hanoi (USTH), 100000 Hanoi, Vietnam
[3]Database Center for Life Science (DBCLS), Chiba 277-0871, Japan

Text mining has become an important research method in biology, with its original purpose to extract biological entities, such as genes, proteins and phenotypic traits, to extend knowledge from scientific papers. However, few thorough studies on text mining and application development, for plant molecular biology data, have been performed, especially for rice, resulting in a lack of datasets available to solve named-entity recognition tasks for this species. Since there are rare benchmarks available for rice, we faced various difficulties in exploiting advanced machine learning methods for accurate analysis of the rice literature. To evaluate several approaches to automatically extract information from gene/protein entities, we built a new dataset for rice as a benchmark. This dataset is composed of a set of titles and abstracts, extracted from scientific papers focusing on the rice species, and is downloaded from PubMed. During the 5th Biomedical Linked Annotation Hackathon, a portion of the dataset was uploaded to PubAnnotation for sharing. Our ultimate goal is to offer a shared task of rice gene/protein name recognition through the BioNLP Open Shared Tasks framework using the dataset, to facilitate an open comparison and evaluation of different approaches to the task.

Keywords: named-entity recognition, natural language processing, Oryza sativa, plant molecular biology, rice, text mining
Availability: A part of the OryzaGP dataset is publicly available through PubAnnotation (http://pubannotation.org/projects/OryzaGP). The full dataset will become available soon after a portion to be a hidden test data set is determined.

Introduction

The last few decades have witnessed a massive explosion of information in the life sciences. However, an important proportion of this information, relevant to this field, is not available from databases, but is instead present in unstructured scientific documents, such as journal articles, reviews, abstracts, and reports. Agronomy is an overarching field that is comprised of diverse domains such as genetics, plant molecular biology, ecology and soil science [1]. Despite advancements in information technology, scientific advancements in agronomy are still commonly based on text. To effectively develop applications to improve crop production through sustainable methods, however, it is important to overlap research findings from these various subdomains, as they are highly interconnected. However, the collection of content is growing continuously, and the information currently available is unstructured text. Using these resources more efficiently, and taking advantage of associated cross-disciplinary research opportunities, poses a major challenge to both biologists and information technologists. One important subtask of information extraction is to

identify biological entities, and their classifications, an endeavor known as named-entity recognition (NER).

Identifying biological entities, from text, is not trivial. Despite the existence of many available approaches to handle this problem in general, and in biomedical domains in particular, few comprehensive studies have been implemented for plants, especially rice. Moreover, we found that rare benchmarks are available for many plant species, but none for rice. Thus, taken together, we faced various difficulties to exploit advanced machine learning methods, for the accurate analysis of rice.

Objective

On the large scale, we are currently building a Resource Description Framework (RDF) knowledge base termed Agronomics-Linked Data (AgroLD [2], http://www.agrold.org). This knowledge base is designed to integrate data from various public, plant-centric databases such as Gramene [3], Oryzabase [4], and TAIR [5], to name a few. The aim of the AgroLD project is to provide an integrated portal for both bioinformatics and domain experts, to exploit a homogenized data model for filling knowledge gaps. Using this landscape, we aim to extract relevant information from the literature, to enrich the content of integrated datasets.

Due to the scope of the project, we exploited information from the Oryzabase database to build a dataset aimed to recognize named text entities such as rice genes and proteins. Our main purpose was to solve NER of rice biological entities, to find the best approach. By sharing this dataset on the PubAnnotation platform and be available at the *BioNLP Open Shared Tasks* (BioNLP-OST, https://2019.bionlp-ost.org), we invited participants to implement their own methods to solve NER tasks for this dataset. Furthermore, to evaluate the performances, we compared their approaches, implemented during the task, with our method [6], implemented before the hackathon.

Contribution

In this project, we used data from Oryzabase (http://pubannotation.org/projects/OryzaGP), a rice comprehensive database for Oryza sativa species published online since 2,000 by Japanese researchers. The latest version of Oryzabase contains 21,739 of rice genes, collected from 44,837 distinct scientific articles. Consequently, we used this information to create the basis of the OryzaGP dataset. Then we used PubMed as a resource to collect the raw data that was later preprocessed to compose the dataset, and developed a custom script implementing the BioPython library to query and retrieve the specific abstracts from PubMed. However, a number of scientific articles were not available in the PubMed database, due to some historical issues and lack of published resources. Due to the limited access of some resources, 10,400 articles were processed after filtering. The detailed raw data is shown in Table 1.

By focusing on the entities of the rice genome, we used the Oryzabase gene list as the ground truth to build up our dataset by keyword matching terms. The first step to preprocessing the data was filtering to remove special characters from the raw data. In fact, due to the number of articles used, the time range of articles was also wide, in that several articles were published in previous decades. To handle the problems of OCR-errors (which appear in the scanned text), we manually removed all the null and nonsense characters in utf-16 in the raw text. All the work was processed by our scripts and then after first step preprocessing. Moreover, we added part-of-speech (POS) tags for each word, to define its type with the aim, to ensure the accuracy of the identification of entities. The POS tagging process is supported by the Natural Language ToolKit (NLTK). To tokenize data, each word was considered a token, giv-

Table 1. Description of the dataset

Name	OryzaGP
Text genre	Article
Text type	Abstract & title
Entity type	Gene, protein
No. of articles	10,400
No. of sentences	75,096
No. of words	2,697,726

```
here            RB   O
we              PRP  O
attempted       VBN  O
biochemical     JJ   O
characterization NN  O
of              IN   O
cyp99a2         NN   B-gene
and             CC   O
cyp99a3         NN   B-gene
which           WDT  O
was             VBD  O
ultimately      RB   O
achieved        VBN  O
```

Fig. 1. Example of data after pre-processing.

https://doi.org/10.5808/GI.2019.17.2.e17

en in the following lines; one token per line, and included three tabs: the word itself, the POS tag, and the entity type (Fig. 1). To minimize the errors of inaccurate tags assignation when running the script, preprocessed data were checked manually, based on the existing resources (Oryzabase gene list, etc.)

During the 5th Biomedical Linked Annotation Hackathon (BLAH5, http://blah5.linkedannotation.org), a portion of the dataset (29,098 annotation instances made to 6,107 abstracts) was uploaded to the PubAnnotation repository (http://pubannotation.org/projects/OryzaGP), which uses JSON (JavaScript Object Notation) as its default format, to store annotations. Resultantly, the dataset is accessible through or downloadable from PubAnnotation. Sharing it through PubAnnotation also means that the dataset can be compared to annotations from other projects if they share the same documents [7].

Future Work

Our ultimate goal of sharing the dataset is to offer a shared task of rice gene/protein NER, through the BioNLP Open Shared Tasks (BioNLP-OST) framework, to facilitate open comparison and evaluation of various approaches to the task. Toward the goal, we will further upload remaining annotation data, while keeping some portion of it hidden for test dataset.

ORCID

Pierre Larmande: https://orcid.org/0000-0002-2923-9790
Huy Do: https://orcid.org/0000-0003-2588-2858
Yue Wang: https://orcid.org/0000-0001-6230-3275

Authors' Contribution

Conceptualization: PL. Formal analysis: PL. Methodology: HD. Writing – original draft: PL, HD. Writing – review & editing: YW.

Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

This work was supported by IRD UMR DIADE and CGIAR CRP RICE.

References

1. Deans AR, Lewis SE, Huala E, Anzaldo SS, Ashburner M, Balhoff JP, et al. Finding our way through phenotypes. PLoS Biol 2015;13:e1002033.
2. Venkatesan A, Tagny Ngompe G, Hassouni NE, Chentli I, Guignon V, Jonquet C, et al. Agronomic Linked Data (AgroLD): a knowledge-based system to enable integrative biology in agronomy. PLoS One 2018;13:e0198270.
3. Gupta P, Naithani S, Tello-Ruiz MK, Chougule K, D'Eustachio P, Fabregat A, et al. Gramene database: navigating plant comparative genomics resources. Curr Plant Biol 2016;7-8:10-15.
4. Yamazaki Y, Sakaniwa S, Tsuchiya R, Nonomura KI, Kurata N. Oryzabase: an integrated information resource for rice science. Breed Sci 2010;60:544-548.
5. Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, Foerster H, et al. The Arabidopsis Information Resource (TAIR): gene structure and function annotation. Nucleic Acids Res 2008;36:D1009-D1014.
6. Do H, Than K, Larmande P. Evaluating named-entity recognition approaches in plant molecular biology. In: 12th Multi-disciplinary International Conference on Artificial Intelligence (MIWAI 2018), 2018 Nov 18-20, Hanoi, Vietnam. Cham: Springer, 2018. pp. 219-225.
7. Kim JD, Wang Y. PubAnnotation: a persistent and shareable corpus and annotation repository. In: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (BioNLP 2012) (Cohen KB, Demner-Fushman D, Ananiadou S, Webber B, Tsujii J, Pestian J, eds.), 2012 Jun 8, Montreal, Canada. Stroudsburg: Association for Computational Linguistics, 2012. pp. 202-205.

Instructions for authors

Enacted January 2003
Recently revised January 9, 2019

Genomics & Informatics (Genomics Inform) is owned and published by the Korea Genome Organization (KOGO). It is published four times per year (Mar, Jun, Sep, and Dec) in an online version. *Genomics & Informatics* welcomes high-quality research papers presenting novel data on the topics of gene discovery, comparative genome analyses, molecular and human evolution, informatics, genome structure and function, technological innovations and applications, statistical and mathematical methods, cuttingedge genetic and physical mapping, and DNA sequencing and other reports that present data where sequence information is used to address biological concerns. The journal publishes papers based on original research that are judged after editorial review to make a substantial contribution to the understanding of any area of genomics or informatics. Only manuscripts written in English under the *Genomics & Informatics* author guidelines are accepted. *Genomics & Informatics* follows the open access journal policy. All of the content of *Genomics & Informatics* is freely available online. Digital files can be read, downloaded, and printed without charge.

Manuscripts for submission to *Genomics & Informatics* should be prepared according to the following instructions. *Genomics & Informatics* follows the Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals (http://www.icmje.org) from ICMJE and Principles of Transparency and Best Practice in Scholarly Publishing (joint statement by COPE, DOAJ, WAME, and OASPA; (http://doaj.org/bestpractice) if otherwise not described below.

Research and publication ethics

For the policies on research and publication ethics that are not stated in these instructions, the Good Publication Practice Guidelines for Medical Journals (http://kamje.or.kr/intro.php?body=publishing_ethics) and the Guidelines on Good Publication(http://publicationethics.org/resources/guidelines) can be applied. The Editor-in-Chief reserves the right to reject manuscripts that do not comply with the below requirements. The author will be held responsible for false statements or failure to fulfill the below requirements.

Statement of Informed Consent

Copies of written informed consent and Institutional Review Board (IRB) approval for clinical research should be kept. If necessary, the editor or reviewers may request copies of these documents to resolve questions about IRB approval or study conduct.

Statement of Human and Animal Rights

All human investigations must be conducted according to the principles expressed in the Declaration of Helsinki. All studies involving animals must state that the guidelines for the use and care of laboratory animals of the authors' institution, or of any national law, were followed. Registration of clinical trial research: Any research that deals with a clinical trial should be registered with the primary national clinical trial registry site, such as the Korea Clinical Research Information Service (CRiS, http://cris.nih.go.kr), other primary national registry sites accredited by the World Health Organization (http://www.who.int/ictrp/network/primary/en/), or ClinicalTrials.gov (http://clinicaltrials.gov/), a service of the United States National Institutes of Health.

Authorship

Authorship credit should be based on 1) substantial contributions to conception and design, acquisition of data, and/or analysis and interpretation of data; 2) drafting the article or revising it critically for important intellectual content; 3) final approval of the version to be published; and 4) agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. Every author should meet all of these four conditions. After the initial submission of a manuscript, any changes whatsoever in authorship (adding author(s), deleting author(s), or re-arranging the order of authors) must be explained by a letter to the editor from the authors concerned. This letter must be signed by all authors of the paper. Copyright assignment must also be completed by every author.

Corresponding author and first author

It does allow multiple corresponding authors for one article. Only one author should correspond with the editorial office. It does accept notice of equal contribution for the first author when the study was clearly performed by co-first authors.

Correction of authorship after publication

It does not correct authorship after publication unless a mistake has been made by the editorial staff. Authorship may be changed before publication but after submission when an authorship correction is requested by all of the authors involved with the manuscript.

Conflict of Interest Statement

The corresponding author must inform the editor of any potential conflicts of interest that could influence the authors' interpretation of the data. Examples of potential conflicts of interest are financial support from or connections to pharmaceutical companies, political pressure from interest groups, and academically related issues. In particular, all sources of funding applicable to the study should be explicitly stated. As a guideline, any affiliation associated with a payment or financial benefit exceeding $10,000 per annum or 5% ownership of a company or research funding by a company with related interests would constitute a conflict that must be declared. This policy applies to all submitted research manuscripts and review material.

Originality and Duplicate Publication

No part of the accepted manuscript should be duplicated in any other scientific journal without the permission of the Editorial Board. If duplicate publication or plagiarism related to the papers of this journal is detected, the authors will be announced in the journal, their institutes will be informed, and the authors will be penalized. All submitted manuscripts are screened by CrossCheck (Similarity Check), a plagiarism detection program provided by iThenticate. The authors assure that no substantial part of the work has been published or is being considered for publication elsewhere. When any of the results is to appear in another journal, details must be submitted to the Editor-in-Chief, together with a copy of the other paper(s) and the expected date(s) of publication.

Secondary Publication

It is possible to republish manuscripts if the manuscripts satisfy the condition of secondary publication of the Uniform Requirements for Manuscripts Submitted to Biomedical Journals by the International Committee of Medical Journal Editors (ICMJE), available from http://www.icmje.org/. These are:

- The authors have received approval from the editors of both journals (the editor concerned with the secondary publication must have access to the primary version).
- The priority for the primary publication is respected by a publication interval negotiated by editors of both journals and the authors.
- The paper for secondary publication is intended for a different group of readers; an abbreviated version could be sufficient.
- The secondary version faithfully reflects the data and interpretations of the primary version.
- The secondary version informs readers, peers, and documenting agencies that the paper has been published in whole or in part elsewhere—for example, with a note that might read, "This article is based on a study first reported in the [journal title, with full reference]"—and the secondary version cites the primary reference.
- The title of the secondary publication should indicate that it is a secondary publication (complete or abridged republication or translation) of a primary publication. Of note, the United States National Library of Medicine (NLM) does not consider translations to be "republications" and does not cite or index them when the original article was published in a journal that is indexed in MEDLINE.

Process to manage research and publication misconduct: When the Journal faces suspected cases of research and publication misconduct, such as a redundant (duplicate) publication, plagiarism, fabricated data, changes in authorship, undisclosed conflicts of interest, an ethical problem discovered with the submitted manuscript, a reviewer who has appropriated an author's idea or data, complaints against editors, and other issues, the resolving process will follow a flowchart provided by the Committee on Publication Ethics (http://publicationethics.org/resources/flowcharts). The discussion and decision on suspected cases are done by the Editorial Board of *Genomics & Informatics*.

Preparation of manuscripts

General requirement

Authors are recommended to keep the length of papers below 10 printed pages (30 typed pages of manuscript, including figures and tables) for original articles, four printed pages for research communications, and two printed pages (approximately 1,400 words or 1,000 words plus one figure) for application notes. All sections of the typescript should be double-spaced on one side of A4 paper (210 × 297 mm), and all pages must be numbered in order.

Manuscript type

Original articles

Original research articles are full scientific reports of original research. The manuscript should be organized as follows: Title Page, Abstract & Keywords, Introduction, Methods, Results,

Discussion, Acknowledgments, References, Tables, and Figure Legends. The Results and Discussion can be combined.

Application notes

Application notes are short communications about novel software, new algorithm implementations, databases, and network services (web servers and interfaces). The manuscripts include the following: Title Page, Abstract & Keywords, Availability, Introduction, Main Text, Conclusion, References, and Supplementary Information.

Clinical genomics

Clinical genomics is for a short report of all kinds of genome analysis data from clinical fields, such as cancer, diverse complex diseases, and genetic diseases. Especially, *Genomics & Informatics* would encourage submitting cancer panel analysis data for a single cancer patient or a group of patients. *Genomics & Informatics* also would encourage depositing genome data into the *Genomics & Informatics* database. The manuscript should be organized as follows: Title Page, Abstract & Keywords, Introduction, Methods, Results, Discussion, Acknowledgments, References, Tables, and Figure Legends. The Introduction, Methods, Results, and Discussion can be combined.

Genome archives

Genome Archives is for a short manuscript announcing the genetic information of recently sequenced prokaryotic and eukaryotic genomes. *Genomics & Informatics* would encourage depositing the genome data into the *Genomics & Informatics* database. These genome archive data can make the rationale for sequencing a specific organism. The manuscript should be organized as follows: Title Page, Abstract & Keywords, Introduction, Methods, Results, Discussion, Acknowledgments, References, Tables, and Figure Legends. The Introduction, Methods, Results, and Discussion can be combined.

Letters to the editor

Critical comments are welcomed for correcting errors of published facts and for providing alternative interpretations of published data. The sequence for a Letter to the Editor is the following: Title Page, Text, References, and Names and Affiliations of Authors. If needed, tables and figures can be included. A Letter to the Editor should not be longer than a printed page.

Review articles

Review Articles are usually solicited by the Editor-in-Chief. Authors wishing to prepare a review article should contact the Editor-in-Chief to discuss the suitability of the subject for the journal. There is no specific requirement for subsections of the body text of the paper.

Opinions/Commentaries

An opinion or commentary piece is a short article that conveys the author's viewpoint on a research publication, including interpretation of data, value of methods used, and strengths/weaknesses, regarding any topic relevant to the field of research. Opinion (or commentary) articles provide insight, interpretation, and evaluation of specific issues, within the scope of the journal. Opinions should explain the implications of the article and describe the most important conclusions of the paper they are commenting on, highlight controversial issues, mention the strengths and weaknesses of the paper, highlight the presenter's omission of key facts, and mention supporting arguments that would create a stronger presentation. Opinions are relatively short articles, around 1000 words, allowing maximum freedom of authors' viewpoints, and are peer-reviewed. The articles are copyedited, citable, published in both PDF and HTML formats, and submitted for indexing in digital archives (e.g., PubMed Central). Authors are not required to pay a fee to publish an opinion (or commentary) article. Commentaries have no set format beyond the basic building blocks of a regular article, i.e., title, manuscript text, subheadings as needed, references, and author information.

Minireviews

Minireview articles are similar to review articles, except for their word limit and references. Minireviews focus on clearly defined topics of current interest, and recent developments in specific fields. Therefore, they offer a fast and easy means to keep abreast of exciting new developments and/or concepts. The word limit for minireview articles is 1000 words (or 2 double-spaced pages), with no more than 30 references. Minireview articles are peer-reviewed, copyedited, citable, published in both PDF and HTML formats, and submitted for indexing in digital archives, such as PubMed Central. Authors are required to pay a fee to publish a minireview.

Manuscript Format

Title

The title page should include (1) the full names of all authors with their Open Researchers and Contributors ID (ORCID), and the name(s) and address(es) of the institution(s) at which the work was carried out; (2) the telephone and fax numbers, and the

E-mail address of the corresponding author; and (3) a running title of no more than 50 characters, including spaces. Place an asterisk (*) after the corresponding author.

Abstract

The abstract should be unstructured and a single paragraph of fewer than 250 words. References should not be cited in the abstract. Six or fewer keywords should be appended to the abstract in alphabetical order. When possible, the keywords should be those found in the Medical Subject Headings of Index Medicus.

Main text:

All papers should be divided into the following sections and appear in this order:

(1) **Introduction:** The paper begins with an introduction without subheadings that reviews the literature and states and justifies the purpose of the research.

(2) **Methods:** This section should contain sufficient detail so that all procedures can be repeated, in conjunction with the cited references. The manufacturer and model number should be stated in this section—for example, as Sigma Chemical Co. (St. Louis, MO, USA).

(3) **Results:** This section should describe the results of the experiments. Extensive interpretation should be reserved for the Discussion section. The results should be presented as concisely as possible. Footnotes should not be used and will be transferred to the text. Gene symbols should be italicized; protein products are not italicized.

(4) **Discussion:** This section should provide an interpretation of the results in relation to previously published work and to the experimental system at hand. The Results and Discussion may be combined.

(5) **Acknowledgments:** Information concerning the sources of financial support should be included in the acknowledgments.

Authors' contribution

If the number of authors is equal to or greater than two, the authors' roles should be described according to their specific role. *Genomics & Informatics* participates in the CRediT standard for author contributions. The contributions of all authors must be described using the CRediT Taxonomy of author roles. For each of the categories below, please enter the initials of the authors who contributed in that category. If listing more than one author in a category, separate each set of initials with a space. If no one contributed in a category, you may leave that box blank. The corresponding author is responsible for completing this information at submission, and it is expected that all authors will have reviewed, discussed, and agreed to their individual contributions ahead of this time.

- Conceptualization: AB
- Data curation: EFG
- Formal analysis: AB
- Funding acquisition: CD
- Methodology: AB, CD, EFG
- Writing – original draft: AB, EFG
- Writing – review & editing: AB, CD, EFG

Reference

The references should include only articles that are published or in press. Unpublished data, submitted manuscripts, abstracts, and personal communications should be cited within the text only. References are to be numbered in the order of citation within the article in brackets. References with up to six authors must list all names; for more than six authors, the first six names should be listed, followed by "et al." Journal name titles should be abbreviated in accordance with the NLM Catalog, available from: https://www.ncbi.nlm.nih.gov/nlmcatalog/journals, or the ISO 4 standard, available from: http://www.issn.org/services/online-services/access-to-the-ltwa/?letter=a.

Examples of references are given below:

Journal article

- Park J, Lappe M, Teichmann SA. Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. J Mol Biol 2001;307:929-938.
- Cho SM, Jung SH, Chung YJ. A variant in RUNX3 is associated with the risk of ankylosing spondylitis in Koreans. Genomics Inform 2017;15:65-68.
- Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, et al. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res 2003;13:2129-2141.

Books

- Cowan WM, Jessell TM, Zipursky SL. Molecular and Cellular Approaches to Neural Development. New York: Oxford University Press, 1997.

Book sections

- Sorenson PW, Caprio JC. Chemoreception. In: The Physiology of Fishes (Evans DH, ed.). Boca Raton: CRC Press, 1998. pp.

375-405.

Online document

- Puniyani AR, Lukose RM. Growing random networks under constraints. Ithaca: Cornell University Library, 2001. Accessed 2011 Oct 3. Available from: http://xxx.lanl.gov/abs/cond-mat/0107391.

Conference paper

- Han H. Nonnegative principle component analysis for mass spectral serum profiles and biomarker discovery. In: The 8th Asia-Pacific Bioinformatics Conference (Parida L, Myers G, eds.), 2010 Jan 18-21, Bangalore.

Dissertation/Thesis

- Hwang KB. Hierarchical probabilistic graphical models for large-scale data analysis. Ph.D. Dissertation. Seoul: Seoul National University, 2005.

Tables and figures

Figure legends and tables should be included in the submitted manuscript as separate sections and should be formatted following the style of the journal. Each figure legend should have a brief, separate title that describes the entire figure without citing specific panels. The manuscript should be submitted with a set of figures of sufficient quality for reviewers to judge the data. All figures may be provided in color for the electronic version of the journal, even if the print version is in black and white. Figures will be printed in color only when in the reviewers' opinions the color is essential.

Photographs and illustrations should be of professional quality. Images should be provided as TIFF files. JPEG is also acceptable when the original format is JPEG. Each figure must be of 300 dpi or higher resolution with good contrast and sharpness. If a figure is to be reduced, all elements, including labels, should be able to withstand reduction and remain legible. Electron and light microscopic figures must be original or scanned copies from the original. The magnification should be indicated on each micrograph with a scale bar.

Tables are to be organized in portrait view and may run, if necessary, to subsequent pages in the vertical direction only. Tables should be designed for printing within two (17.5 cm) columns of width in no less than 10-point font and should not exceed more than the width of a journal page. If a table does not fit into this format, consider shortening row or column labels, using more than one table to display the data, eliminating unnecessary data, or converting table data into a figure or transferring part of the table data to the supplement.

Scientific names

The full formal Latin name for a taxon (e.g., Homo sapiens) should be provided the first time that the taxon is mentioned and should be italicized. In subsequent sentences, the scientific name of all taxa in the same genus should be abbreviated to the first initial of the generic name and the species name (e.g., H. sapiens), except where this usage creates confusion or ambiguity. When common names are used, the scientific name should be provided the first time the taxon is mentioned in the abstract and again the first time that taxon is mentioned in the main manuscript [e.g., "red pine (Pinus densiflora)..."]. Other taxonomic designations (e.g., family names) should not be italicized, and common names should not be capitalized.

Units and equations

Standard metric units should be used for describing length, height, weight, and volume. The unit of temperature is given in degrees Celsius (°C). All others are in terms of the International System of Units (SI). All unit symbols must be preceded by one space except percentage (%) and temperature (°C). All equations should be numbered in Arabic numerals.

Abbreviations

Abbreviations must be used as an aid to the reader, rather than as a convenience of the author, and therefore, their use should be limited. Generally, avoid abbreviations that are used less than 3 times in the text, including the tables and figure legends. In addition to abbreviations for SI units, common molecular, chemical, immunological, and hematological terms can be used without definition in the title, abstract, text, tables, and figure legends—e.g., bp, kb, kDa, DNA, cDNA, RNA, mRNA, and PCR. Other common abbreviations are as follows (the same abbreviations are used for plural forms): h (hour; use 0-24:00 h for time), s (second), min (minute), day (not abbreviated), week (not abbreviated), month (not abbreviated), year (not abbreviated), L (liter), mL (milliliter), μL (microliter), g (gram), kg (kilogram), mg (milligram), μg (microgram), ng (nanogram), pg (picogram), g (gravity; not ×g), n (sample size), SD (standard deviation of the mean), and SE (standard error of the mean).

Supplementary materials

Supplementary materials can be provided to support and enhance scientific information. Supplementary files offer additional possibilities for publishing supporting applications, sequence alignment, background datasets, microarray hybridization

experiments, high-resolution images, movies, sound clips, and more. Supplementary files will be published alongside the online version of the article on the *Genomics & Informatics* web site. This material will not be edited or formatted; thus, the authors are responsible for the accuracy and presentation of all such material.

Accepted file formats for supplementary materials:

- Quick Time files (.mov)
- Graphical image files (.gif)
- HTML files (.html)
- MPEG movie files (.mpg)
- JPEG image files (.jpg)
- Sound files (.wav)
- Plain ASCII text (.txt)
- Acrobat files (.pdf)
- MS Word documents (.doc)
- Postscript files (.ps)
- MS Excel spreadsheet documents (.xls)
- PowerPoint (.ppt)
- TeX and LaTeX

File sizes must be as small as possible, for quick downloading. Recommended specifics are:

- Videos
  - File size: <150 MB
  - Frame rate: 30 frames per second
  - Field order: none (progressive, not interlaced)
  - Aspect ratio: widescreen 16:9
  - Video codec: H.264
  - Video bitrate: 2 Mbps
  - Audio codec: AAC
  - Audio bitrate: 128 kbps
- Images
  - Frame size: 300 dpi in resolution
  - Frame rate: 300 dpi in resolution and 10-15cm in width

Please seek advice from the editorial office before sending files larger than our recommended size to avoid delays in publication.

Accession numbers

Please provide accession numbers for any new data (SNPs, gene sequences, protein sequences, CNVs, microarray data, or structures), which must be deposited in the appropriate genome- or locusspecific database, in a separate section entitled "Accession Numbers," following the Web Resources section (or the Acknowledgments section if no online resources or appendices have been used), directly above the reference list. Please use the following format to list accession numbers: "The accession number(s) for the _____ sequence(s) reported in this paper is/are [database]: [accession number]."

Gender equity (Described according to ICMJE recommendation available from

http://www.icmje.org/recommendations/browse/manuscript-preparation/preparing-for-submission.html)

Selection and Description of Participants

Clearly describe the selection of observational or experimental participants (healthy individuals or patients, including controls), including eligibility and exclusion criteria and a description of the source population. Because the relevance of such variables as age, sex, or ethnicity is not always known at the time of study design, researchers should aim for inclusion of representative populations into all study types and at a minimum provide descriptive data for these and other relevant demographic variables. Ensure correct use of the terms sex (when reporting biological factors) and gender (identity, psychosocial or cultural factors), and, unless inappropriate, report the sex and/or gender of study participants, the sex of animals or cells, and describe the methods used to determine sex and gender. If the study was done involving an exclusive population, for example in only one sex, authors should justify why, except in obvious cases, (e.g., prostate cancer)." Authors should define how they determined race or ethnicity and justify their relevance.

Submission of Manuscript

The manuscript should be submitted in MS Word file format. The recommended font is Times New Roman with a 11-point font size. All manuscripts must be submitted online through the *Genomics & Informatics* e-submission system at http://submit.genominfo.org. Any questions concerning manuscript submission should be directed to: Editor, *Genomics & Informatics*, Korea Genome Organization, The Korean Federation of Science and Technology Societies, Room number 1011, 22, Teheran-ro 7-gil, Gangnam-gu, Seoul, 06130, Korea (http://www.kogo.or.kr, Tel: +82-2-558-9394, Fax: +82-2-558-9434, E-mail: kogo@kogo.or.kr).

Peer review and revision of manuscripts

Peer review

A manuscript is generally reviewed by at least two peer reviewers qualified to evaluate the manuscript. It is a single blind peer

review. An initial decision will normally be made within one month of receipt of a manuscript. A manuscript that has been published or of which a substantial portion has been published elsewhere will not be accepted. The Editor-in-Chief is responsible for final decisions regarding the acceptance of a peer-reviewed paper.

Manuscript revision

When a manuscript is returned to the corresponding author for revision, the reviewed manuscript must be re-submitted within one month, unless the authors request an extension. A galley proof and reprint order form will be sent to the corresponding author. The corresponding author is responsible for communicating with the other authors about revisions and final approval of the proofs. The first proofreading is the author's responsibility, and the proof should be returned within three days from the date of receipt.

Copyrights, open access policy and open data policy

Copyright

The regulations for acceptance of a manuscript for publication automatically include the consent of the author(s) to transfer the copyright or license to KOGO. Authors should complete a Copyright Agreement Form (CAF) at the time of proofreading. The corresponding author can sign on behalf of any co-authors. The CAF can be obtained from the editorial office. Acceptance of the agreement will ensure full copyright protection and help to disseminate the article to the widest possible readership in print and electronic formats. The authors are responsible for obtaining permission to reproduce copyrighted material from other sources

Open access policy

Genomics & Informatics is an open access journal. Articles are distributed under the terms of the Creative Commons Attribution license (http://creativecommons.org/licenses/by-nc/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited for non-commercial purposes. To use the tables or figures of *Genomics & Informatics* in other periodicals, books, or media for scholarly, educational, or even commercial purposes, the process of permission request to the Publisher is not necessary. This is in accordance with the Budapest Open Access Initiative definition of open access. It also follows the open access policy of PubMed Central at the United States National Library of Medicine (http://www.ncbi.nlm.nih.gov/pmc/). All of the content of the journal is available immediately upon publication without an embargo period.

Archiving policy

It is accessible without barrier from Korea Citation Index (https://kci.go.kr), National Library of Korea (http://nl.go.kr), or PubMed Central (https://www.ncbi.nlm.nih.gov/pmc/journals/1928/) in the event a journal is no longer published.

Deposit policy (Self-archiving policy) according to Sherpa/Romeo

(http://www.sherpa.ac.uk/): Author can not archive pre-print (i.e., pre-refereeing). Author can archive post-print (i.e., final draft post-refereeing).

Author can archive publisher's version/PDF.

Open data policy

Data sharing is recommended. If the data are already public, the URL site or sources should be disclosed. If data can not be publicized, it can be negotiated with the editor. If there are any inquiries on depositing data, author s should contact the editorial office.

Clinical data sharing policy

This journal follows the data sharing policy described in "Data Sharing Statements for Clinical Trials: A Requirement of the International Committee of Medical Journal Editors" (https://doi.org/10.3346/jkms.2017.32.7.1051). As of July 1, 2018, manuscripts submitted to ICMJE journals that report the results of clinical trials must contain a data sharing statement as described below. Clinical trials that begin enrolling participants on or after January 1, 2019 must include a data sharing plan in the trial's registration. The ICMJE's policy regarding trial registration is explained at www.icmje.org/recommendations/browse/publishingand-editorial-issues/clinical-trial-registration.html. If the data sharing plan changes after registration, this should be reflected in the statement submitted and published with the manuscript and updated in the registry record. Data sharing statements must indicate the following: whether individual deidentified participant data (including data dictionaries) will be shared; what data in particular will be shared; whether additional, related documents will be available (e.g., study protocol, statistical analysis plan, etc.); and when the data will become available and for how long; by what access criteria data will be shared (including with whom, for what types of analyses, and by what mechanism). Illustrative examples of data sharing statements that would meet these requirements are in Table 1.

Table 1. Examples of data sharing statements that fulfill ICMJE requirements[a]

Element	Example 1	Example 2	Example 3	Example 4
Will individual participant data be available (including data dictionaries)?	Yes	Yes	Yes	No
What data in particular will be shared?	All of the individual participant data collected during the trial, after deidentification.	Individual participant data that underlie the results reported in this article, after deidentification (text, tables, figures, and appendices).	Individual participant data that underlie the results reported in this article, after deidentification (text, tables, figures, and appendices).	Not available
What other documents will be available?	Study protocol, statistical analysis plan, informed consent form, clinical study report, analytic code	Study protocol, statistical analysis plan, analytic code	Study protocol	Not available
When will data be available (start and end dates)?	Immediately following publication. No end date.	Beginning 3 months and ending 5 years following article publication.	Beginning 9 months and ending 36 months following article publication.	Not applicable
With whom?	Anyone who wishes to access the data.	Researchers who provide a methodologically sound proposal.	Investigators whose proposed use of the data has been approved by an independent review committee ("learned intermediary") identified for this purpose.	Not applicable
For what types of analyses?	Any purpose	To achieve aims in the approved proposal.	For individual participant data meta-analysis.	Not applicable
By what mechanism will data be made available?	Data are available indefinitely at (link to be included).	Proposals should be directed to xxx@yyy. To gain access, data requestors will need to sign a data access agreement.	Proposals may be submitted up to 36 months following article publication. After 36 months, the data will be available in our University's data warehouse but without investigator support other than deposited metadata.	Not applicable
			Information regarding submitting proposals and accessing data may be found at (link to be provided).	
		Data are available for 5 years at a third-party website (link to be included).		

ICMJE, International Committee of Medical Journal Editors.
[a]These examples are meant to illustrate a range of, but not all, data sharing options.

Detailed Description of Use of Articles of *Genomics & Informatics*

Reader benefit

Publisher applies the Creative Commons Attribution Non-Commercial license to works it publishes and allows free immediate access to, and unrestricted reuse of, original works of all types.

Reuse benefit

Publisher applies the Creative Commons Attribution Non-Commercial license to works it publishes and allows free immediate access to, and unrestricted reuse of, original works of all types.

Copyrights

Publisher applies the Creative Commons Attribution Non-Commercial license to works it publishes. Under this license, although publisher retains ownership of the copyright for content, it allows anyone to download, reuse, reprint, modify, distribute, and/or copy the content.

Author posting benefit:

Publisher applies the Creative Commons Non-Commercial Attribution license to works it publishes. Under this license, although publisher retains ownership of the copyright for content, it allows anyone, including author, to download, reuse, reprint, modify, distribute, and/or copy the content.

Automatic Posting:

Publisher immediately deposits the accepted articles in PubMed Central (http://pubmedcentral.org/) and journal homepage (https://genominfo.org/) upon publication.

Machine readability:

Genomics & Informatics articles can be accessed programmatically

through PubMed Central or Europe PMC's RESTful Web Service (https://europepmc.org/RestfulWebService). For inquiries, please contact editorial office, as below:

Article processing charge

Neither page charge, article processing fee nor submission fee will be applied since 2019. It is the platinum open access journal

Contact address

Editorial office

Rm.1011, The Korea Science & Technology Center, 22, Teheran-ro, 7-gil, Gangnam-gu, Seoul, 06130, Korea

Tel: +82-2-558-9394

Fax: +82-2-558-9434

E-mail: kogo3@kogo.or.kr

Copyright transfer agreement

The copyright to this article is transferred to Genomics & Informatics effective if and when the article is accepted for publication. The author warrants that his/her contribution is original and that he/she has full power to make this grant. The author signs for and accepts responsibility for releasing this material on behalf of any and all co-authors. The copyright transfer covers the exclusive right to reproduce and distribute the article, including reprints, translations, photographic reproductions, microform, electronic form (offline, online) or any other reproductions of similar nature.

According to the deposit policy (self-archiving policy) of Sherpa/Romeo (http://www.sherpa.ac.uk), authors cannot archive pre-print (i.e. pre-refereeing), but they can archive post-print (i.e. final draft post-refereeing). Authors can archive publisher's version/PDF.

Title of article	
Author(s)	
Author's signature	
Date	

Taesung Park
Editor in Chief
Genomics & Informatics
Korea Genome Organization (KOGO)

Publication ethics

For the policies on research and publication ethics that are not stated in these instructions, the Good Publication Practice Guidelines for Medical Journals (http://kamje.or.kr/intro.php?body=publishing_ethics) and the Guidelines on Good Publication(http://publicationethics.org/resources/guidelines) can be applied. The Editor-in-Chief reserves the right to reject manuscripts that do not comply with the below requirements. The author will be held responsible for false statements or failure to fulfill the below requirements.

Statement of Informed Consent

Copies of written informed consent and Institutional Review Board (IRB) approval for clinical research should be kept. If necessary, the editor or reviewers may request copies of these documents to resolve questions about IRB approval or study conduct.

Statement of Human and Animal Rights

All human investigations must be conducted according to the principles expressed in the Declaration of Helsinki. All studies involving animals must state that the guidelines for the use and care of laboratory animals of the authors' institution, or of any national law, were followed. Registration of clinical trial research: Any research that deals with a clinical trial should be registered with the primary national clinical trial registry site, such as the Korea Clinical Research Information Service (CRiS, http://cris.nih.go.kr), other primary national registry sites accredited by the World Health Organization (http://www.who.int/ictrp/network/primary/en/), or ClinicalTrials.gov (http://clinicaltrials.gov/), a service of the United States National Institutes of Health.

Authorship

Authorship credit should be based on 1) substantial contributions to conception and design, acquisition of data, and/or analysis and interpretation of data; 2) drafting the article or revising it critically for important intellectual content; 3) final approval of the version to be published; and 4) agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. Every author should meet all of these four conditions.

After the initial submission of a manuscript, any changes whatsoever in authorship (adding author(s), deleting author(s), or re-arranging the order of authors) must be explained by a letter to the editor from the authors concerned. This letter must be signed by all authors of the paper. Copyright assignment must also be completed by every author.

Corresponding author and first author

It does allow multiple corresponding authors for one article. Only one author should correspond with the editorial office. It does accept notice of equal contribution for the first author when the study was clearly performed by co-first authors.

Correction of authorship after publication

It does not correct authorship after publication unless a mistake has been made by the editorial staff. Authorship may be changed before publication but after submission when an authorship correction is requested by all of the authors involved with the manuscript.

Conflict of Interest Statement

The corresponding author must inform the editor of any potential conflicts of interest that could influence the authors' interpretation of the data. Examples of potential conflicts of interest are financial support from or connections to pharmaceutical companies, political pressure from interest groups, and academically related issues. In particular, all sources of funding applicable to the study should be explicitly stated. As a guideline, any affiliation associated with a payment or financial benefit exceeding $10,000 per annum or 5% ownership of a company or research funding by a company with related interests would constitute a conflict that must be declared. This policy applies to all submitted research manuscripts and review material.

Originality and Duplicate Publication

No part of the accepted manuscript should be duplicated in any other scientific journal without the permission of the Editorial Board. If duplicate publication or plagiarism related to the papers of this journal is detected, the authors will be announced in the journal, their institutes will be informed, and the authors will be penalized. All submitted manuscripts are screened by CrossCheck

(Similarity Check), a plagiarism detection program provided by iThenticate. The authors assure that no substantial part of the work has been published or is being considered for publication elsewhere. When any of the results is to appear in another journal, details must be submitted to the Editor-in-Chief, together with a copy of the other paper(s) and the expected date(s) of publication.

Secondary Publication

It is possible to republish manuscripts if the manuscripts satisfy the condition of secondary publication of the Uniform Requirements for Manuscripts Submitted to Biomedical Journals by the International Committee of Medical Journal Editors (ICMJE), available from http://www.icmje.org/. These are:

- The authors have received approval from the editors of both journals (the editor concerned with the secondary publication must have access to the primary version).
- The priority for the primary publication is respected by a publication interval negotiated by editors of both journals and the authors.
- The paper for secondary publication is intended for a different group of readers; an abbreviated version could be sufficient.
- The secondary version faithfully reflects the data and interpretations of the primary version.
- The secondary version informs readers, peers, and documenting agencies that the paper has been published in whole or in part elsewhere—for example, with a note that might read, "This article is based on a study first reported in the [journal title, with full reference]"—and the secondary version cites the primary reference.
- The title of the secondary publication should indicate that it is a secondary publication (complete or abridged republication or translation) of a primary publication. Of note, the United States National Library of Medicine (NLM) does not consider translations to be "republications" and does not cite or index them when the original article was published in a journal that is indexed in MEDLINE.

Process to manage research and publication misconduct: When the Journal faces suspected cases of research and publication misconduct, such as a redundant (duplicate) publication, plagiarism, fabricated data, changes in authorship, undisclosed conflicts of interest, an ethical problem discovered with the submitted manuscript, a reviewer who has appropriated an author's idea or data, complaints against editors, and other issues, the resolving process will follow a flowchart provided by the Committee on Publication Ethics (http://publicationethics.org/resources/flowcharts). The discussion and decision on suspected cases are done by the Editorial Board of Genomics & Informatics.

Author's checklist

☐ 1. Typed double-spaced with 12-point font in Times New Roman font on A4 sized paper and prepared with an MS-word file.

☐ 2. Title page: (1) complete title, (2) manuscript type, (3) authors' name, (4) affiliations, (5) telephone, facsimle and E-mail address of corresponding author, (6) running title (no more than 50 characters).

☐ 3. Abstract in unstructured format within 250 words.

☐ 4. Six or fewer keywords, preferably MeSH terms.

☐ 5. Manuscript is structured as follows:
Original Article: Abstract, Keywords, Introduction, Materials and Methods, Results, Discussion, References, Table and Figure.
Research Communication: Abstract, Keywords, Main Text, and Conclusion (if applicable), References, Table and Figure.
Application Note: Abstract, Keywords, Availability, Introduction, Main Text, and Supplementary Information, References, Table and Figure.

☐ 6. Reference in proper format. Check that all references listed in the references section are cited in the text and vice versa.

☐ 7. All figures and tables referenced in the text and numbered in order of its appearance in the text.

☐ 8. Figures as a separate files, in TIFF or JPG format, minimum 300 dpi.

☐ 9. Each necessary permission statement signed by the appropriate source.

☐ 10. Elucidation of research or project support/funding.