www.genominfo.org

eISSN 2234-0742

GENOMICS & INFORMATICS

Vol. 18 - No. 2, June 30 2020

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

www.genominfo.org
eISSN 2234-0742

G&I
Genomics & Informatics

Volume 18 number 2, June 30, 2020

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, Scopus, 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 No. 806, 193 Mallijae-ro, Jung-gu, Seoul 04501, Korea
Tel: +82-2-558-9394, Fax: +82-2-558-9434, email: kogo3@kogo.or.kr, URL address: https://genominfo.org

G&I
Genomics & Informatics

Editorial Board

G&I
Genomics & Informatics

Volume 18 number 2, June 30, 2020

G&I
Genomics & Informatics

Editorial

eISSN 2234-0742
Genomics Inform 2020;18(2):e12
https://doi.org/10.5808/GI.2020.18.2.e12

*Corresponding author:
E-mail: jindong.kim@gmail.com

Editor's introduction to the special issue of the 6th Biomedical Linked Annotation Hackathon (BLAH6)

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

[1]Database Center for Life Science (DBCLS), Research Organization of Information and Systems (ROIS), Kashiwa, Chiba 277-0871, Japan
[2]School of Medicine, University of Colorado, Aurora, CO 80045, USA
[3]Dalle Molle Institute for Artificial Intelligence Research (IDSIA), 6928 Manno, Switzerland
[4]National Center for Biotechnology Information (NCBI), National Institutes of Health (NIH), Bethesda, MD 20894, USA
[5]Faculty of Modern & Medieval Languages, University of Cambridge, Cambridge CB3 9DP, UK
[6]Center for Convergence Research of Advanced Technologies, Ewha Womans University, Seoul 03760, Korea

As data science gains in importance and popularity, the need for accessing data in scientific literature is rapidly increasing. While structured databases are supposed to supply readily machine-readable data, unstructured contents, particularly scientific literature, are recognized as a biggest source of data with comprehensive details, e.g., experimental environments and actual observations.

Since the importance of scientific literature for data science has been widely recognized, several groups have invested to develop various text mining resources. While many of them are publicly available, interoperability of them remains a critical issue, hindering efficient use or reuse of them, particularly in mix with others.

The Biomedical Linked Annotation Hackathon (BLAH) series is annually organized to join forces of biomedical text mining for the goal to promote interoperability among text mining resources. The sixth edition of it was held in Tokyo, February 4– 7, 2020, with 52 participants from 9 countries. The first day was held as a symposium to exchange and publicise the activities and ideas of the participants, and the following three days was held as a hackathon: the participants worked on implementing their ideas with collaboration with other participants.

While the main theme of the event was improving interoperability of biomedical literature mining, which include annotation datasets, tools, platforms, terminology resources, and so on, this year, "social media mining" was also explored as a special theme. Social media is recognized as a good source of raw signals on how people are thinking about what is going on in the world, which are largely missing in scientific literature. Therefore, social media mining is expected to complement literature mining.

This special issue is a collection of the reports on achievements from the hackathon, which address various issues of biomedical literature and social media mining, including document collection, automatic annotation, manual annotation, annotation platform, translation, terminology, ontology, and so on. Note that, except a few, many of the works began just before or even during the hackathon, and due to the limited time for work,

they are often small-sized works, which are expected to benefit from collaboration with other participants. Readers will find that many of the articles have co-authorship with, or acknowledgment of other participants, which is a typical nature of hackathon-oriented publications.

We hope that this will be an opportunity for the readers of the journal Genomics & Informatics to get aware of the state-of-the-art activities regarding interoperability of biomedical text mining, and at the same time to observe activities of hackathons like BLAH.

ORCID

Jin-Dong Kim: https://orcid.org/0000-0002-8877-3248
Kevin Bretonnel Cohen: https://orcid.org/0000-0003-1749-8290
Fabio Rinaldi: https://orcid.org/0000-0001-5718-5462
Zhiyong Lu: https://orcid.org/0000-0002-8301-9553
Nigel Collier: https://orcid.org/0000-0002-7230-4164
Hyun-Seok Park: https://orcid.org/0000-0002-6617-2740

Acknowledgments

The 6th Biomedical Linked Annotation Hackathon was held with financial support of National Bioscience Database Center (NBDC) of Japan Science and Technology Agency (JST) and Research Organization of Information and Systems (ROIS).

G&I
Genomics & Informatics

Review article

eISSN 2234-0742
Genomics Inform 2020;18(2):e13
https://doi.org/10.5808/GI.2020.18.2.e13

Received: March 25, 2020
Revised: May 29, 2020
Accepted: June 1, 2020

*Corresponding author:
neo@ewha.ac.kr

Using the PubAnnotation ecosystem to perform agile text mining on *Genomics & Informatics*: a tutorial review

Hee-Jo Nam[1], Ryota Yamada[2], Hyun-Seok Park[1,3*]

[1]Bioinformatics Laboratory, ELTEC College of Engineering, Ewha Womans University, Seoul 03760, Korea
[2]Fuku Corporation, Tokyo 113-0033, Japan
[3]Center for Convergence Research of Advanced Technologies, Ewha Womans University, Seoul 03760, Korea

The prototype version of the full-text corpus of *Genomics & Informatics* has recently been archived in a GitHub repository. The full-text publications of volumes 10 through 17 are also directly downloadable from PubMed Central (PMC) as XML files. During the Biomedical Linked Annotation Hackathon 6 (BLAH6), we experimented with converting, annotating, and updating 301 PMC full-text articles of *Genomics & Informatics* using PubAnnotation, a system that provides a convenient way to add PMC publications based on PMCID. Thus, this review aims to provide a tutorial overview of practicing the iterative task of named entity recognition with the PubAnnotation/PubDictionaries/TextAE ecosystem. We also describe developing a conversion tool between the Genia tagger output and the JSON format of PubAnnotation during the hackathon.

Keywords: named entity recognition, natural language processing, text mining

Introduction

Genomics & Informatics is the official journal of the Korea Genome Organization. The prototype version of the full-text corpus of *Genomics & Informatics* (GNI version 1.0) has recently been archived in a GitHub repository [1,2]. Further preprocessing and semi-automatic editing are underway to prepare the next version of GNI. As the volume numbers of *Genomics & Informatics* are growing, we needed a persistent and sharable repository to annotate and to upload the PMC articles of *Genomics & Informatics*.

During the Biomedical Linked Annotation Hackathon 6 (BLAH6), we experimented with annotating the PMC articles of *Genomics & Informatics*, making a custom dictionary using PubDictionaries, and uploading the annotation results into PubAnnotation. PubDictionaries is a public repository of dictionaries and PubAnnotation is a public repository of text annotations; these resources are primarily developed and maintained by the Database Center for Life Science (DBCLS), Japan [3,4]. PubAnnotation and PubDictionaries adopt a dictionary-based agile text mining approach, wherein iterative development cycles can be carried out by modifying a dictionary, manually reannotating, and automatically reannotating [5].

Thus, the purpose of this interdisciplinary tutorial review is to share our experiences of using the PubAnnotation ecosystem and writing a conversion script to apply to the *Genomics & Informatics* corpus [6,7]. We provide an introductory overview to briefly intro-

duce basic information extraction tasks and dictionary-based named entity recognition (NER) for non-experts in the field, and to provide some helpful pointers to start a deeper investigation into agile text mining and corpus annotation techniques in general.

The conversion code between the Genia tagger output and JSON files and the indexed XML files for *Genomics & Informatics* during BLAH6 are both available through GitHub (https://github.com/Ewha-Bio/Genomics-Informatics-Corpus/ tree/master/code/BLAH; https://github.com/Ewha-Bio/Genomics-Informatics-Corpus/tree/master/XML).

Creating a PubAnnotation Pilot Project

PubAnnotation supports an agile approach to text mining by instantiating software components that allow for decomposed parallel development, while also facilitating continuous integration [5].

The PubAnnotation ecosystem is designed to be an open, API-driven system, and to harness changes for the user's advantage. Annotations can be obtained from an external web service, which is called an annotation server. Through this principle, potential users are able to use the system to fine-tune and adjust their existing project [8].

During BLAH6, we created a PubAnnotation pilot project, called BLAH6-GNI-Corpus (http://pubannotation.org/projects/BLAH6-GNI-Corpus), initially to upload the *Genomics & Informatics* corpus. PubAnnotation provides a convenient way to add, annotate, and edit PMC publications based on PMCID. We specified the PMCID and uploaded the text files of *Genomics & Informatics*. In total, 301 documents were imported into the project.

Three components were used to implement the iterations of agile development, as shown in Fig. 1: PubAnnotation, a storage component for regression testing; TextAE, a manual annotation tool; and PubDictionaries, a dictionary-based annotator. These three components of the PubAnnotation ecosystem provide many ways to proceed with NER projects. The following sections provide one scenario, in which we conducted agile text mining with these components by adding the PMC publications of *Genomics & Informatics* to a PubAnnotation project, writing a script to upload the existing tagged documents, creating a PubDictionaries project to obtain annotations, and editing the annotations manually with TextAE [5].

A Tutorial Example

We initially used the GENIA tagger to annotate biological terms when developing the GNI corpus 1.0 [9,10]. It is easiest to under-

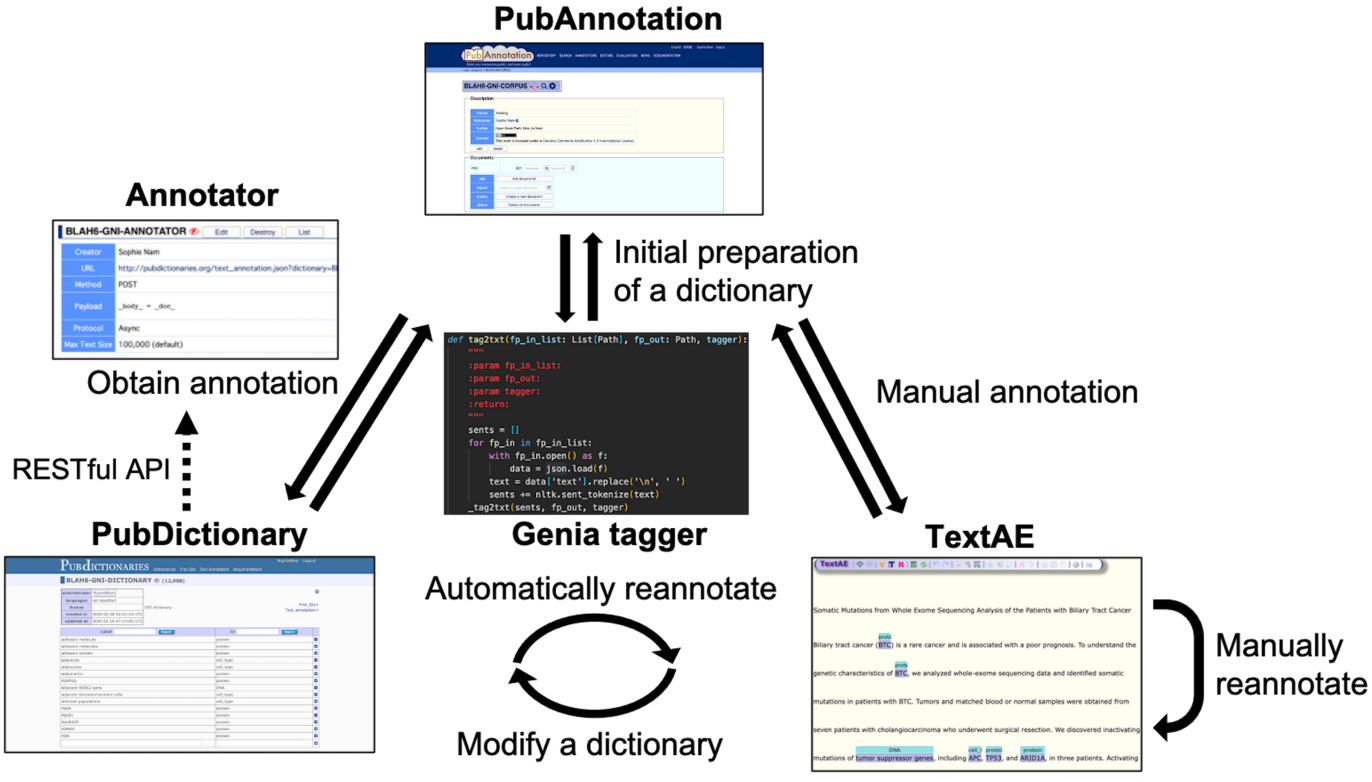

Fig. 1. An Agile approach to text mining with PubAnnotation, PubDictionaries, and TextAE.

stand how PubAnnotation and PubDictionaries might be used to integrate the GNI corpus 1.0 on the basis of an example.

An exemplary output format of the GENIA tagger

The annotation result of PMCID 6440663—"We discovered inactivating mutations of tumor suppressor genes, including APC, TP53, and ARID1A, in three patients."—as shown in Fig. 2, is used as an example sentence.

The GENIA tagger outputs the base forms, part-of-speech (POS) tags, chunk tags, and named entity tags. The tagger is specifically tuned for biomedical texts such as MEDLINE abstracts. Fig. 2A is a direct output from the GENIA tagger, and 2B is a visualization of NER generated by TextAE [10], the default viewer and editor of PubAnnotation. Four different levels of tags are attached for each word in the example sentence: base forms, POS tags, chunk tags, and named-entity tags. For example, "TP53", "TP53", "NN", "B-NP", and "B-protein" indicate that the part of speech of the word "TP53" is a noun ('NN'), that the word begins a noun phrase ('B-NP'), and that it begins a phrase of a protein name ("B-protein").

The last tag is a semantic-level tag to classify named entities in the text into pre-defined categories such as proteins, DNAs, RNAs, cell lines, and cell types. For named-entity tags, B/I/O notation was used, wherein the B/I/O terminology refers to the beginning of the phrase (B), internal to the phrase (I), and outside of the phrase (O).

Fig. 2 shows that APC was wrongly classified, because "APC" could refer to the adenomatous polyposis coli gene or to an antigen-presenting cell. Generally, biomedical NER faces difficulties for many reasons, prominent among which are the often-ambiguous abbreviations that are frequently used in the biomedical field.

Wring a Python script to convert GENIA tagging results into a PubAnnotation format

The desired result of a dictionary-based text annotation task would be an index of the dictionary entities corresponding to the referenced target texts. PubAnnotation's text sequencer turns a document into a sequence of characters, so that positions in the document can be specified unambiguously by character offsets. For this reason, a conversion tool between the Genia tagger output and the JSON import/export format of PubAnnotation was written in Python during BLAH6. As shown in Fig. 3, we extracted the text field from a JSON file, and tokenized it by sentence, using the Natural Language Toolkit (NLTK) package, as in lines 67–73 [11]. This tokenizer divides a text into a list of sentences by using an unsupervised algorithm to build a model for abbreviation words, colloca-

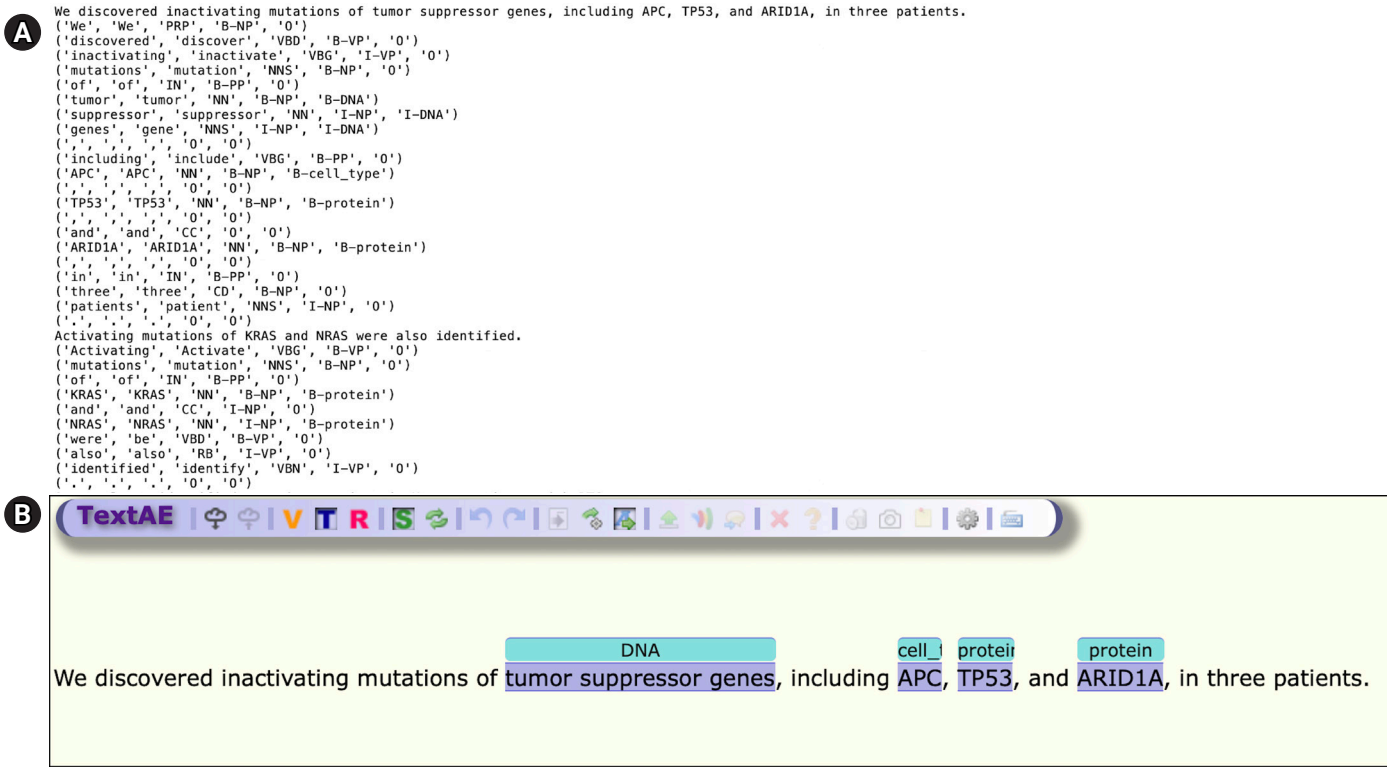

Fig. 2. (A) Initial NER result by GENIA tagger. (B) A visualization of NER generated by TextAE.

```python
def _tag2json(sents, fp_in, fp_json, tagger):
    _, pmcid, divid, sec = fp_in.stem.split('-', maxsplit=3)
    data = {
        'text': None,
        'sourcedb': 'PMC',
        'sourceid': pmcid,
        'divid': divid,
        'denotations': []
    }

    # tagging by GENIA Tagger
    anns = []
    begin, end = 0, 0
    for sent in sents:
        for word, word2, _, _, tag in tagger.parse(sent):
            end = begin + len(word)
            ann = [
                begin,
                end,
                tag,
                word,
            ]
            begin = end + 1
            anns.append(ann)
        begin = end + 1
    text = ' '.join([w for _, _, _, w in anns])
    data['text'] = text

    # combine B-I terms
    reversed_anns = []
    is_continue = False
    _end = None
    word_list = []
    for begin, end, tag, word in reversed(anns):
        if tag.startswith('O'):
            continue
        word_list.append(word)
        if tag.startswith('I') and not is_continue:
            is_continue = True
            _end = end
        elif tag.startswith('B'):
            if _end is None:
                _end = end
            ner_tag = tag.split('-', maxsplit=1)[-1]
            word = ' '.join(reversed(word_list))
            reversed_anns.append([begin, _end, ner_tag, word])
            is_continue = False
            _end = None
            word_list = []

    # append annotaions to data
    _id = 1
    for begin, end, tag, word in reversed(reversed_anns):
        data['denotations'].append({
            'id': 'T{}'.format(_id),
            'span': {'begin': begin, 'end': end},
            'obj': tag,
            'text': text[begin:end],
        })
        _id += 1

    fp_out = fp_json.parent / '{}-{}.json'.format(fp_json.stem, divid)
    with fp_out.open(mode='w') as f:
        json.dump(data, f, indent=4)


def tag2json(fp_in_list: List[Path], fp_out: Path, tagger):
    for fp_in in fp_in_list:
        with fp_in.open() as f:
            data = json.load(f)
        text = data['text'].replace('\n', ' ')
        _sents = nltk.sent_tokenize(text)
        _tag2json(_sents, fp_in, fp_out, tagger)


if __name__ == '__main__':
    parser = argparse.ArgumentParser()
    parser.add_argument('dir_in', type=str)
    parser.add_argument('dir_out', type=str)
    args = parser.parse_args()
    ext = args.ext
    dir_in = Path(args.dir_in)
    dir_out = Path(args.dir_out)
    if not dir_out.exists():
        dir_out.mkdir()

    tagger = load_tagger(GENIA_FP)
    binded_fps = get_binded_fps(dir_in)

    for pmcid, fp_in_list in tqdm(binded_fps.items()):
        fp_out = dir_out / 'tagged_{}.{}'.format(pmcid, ext)
        tag2json(fp_in_list, fp_out, tagger)
```

Fig. 3. A conversion tool written in Python.


```
{"target":"http://pubannotation.org/docs/sourcedb/PMC/sourceid/6440663/divs/0",
  "sourcedb":"PMC","sourceid":"6440663",
  "text":"We discovered inactivating mutations of tumor suppressor genes,
  including APC, TP53, and ARID1A, in three patients.",
  "divid":0,"project":"BLAH6-GNI-CORPUS",
  "denotations":[
    {"id":"T1","span":{"begin":40,"end":62},"obj":"DNA"},
    {"id":"T2","span":{"begin":74,"end":77},"obj":"cell_type"},
    {"id":"T3","span":{"begin":79,"end":83},"obj":"protein"},
    {"id":"T4","span":{"begin":89,"end":95},"obj":"protein"}
    ]
  }
```


Fig. 4. PubAnnotation JSON format with its tag and span indexing information.

https://doi.org/10.5808/GI.2020.18.2.e13

```
{"target":"http://pubannotation.org/docs/sourcedb/PMC/sourceid/6440663/divs/0",
  "sourcedb":"PMC","sourceid":"6440663",
  "text":"We discovered inactivating mutations of tumor suppressor genes,
  including APC, TP53, and ARID1A, in three patients.",
  "divid":0,"project":"BLAH6-GNI-CORPUS2",
  "denotations":[
    {"id":"BLAH6-GNI-ANNOTATOR_T1","span":{"begin":40,"end":62},"obj":"gene"},
    {"id":"BLAH6-GNI-ANNOTATOR_T2","span":{"begin":74,"end":77},"obj":"gene"},
    {"id":"BLAH6-GNI-ANNOTATOR_T3","span":{"begin":79,"end":83},"obj":"gene"},
    {"id":"BLAH6-GNI-ANNOTATOR_T4","span":{"begin":89,"end":95},"obj":"gene"}
    ],
    "relations":[
      {"id":"R1","pred":"contains","subj":"BLAH6-GNI-ANNOTATOR_T1","obj":"BLAH6-GNI-ANNOTATOR_T2"},
      {"id":"R2","pred":"contains","subj":"BLAH6-GNI-ANNOTATOR_T1","obj":"BLAH6-GNI-ANNOTATOR_T3"},
      {"id":"R3","pred":"contains","subj":"BLAH6-GNI-ANNOTATOR_T1","obj":"BLAH6-GNI-ANNOTATOR_T4"}
      ]
    }
```


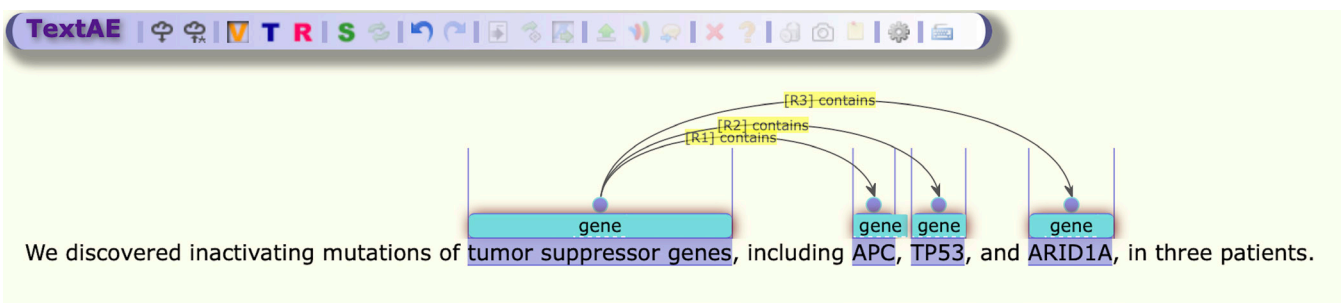

Fig. 5. (A) JSON format of the example sentence. (B) A visualization of named entity recognition generated by TextAE.

tions, and words that start sentences. In line 12–27, indexes for each word in the Genia output are calculated by adding up the white spaces and character lengths. In lines 29–60, a new list is created, containing the begin index, end index, named entity tag, and word; B-tags and I-tags are combined, and the indexes are recalculated. Finally, the list to the denotation field of the dictionary is appended and converted into JSON.

Manual editing using TextAE

A facility of visualization and manual editing is one of the primary aspects of making the PubAnnotation ecosystem adoptable by end-users. A user can easily add a new entry or delete an entry, in a try-and-revise manner.

In Fig. 4, there are four denotations for our example sentence, T1 through T4, with its tag and span information. The first one connects span 40–62 (the text spanning from the 40th to 62nd characters) to DNA, while the fourth connects span 89–95 to Protein. The default interpretation of T4 is as follows: the text span between *"span":{"begin":89, "end":95}* denotes an entity *T1 "id":"T1"* of which the type is *Protein*.

Once an annotation file is prepared, TextAE can be used for manual editing of NER [12], as in Fig. 5. TextAE is a web-based graphical annotation editor, which was developed as an open-source project. APC is now tagged as a "gene," as shown in Fig. 5, after manual editing. In addition to NER tagging, the example also presents the ease of using TextAE for manual editing of relation annotations, showing that the two entities, T1 and T2, that are introduced by the two denotations, are related to each other by the predicate "contains", specified by the two different keys, so the relationship is directional.

Summary

In this tutorial review, we presented our experiences of conducting and agile text mining. During BLAH6, we created two separate PubAnnotation projects (BLAH6-GNI-Corpus and BLAH6-GNI-Corpus2), a dictionary (BLAH6-GNI-Dictionary), and an annotator (BLAH6-GNI-Annotator). A total of 12,908 labels were registered in the PubAnnotation ecosystem (http:// pubannotation.org/annotators/BLAH6-GNI-ANNOTATOR).

While developing a conversion tool during BLAH6, indexing and calculating spans was a non-trivial task, as PubAnnotation uti-

lizes character-based indexing; enforcement of a fixed tokenization of the text is technically expensive.

Some minor suggestions relate to the user interface. In some menus, it was not fully obvious for a first-time user of the system what was clickable. We also had to create two separate projects, simply to utilize PubAnnotation's text sequencer.

We assume that there are many categories of users with different levels of experience and familiarity with PubAnnotation, ranging from pure natural language processing specialists to biomedical research end users. We hope that some additional features will be added to the PubAnnotation ecosystem, to provide diversified access to different groups of users, who have different needs regarding workflow and information density.

ORCID

Hee-Jo Nam: https://orcid.org/0000-0001-6184-6737
Ryota Yamada: https://orcid.org/0000-0003-2237-5025
Hyun-Seok Park: https://orcid.org/0000-0002-6617-2740

Authors' Contribution

Conceptualization: HSP. Data curation: HJN. Methodology: HJN, RY. Writing – original draft: HSP.

Conflicts of Interest

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

Acknowledgments

This work was supported by a National Research Foundation of Korea grant (NRF-2019R1F1A1058858) funded by the Korean government (MSIT).

References

1. Genomics and Informatics archives. Seoul: Korea Genome Organization, 2018. Accessed 2020 Jun 17. Available from: https://genominfo.org/articles/archive.php.
2. Oh SY, Kim JH, Kim SJ, Nam HJ, Park HS. GNI Corpus Version 1.0: annotated full-text corpus of Genomics & Informatics to support biomedical information extraction. Genomics Inform 2018;16:75-77.
3. Kim JD, Wang Y. PubAnnotation: a persistent and sharable corpus and annotation repository. In: BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (Cohen KB, Demner-Fushman D, Ananiadou S, Webber B, Tsukii J, Pestian J, eds.), 2012 Jun 8, Montreal, Canada. Stroudsburg: Association for Computational Linguistics, 2012. pp. 202-205.
4. Kim JD, Cohen KB, Kim JJ. PubAnnotation-query: a search tool for corpora with multi-layers of annotation. BMC Proc 2015;9:A3.
5. Kim JD, Wang Y, Fujiwara T, Okuda S, Callahan T, Cohen KB. Open Agile text mining for bioinformatics: the PubAnnotation ecosystem. Bioinformatics 2019;35:4372-4380.
6. Chinchor N, Robinson P. MUC-7 named entity task definition. In: Proceedings of the 7th Conference on Message Understanding, 1997 Sep 17, Fairfax, VA, USA. pp. 1-21.
7. Song HJ, Jo BC, Park CY, Kim JD, Kim YS. Comparison of named entity recognition methodologies in biomedical documents. Biomed Eng Online 2018;17:158.
8. Beck K, Grenning J, Martin RC, Beedle M, Highsmith J, Mellor S, et al. Manifesto for agile software development. The Author, 2001. Accessed 2020 Jun 17. Available from: http://agilemanifesto.org.
9. Tsuruoka Y, Tateishi Y, Kim JD, Ohta T, McNaught J, Ananiadou S, et al. Developing a robust part-of-speech tagger for biomedical text. In: Advances in Informatics. PCI 2005. Lecture Notes in Computer Science, Vol. 3746 (Bozanis P, Houstis EN, eds.). Berlin: Springer, 2005. pp. 382-392.
10. Tsuruoka Y. GENIA tagger. Tokyo: The Author, 2010. Accessed 2020 Jun 17. Available from: http://www.nactem.ac.uk/GENIA/tagger.
11. Loper E, Bird S. NLTK: the natural language toolkit. Preprint at https://arxiv.org/abs/cs/0205028 (2002).
12. Kim JD, Wang Y, Nakajima S. TextAE. The Author, 2015. Accessed 2020 Jun 17. Available from: http://textae.pubannotation.org/.

https://doi.org/10.5808/GI.2020.18.2.e13

G&I
Genomics & Informatics

Original article

eISSN 2234-0742
Genomics Inform 2020;18(2):e14
https://doi.org/10.5808/GI.2020.18.2.e14

Received: March 27, 2020
Revised: June 9, 2020
Accepted: June 10, 2020

*Corresponding author:
claire.nedellec@inrae.fr

WTO, an ontology for wheat traits and phenotypes in scientific publications

Claire Nédellec[1*], Liliana Ibanescu[2], Robert Bossy[1], Pierre Sourdille[3]

[1]Paris-Saclay University, INRAE, MaIAGE, F-78350 Jouy-en-Josas, France
[2]Paris-Saclay University, INRAE, UMR MIA-Paris, AgroParisTech, F-75005, Paris, France
[3]University Clermont-Auvergne, INRAE, UMR 1095 GDEC, F-63000 Clermont-Ferrand, France

Phenotyping is a major issue for wheat agriculture to meet the challenges of adaptation of wheat varieties to climate change and chemical input reduction in crop. The need to improve the reuse of observations and experimental data has led to the creation of reference ontologies to standardize descriptions of phenotypes and to facilitate their comparison. The scientific literature is largely under-exploited, although extremely rich in phenotype descriptions associated with cultivars and genetic information. In this paper we propose the Wheat Trait Ontology (WTO) that is suitable for the extraction and management of scientific information from scientific papers, and its combination with data from genomic and experimental databases. We describe the principles of WTO construction and show examples of WTO use for the extraction and management of phenotype descriptions obtained from scientific documents.

Keywords: ontology, text mining, wheat trait and phenotype
Availability: WTO ontology (https://doi.org/10.15454/1.4382637738008071E12) is available on AgroPortal: http:// agroportal.lirmm.fr/ontologies/WHEATPHENOTYPE under the license Creative Commons Attribution International 4.0 International (CC BY 4.0); The wheat trait bibliographic search engine SAMBlé AlvisIR is available at: http:// bibliome.jouy.inra.fr/demo/wheat/alvisir/webapi/search.

Introduction

Improvement of most animal and plant species of agronomical interest has become an international stake because of the increasing demand for feeding a growing world population. The new environmental constraints such as the reduction of inputs (water, fertilizers, and pesticides) and the reduction of acreages involve the development of new breeding schemes that must be shorter and more powerful. This requires a significant improvement of the agronomical potential of the species through breeding. This is especially true for bread wheat (*Triticum aestivum* L.) which is the most widely grown crop worldwide.

The recent advent of genomic tools contributed to a better understanding of the biological mechanisms underlying the expression of phenotypes of agronomical interest. The availability of genetic information linked to genotyping and phenotyping experimental data obtained from fields and controlled environments has never been greater for understanding biological mechanisms and hypothesizing new models of plant biology [1].

As a consequence, reusing data from different platforms that are obtained through different methods, sensors and protocols, has become a major challenge. The standardization of the information for semantic interoperability of heterogeneous datasets is a key

milestone [2]. An ontologie, as defined in [3], is designed to represent the knowledge from one domain by concepts (or classes), relationships among these concepts and instances of these concepts. Therefore ontologies have long been identified as a critical tool for managing information systems in the fields of integrative plant biology, genetics and phenomics [4]: among others Gene Ontology [5,6] defines gene functions, biological processes and cellular components ; the Plant Ontology (PO) Database [7,8] developed by the Planteome Project is a community resource for plant structure and developmental stages controlled vocabulary and annotations [9]. PO links plant anatomy, morphology and growth and development to plant genomics data.

Dedicated ontologies focus on controlled vocabulary for the description of the phenotypic information. The Plant Trait Ontology (TO) [10] of the Planteome project [11] defines general phenotypic traits in plants. Each trait is a distinguishable feature, characteristic, quality, or phenotypic feature of a developing or mature plant independently of the species. The Crop Ontology (CO) [12,13] is developed by several centers of the Consultative Group on International Agricultural Research (CGIAR) Biodiversity and their partners (Elixir, INRAE, iBET). This ontology focuses on the documentation of phenotype observations as variables that are grouped in nine high-level trait classes. The variables are triplets of observation methods, units of measurement, and traits that encompass the observed entity (e.g., grain, plant). CO distinguishes specific traits for 31 economically important plant species. Their vocabularies have reached different stages of development, ranging from pearl millet (52 variables) to wheat, the richest, with 498 variables.

Beside observation and experimental data, scientific literature is a significant source of genetic and phenotypic information on plants [14,15]. Automatic information retrieval and information extraction have been acknowledged as major challenges in Life Science for assisting manual biocuration, either to assess experimental or inferred data quality or to fill databases with complementary information [16,17]. However, most work focuses on molecular biology, functional and comparative genomics resource development, and phenotypic-related human health, as the Biocreative Track III interactive text mining task in 2012 [18].

In the plant biology domain, information extraction from text has attracted less attention [19], even though the quality and the abstraction of the textual information confer it a significant value for breeding. General properties of plant cultivars as described in the literature are of great interest for many research and innovation studies that are complementary to the detailed and partially unrelated phenotypic observations. Scientific literature summarizes, synthetizes, abstracts and explains experimental results, filtering out spurious observations and highlighting important outcomes. As such it constitutes a valuable source of knowledge for the interpretation of phenotyping experimental results, as well as for the design of plant system biology models able to explain, predict, or simulate genotypic-phenotypic relationships.

Information extraction from text requires the establishment of dedicated ontologies and of text mining pipelines as largely recognized in the biomedical domain [20]. Ontologies improve text mining performances and conversely the information extracted is more reusable when linked to a reference resource such as an ontology through the normalization process. Normalization consists in assigning a class or a category from a controlled vocabulary to text mentions. It is a key step for the semantic interoperability of textual information and other sources of data and a major text mining challenge [21]. Plant traits and phenotypes expressed in textual sources are characterized by a great variability of the lexicon [15]. The text carries information at various levels of generality with different assessment status, ranging from experimental fine-grained data to general expert knowledge, through intermediate levels of synthesis and abstraction. The examples in Fig. 1 illustrate the variability of trait expressions in scientific documents in descending order of generality. In example (1) the trait "resistance to fungal and viral diseases" is a general trait. Example (2) mentions "FHB resistance" (i.e., Fusarium head blight resistance) which reflects resistance to a specific fungal disease FHB, and its effect on the related observations of six specific traits (e.g., plant height). Example (3) is the most specific: it is about the severity score of the trait "Russian wheat aphid resistance" observed for a given cultivar (i.e., Hatcher) whose value is 1.9.

This varying scope of phenotypic information in scientific papers answers to different needs and usages. It ranges from detailed documentation of experiments and inferred data, to review of shared and well-acknowledged bodies of knowledge supported by large sets of experimental and scientific results.

We have been developing the Wheat Trait Ontology (WTO) since 2010 to answer breeders and scientists' needs for wheat trait and phenotype information management and retrieval at varying abstraction scales. WTO supports two objectives: (1) building a formal shared representation of wheat trait whose knowledge organization closely reflects the expert knowledge model and (2) making phenotypic information extraction from text easier. To achieve both objectives, the sources for building WTO include expert knowledge and textual documents: expert interviews, terminology analysis from the literature and gene catalogs. The richness of the WTO vocabulary, its similarity with scientific literature lexicon and its deep hierarchies make it a useful resource for both text mining and information management.

a. Thinopyrum ponticum *and* Th. intermedium*: the promising source of* ***resistance to fungal and viral diseases*** *of wheat*

b. ***FHB resistance*** *QTL alleles from Nyuubai, Sumai-3, and Wuhan-1 were evaluated for their effect on* ***Fusarium head blight (FHB) index***, ***Fusarium damaged kernels (FDK), deoxynivalenol (DON) accumulation, plant height, anthesis date,*** *and numerous* ***grain quality traits***

c. ***Russian wheat aphid resistance*** *scores for Hatcher (****1.9 score****, 1 = very resistant to 5 = very susceptible, n = 12 observations) in standard greenhouse seedling screening tests*

Fig. 1. Examples of phenotype mentions from scientific papers. Traits and phenotypes are in bold.

The paper is organized as follows. Section 2 describes the WTO. Section 3 presents the motivation and method for building WTO. Section 4 illustrates WTO usage through an application. Section 5 discusses WTO characteristics compared to other semantic resources and presents future work.

WTO Description

The WTO covers a wide range of bread wheat traits (e.g., observable physical plant properties), phenotypes (e.g., trait values) and their related environmental conditions (e.g., disease, extreme temperature) organized in three trees. The current version contains 596 classes. The population of the main classes and their subclasses is given in Table 1.

The maximum depth of WTO is 9 and the average number of children per class is 3. We chose a deep and balanced structure because the breeder's needs to manage data at different levels of aggregation. Classes at intermediate levels support synthetic queries for searching high-level correlations between genetic, phenotypic, and physiological phenomena.

The classes of the 'Trait' subtree are linked to the corresponding phenotypes by the "Trait_has_value" relationship. For instance, 'ear emergence time' trait class is linked to the 'late heading' phenotype class.

The 'Environmental condition' subtree mainly represents abiotic conditions and biotic conditions that are linked to the corresponding responses of the plant to abiotic stresses and biotic stresses. The main root classes of 'Response to environmental conditions' range from response to chemical, radiation, temperature, to a large range of responses to biotic stresses as shown in Fig. 2.

Response to biotic stress is indeed a major concern for wheat breeding. Wheat is affected by several microbial, bacterial, viral and mainly fungal diseases that cause major crop loss [22]. WTO accounts for this situation with two large subtrees 'Disease' (58 classes) and 'Pest' (103 classes) of 'Environmental condition' (Table 1). The relation 'Causes' between the 'Pathogen' classes and the infectious "Disease" classes represents the causative link between the agent and the disease. WTO distinguishes between disease of bacterial, viral and fungal causal agents. A total of 55 different fungal species causing 44 diseases is described in WTO.

Table 1. Main classes of WTO with the number of some subclasses

	No.
Environmental condition	221
Abiotic condition (e.g., chemical, nutrient, water, wind)	51
Biotic condition	171
Biotic stress	170
Disease	58
Bacterial disease	6
Fungal disease	44
Viral disease	6
Diseased caused by nematode	2
Pest	103
Insects	21
Plant property	374
Phenotype	45
Trait	326
Development (plant habit, precocity, vernalization)	19
Growth (crop yield, nutrient use efficiency, density)	41
Morphology (of awn, glume, grain, spike)	23
Quality	58
Food property	30
Grain composition	12
Grain quality	13
Milling quality	4
Reproduction	5,173
Response to environmental conditions	64
Response to abiotic stress	104
Response to biotic stress	

WTO, Wheat Trait Ontology.

In a similar way, the 'Response to biotic stress' subtree finely distinguishes between the causal stress factors as shown in Figs. 3 and 4.

The responses of the plant to biotic stresses are expressed in two ways: either by the disease name or by the causative agent names. A given disease name may have synonyms and a disease may be caused by more than an agent (Fig. 5).

Moreover, fungi naming in scientific paper do not always strictly follows the nomenclature standard imposed by the mycologist community. For instance, names corresponding to different life stages can be found. For each resistance trait, the causative agents are given with their standard names and the other names as in Fig. 6.

WTO lexical synonymy relations and conceptual relations are complementary with respect to the intended uses to reflect expert knowledge model and make phenotypic information extraction from text easier.

As summarized by Fig. 7 WTO structure is mainly hierarchical with two transversal relations: a domain-specific causal one and a variable-value relation.

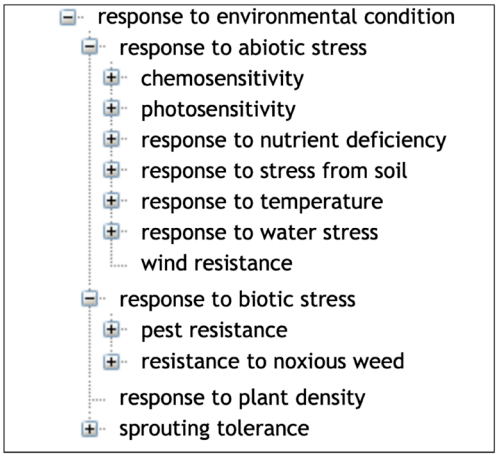

Fig. 2. Wheat Trait Ontology (WTO) subclasses for 'response to environmental condition'.

Wheat Trait Ontology Building

WTO was built using the NeOn Methodology [23], a scenario-based methodology that supports the collaborative aspects of ontology development and reuse. The WTO development process followed successively Scenario 1, From specification to implementation, and Scenario 2, Reusing non-ontological resources of the NeOn methodology. The first step is to specify the ontology requirements, provided in the next subsection. Then we present insights and rationales for design choices in the following subsection.

Ontology Requirements Specification

The needs for the development of shorter and more powerful breeding schemes is a strong motivation for sharing phenotypic information linked to genes of interest and traits. Building an open and shared database for marker-based assisted selection (MAS) in bread wheat was the SAMblé project objective (2010–2014) [24]. The SAMBlé database should support both short-term MAS-related goals of breeders, the intended users, and long-term research goals of researchers on underlying biological mechanisms of phenotypes. The information considered for the database was the existence of links between one or more markers and genes of agronomic interest in bread wheat. The information sources were the scientific literature, gene catalogs and in-field and high-throughput phenotyping experiments.

In scientific papers, phenotypic information is frequently linked to varieties, genes or markers and traits as in Fig. 8, which makes it extremely relevant for breeding [15].

This information was first automatically extracted from the literature, then assessed against reference material and elite material (335 varieties) under field conditions for different traits of interest. Finally, the markers that gave the best results and could be used in breeding selection schemes were recorded in the database to be queried by the partner breeders [24]. The traits considered in the SAMBlé project were related to four main large topics, namely, disease resistance, resistance to abiotic stress, plant development, and baking quality.

Representative queries of the breeders were, "which alleles and markers are involved in resistance to rust (e.g., leaf rust, stripe rust, stem rust)" and "what are the varieties tested." Same question arises for "bread making quality (e.g., flour quality, color, composition, mechanical property, crumb firmness)?".

A general objective of SAMBlé was to develop a shared database with the information collected by the project that would be easily searchable. The WTO was designed to support this goal. The ontology should support queries on traits and phenotypes at various levels of aggregation combinied with other criteria on markers, genes, and varieties.

To this purpose we created WTO as deep non-strict hierarchies of traits, phenotypes, and environment factors. Non-strict means here that one concept may have several direct parents forming a direct acyclic oriented graph. It covers the large set of topics of the SAMBlé database, ranging from development or resistance to stress to food quality.

Design and implementation

The design of WTO followed a top-down approach where the core model was first established based on project partner expertise on wheat phenotyping: the SAMblé project gathered breeders from French breeding companies, the French union of breeders (UFS) and Arvalis, it was led by the research unit GDEC-INRAE (Genetics, Diversity and Ecophysiology of Cereals). Text mining and plant information management were provided by Mathematics, Informatics and Genomics Laboratory, French National Research Institute for Agriculture, Food and the Environment (MIG-INRA) and Unité de Recherche Génomique Info, French National Research Institute for Agriculture, Food and the Environment (URGI-INRAE). The main classes of the WTO core model were similar as presented in Table 1 of Section 2. The core model was then extended by reusing information from three external sources: scientific literature, the Catalog of Gene Symbols for Wheat [25] and GrainGene database [26]. The biotic stress response, diseases, and pathogen WTO subtrees (see Section 2) were then significantly restructured by wheat disease experts. We adopted the Obo-Edit tool as ontology editor, to make it easier for biologists and breeders to revise and enrich WTO, compared to more powerful but less user friendly tools.

https://doi.org/10.5808/GI.2020.18.2.e14

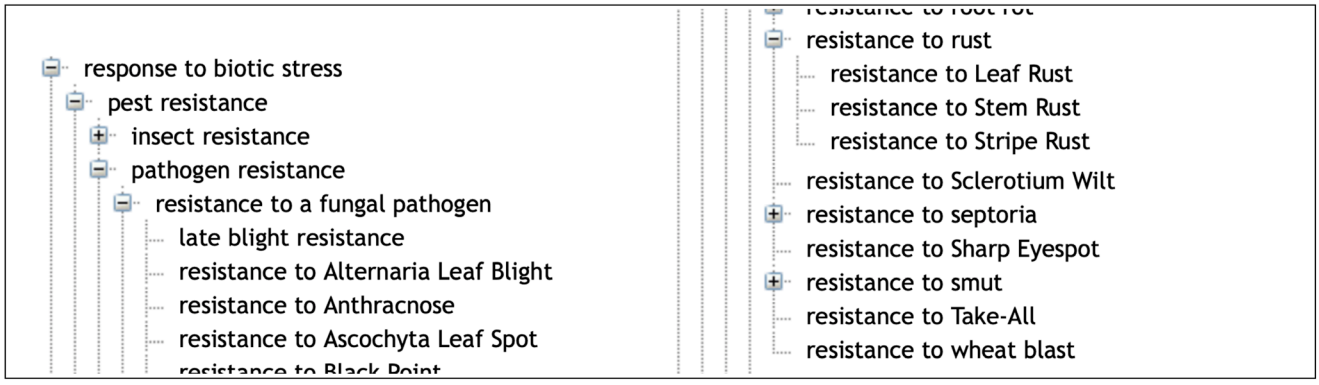

Fig. 3. An excerpt of the different 'resistance to a fungal pathogen' in Wheat Trait Ontology (WTO).

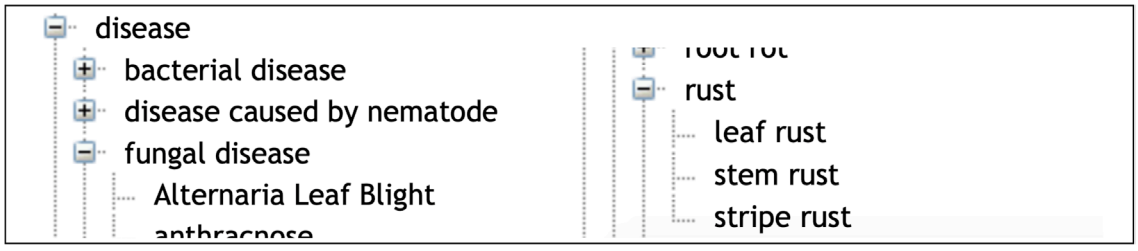

Fig. 4. Example of the rust disease family in Wheat Trait Ontology (WTO).

Preferred Name	resistance to Leaf Rust
Synonyms	resistance to brown rust resistance to Puccinia recondita leaf rust resistance resistance to Puccinia triticina

Fig. 5. Synonyms of the 'Resistance to Leaf Rust' label in Wheat Trait Ontology (WTO). Leaf Rust disease is caused by different fungi, namely 'Puccinia recondita' and 'Puccinia tricina'.

Preferred Name		Parastagonospora nodorum
Synonyms	Fungi	Septoria nodorum Leptosphaeria nodorum
Preferred Name	Disease	Septoria nodorum blotch
causes		Parastagonospora nodorum

Plant response

Preferred Name	resistance to Stagonospora nodorum
Synonyms	resistance to Septoria nodorum blotch resistance to Leptosphaeria nodorum resistance to Septoria nodorum

Fig. 6. Example of various names of *Parastagonospora nodorum* fungus in Wheat Trait Ontology (WTO).

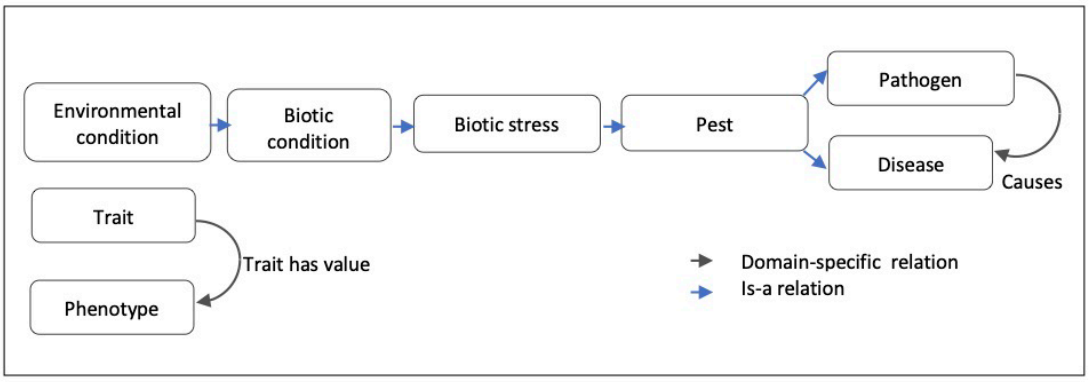

Fig. 7. Wheat Trait Ontology (WTO) relations.

*The **powdery mildew resistance** in **Suwon 92** is most likely conditioned by the **Pm3** locus [..] located on the short **arm of chromosome 1A** [..] The gene markers developed herein can be directly used for MAS of some of the Pm3 alleles in breeding programs. [PMID 18944309].*

Fig. 8. Example of marker in a cultivar related to disease resistance phenotype.

Scientific literature as a source of concepts

An Ontology Acquisition approach was first used to extract and conceptualize WTO concepts and relationships from scientific text expressed in natural language, following the same methodology as described in Nedellec et al. study [27].

We applied the term extractor BioYateA [28] to a scientific corpus to automatically extract relevant domain-specific terms. BioYatea's strength over other term extractors is the ability to extract prepositional phrases that are frequent in wheat trait terms, e.g., response *to* vernalization, florets *without* grain [29]. The scientific corpus was composed of the abstracts and titles of articles. They were obtained from the Web of Science (WoS) bibliographical search engine with the keywords '*wheat* or Triticum aestivum and *marker* and *gene*'. It yielded 3,170 references (see Nedellec et al. study [15] for more details).

The candidate terms extracted by BioYateA were then used to derive concepts using the Terminology Design Interface (TyDI) tool. TyDI supports term collaborative assessment and structuring [27]. First, relevant terms were selected among candidate terms by manual screening. Validated terms were grouped in semantic classes of preferred terms, synonyms and typographic and acronym variations. They were structured in hypernym hierarchies consistent with the core model. Concepts and concept hierarchies were then derived from these semantic classes and hypernym trees to populate the core model. The preferred terms were kept as concept labels. This literature term analysis approach sped-up the discovery process of a very large set of trait, phenotype, disease, and pathogen related concepts and subsumption relationships.

Other external sources of wheat trait terms

To identify complementary relevant trait terms, we also used the Catalog of Gene Symbols for Wheat (WGC) [30] available online at the Wheat Genetics Resources Database of Japan as a PDF file at the date of WTO building in 2011. The main contribution to WTO from the catalog was related to plant morphology (e.g., plant height) and physiology (e.g., response to photoperiod).

The GrainGenes [26] database was also used for the study of biotic stress response. GrainGenes is a comprehensive resource for molecular and phenotypic information for wheat maintained by U.S. Department of Agriculture and mirrored by MaIAGE. GrainGenes web pages listed general traits and specific traits for wheat, barley, and oat species from which we identified some wheat disease names and their pathogen agents. INRAE experts of wheat diseases then controlled the naming because for some diseases American vernacular naming was not consistent with European naming.

WTO evolution

Fig. 9 displays the evolution of the WTO (formerly named Wheat Phenotype Ontology) between 2010 and 2020. The first public

https://doi.org/10.5808/GI.2020.18.2.e14

version of WTO was released in August 2011. It contained 460 classes and 260 synonyms of labels. The 2011 version of WTO was revised in 2014. Confusions between pathogen names and synonyms were corrected which resulted in an increased number of synonyms and decreased number of classes. In 2015, the classes 'Fiber quality', 'Food property', 'Milling quality', 'Grain composition' and 'Grain quality' were grouped in a new 'Quality' class in order to reduce the number of root classes and to increase WTO readability. Conversely the 'Development' class that mixed phenological phenotypes, morphology (e.g., color, length) and growth (related to yield) was split into three distinct classes: 'Development', 'Morphology' and 'Growth'.

In 2017, with the purpose of using WTO for managing other phenotypic databases than the SAMBlé one, we evaluated WTO scope with respect to two external resources. The WIPO (the Wheat INRA Phenotype Ontology formerly named the INRA Wheat Ontology) [31] developed by URGI-INRAE and the "list of wheat descriptors for Characterization and Evaluation" of the NARO GeneBank project. A few more morphology terms such as 'presence of awn', 'glume pubescence', 'glume color' were then added to WTO.

WTO for marker-assisted selection

To be used as the conceptual formalization of the SAMBlé database schema, WTO was integrated into the MAS (Marker Assisted Selection) knowledge model detailed in Nedellec et al.'s study [15]. The MAS model was designed to manage the entities and relations of the SAMBlé database. It contains 8 entity types and 14 n-ary relationships for the representation of the genotypic and phenotypic information and relationships collected from the literature and experiments of the SAMBlé project. The main MAS model entities are 'Marker', 'Type', 'Allele' and 'Gene', 'Trait' and 'Phenotype' and 'Variety'. 'Type' represents the type of method used to identify the marker, e.g., amplified fragment length polymorphism, microsatellite. The main relationships are 'Marker tags Gene in Variety' between markers, genes and varieties, 'Trait has Phenotype in Variety' between traits, phenotypes and varieties and 'Gene expresses Phenotype in Variety' between genes, phenotypes, and varieties.

The connection of the MAS model to WTO is achieved through the straight forward alignment of two pairs of MAS and WTO classes: (1) the MAS 'Trait' class is aligned with WTO 'Plant property' class, the root of the trait subtree (2) the MAS 'Phenotype' class is aligned with WTO 'Phenotype' class, the root of WTO phenotype subtree. The other MAS classes (e.g., Gene, Marker, Variety) are also connected to nomenclatures and catalogs (e.g., Genes nomenclatures, Markers lists, and Variety catalogs) for data standardization. The integrated MAS and WTO model was successfully used for the management of SAMBlé database information and for information extraction from text.

Wheat Trait Ontology Usage

WTO has been validated through the use by breeders and researchers involved in the SAMBlé project of two end-user applications, the SAMBlé database interface [24] and the Wheat literature semantic search engine AlvisIR. AlvisIR supports queries on genes, varieties, markers, phenotypes and traits extracted from PubMed references. Phenotype and trait expressions in text are normalized by WTO concepts.

Fig. 10 gives an example of a semantic search for phenotypic information. The example query asks for documents where the gene 'Lr34' is mentioned in relation to the trait 'resistance to rust' in 'wheat' by combining the three keywords, 'lr34', 'resistance to rust', and 'wheat'. The first hit displays a document extract where 'adult plant stripe rust resistance' (underlined in green) is tagged by the query 'resistance to rust' keyword.

'Resistance to rust' in the user query has been interpreted by three complementary mechanisms. A text mining workflow run in batch mode has first automatically extracted all terms from the documents, among which the term 'adult plant stripe rust resistance', and automatically mapped it to the relevant WTO class 'Resistance to stripe rust'. The query interpreter executed on the fly has segmented the user query and mapped the query term 'resistance to rust' to the corresponding WTO class. The subsumption relation between the query class and the document class has then been verified. The document term is therefore validated as an instance of the query term and the document is displayed as a hit. Fig. 11 shows the corresponding subpart of WTO with the two mapped classes. A navigation tool (Fig. 11) supports the expression of the query by the user by the combination of selected classes. The users from the SAMBlé project are satisfied with the balanced and deep tree structure of WTO that makes ontology browsing and class selection much easier than a flat and large list of classes would.

High-level queries as exemplified here are powerful for combining criteria on phenotypes with other genetic or environmental information as requested by the SAMBlé project.

The online version of AlvisIR indexes PubMed abstracts. PubMed has been preferred over WoS for its Open Access license to references. Current work includes the extension of the corpus to full papers of main scientific journals. Eighteen thousand papers have been identified among which half are available through Open Access and 1,361 journals targeted. The text mining workflow

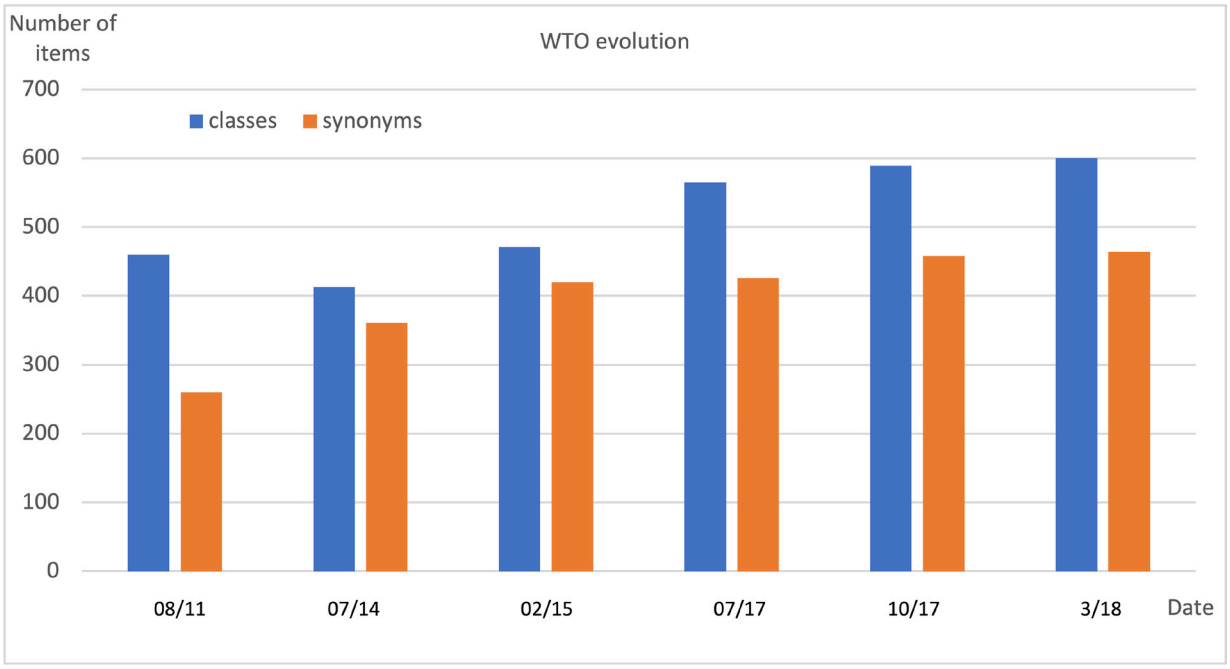

Fig. 9. Statistics of the Wheat Trait Ontology (WTO) between 2010 and 2020.

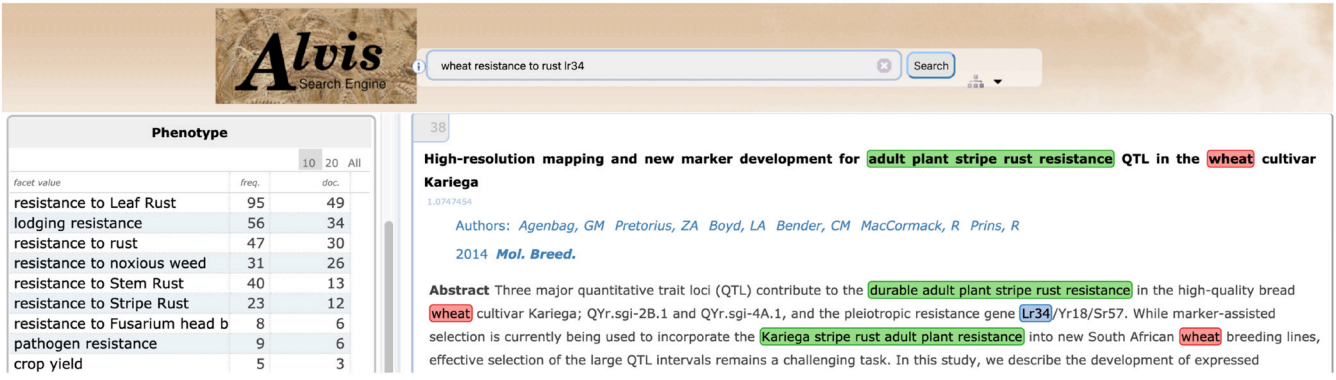

Fig. 10. Screenshot of AlvisIR semantic search engine query web page.

named WheatLiterature used to fill in the database from the scientific literature is based on the AlvisNLP technology (AlvisNLP on Github). It is distributed as a component of the European text mining OpenMinTeD platform [32].

Discussion

Beyond semantic search, ontology-based fine-grained information extraction is a key component of the integration of textual information with experimental and genetic data. However, the reference knowledge models often differ with the sources and the nature of the information. Their alignment and user query rewriting are a major challenge for data integration [33].

Significant work has been done on Wheat Data Interoperability Guidelines [34] that focuses on Minimum Information About a Plant Phenotyping Experiment (MIAPPE) [2,35]. For experimental data, observation variables including traits but also observation protocol, unit of measure and development stage are critical for properly documenting the observations and determining if observations are comparable or not. This leads to building trait ontologies as WIPO [31] or Crop Ontology [13] where the trait leaves are database variable traits (e.g., 'Susceptibility to leaf rust' in WIPO,' Leaf rust severity' in Crop Ontology). Phenotypes, the values of the traits, (e.g., '*Susceptible* to leaf rust') are not conceptualized as classes but represented by the database numerical data as values of the trait variable. For instance, in WIPO the trait Disease

https://doi.org/10.5808/GI.2020.18.2.e14

intensity score takes values on a 1 to 9 increasing scale (1, no disease; 9, very severe). Such ontologies are suitable for accurately documenting observations and for the computation of correlations by statistical tools.

Conversely, WTO aimed at managing both traits and phenotype values represented by expressions, as they occur in the scientific literature. For instance, in the *the leaf rust susceptible cultivar 'GA 100'* phrase, the phenotype value is *'leaf rust susceptible'* and the variety is *'GA 100'*. In this way, WTO representation then supports SAMblé data discovery by direct queries on traits and phenotype values (e.g., 'Leaf rust susceptibility') at various levels of generality (e.g., 'Rust susceptibility', 'Fungal disease susceptibility') and their relation to other information (e.g., cultivars).

Similar queries on observation databases that follow MIAPPE recommendations would require the translation of numerical values by using value domains or thresholds, i.e., discretization and hierarchization of the phenotypes. Moreover, the lack of depth of ontologies such as WIPO or Crop Ontology with a comb-like structure does not allow high-level queries. An example in WIPO, is the trait 'Susceptibility to leaf rust', which is a direct subclass of the high level 'biotic stress trait' without intermediate levels. Similarly, in the Crop Ontology the trait 'Fusarium head blight AUDPC' is a direct subclass of 'biotic stress trait'.

Another representative example is 'Nitrogen harvest index'. In WIPO and Crop Ontology, it has only one direct ancestor, which is 'Quality trait', with 51 other sibling traits in Crop Ontology. In WTO, 'Nitrogen harvest index' has five successive ancestors: 'Nitrogen use efficiency', 'Macronutrient use efficiency', 'Nutrient use efficiency', 'Growth' by increasing order of generality.

The integration of the two sources of data, observations, and synthetic information from text in a same data management system should preserve the best of the two approaches. It would require the alignment of the ontology classes and the rewriting of the phenotype variable values to map them to qualitative descriptors.

In the SAMBlé project and for development of the OpenMinTeD Wheat use case, we experienced this situation with the two ontologies: WIPO, which indexes experimental phenotype data, and WTO, which indexes PubMed phenotypic information. Their classes are not mappable in a straight-forward one-to-one way. It is noteworthy that the types of alignments and rewriting identified during these projects are not specific to wheat or even to plants, but are general to any phenotype observation data. Further investigation of this question is a future challenge for the integration of phenotype data from different sources allowing a better exploitation of textual data.

Conclusion

We proposed WTO, a reusable ontology of bread wheat traits and phenotypes and related environmental factors. The design of the model relies both on domain expert knowledge and the analysis of evidence published in the scientific literature. The WTO model is

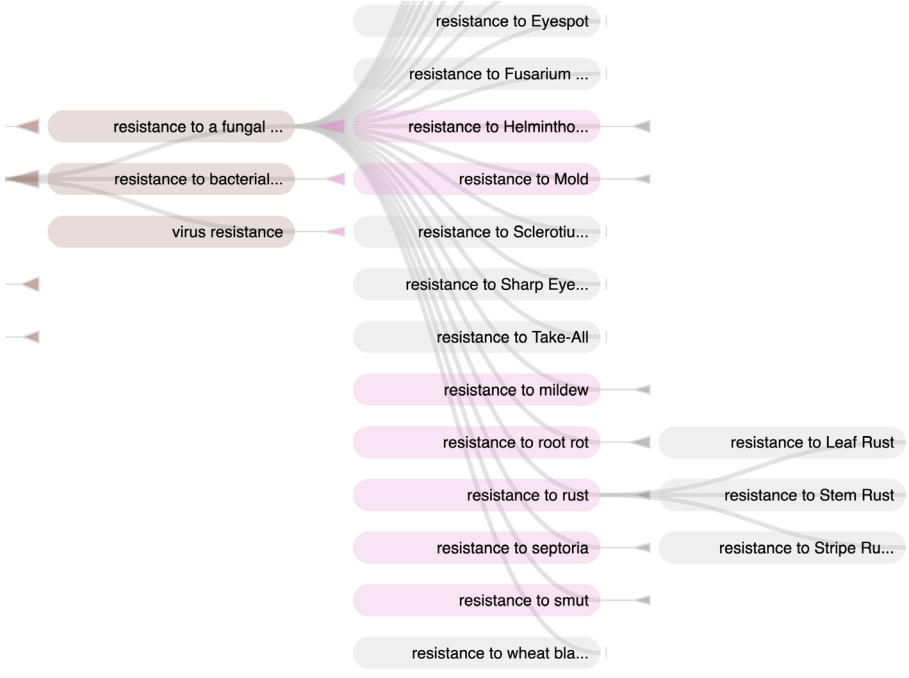

Fig. 11. Screenshot of ontology navigation in AlvisIR semantic search engine.

deeply structured, well reflecting the domain knowledge. It facilitates navigation and reuse for data and knowledge discovery. The model was designed to support the extraction and the management of marker-assisted selection information. WTO is also a contribution to the description of the link between genetic and phenotypic information. Concept synonyms were directly extracted from the literature, which turns WTO a suitable resource for Information Extraction and Information Retrieval. WTO has been assessed for its consistency through its use. WTO is complementary to other ontologies dedicated to the documentation of phenotypic observations. We believe that future work on their alignment and mapping will favor data semantic interoperability from the literature and experimental sources.

ORCID

Claire Nédellec: https://orcid.org/0000-0002-0577-0595
Liliana Ibanescu: https://orcid.org/0000-0003-3373-437X
Robert Bossy: https://orcid.org/0000-0001-6652-9319
Pierre Sourdille: https://orcid.org/0000-0002-1027-2224

Authors' Contribution

Conceptualization: CN, RB, PS. Funding acquisition: CN, PS. Methodology: CN, RB, LI. Writing – original draft: CN. Writing – review & editing: LI.

Conflicts of Interest

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

Acknowledgments

The authors would like to thank INRAE biologist researchers, especially Jacques Legouis and Thierry Marcel for their contribution to WTO extension. This work was partially funded by Oséo through the Quaero project, by FSOV through the SAM blé project. It has also received support from the wheat use case of the EU OpenMinTeD project under the H2020-EINFRA-2014-2 call, Project ID: 654021. The author work at the 6th Biomedical Linked Annotation Hackathon (BLAH6) was supported by the Database Center for Life Science (DBCLS).

References

1. Tardieu F, Cabrera-Bosquet L, Pridmore T, Bennett M. Plant phenomics, from sensors to knowledge. Curr Biol 2017;27:R770-R783.
2. Cwiek-Kupczynska H, Altmann T, Arend D, Arnaud E, Chen D, Cornut G, et al. Measures for interoperability of phenotypic data: minimum information requirements and formatting. Plant Methods 2016;12:44.
3. Guarino N, Oberle D, Staab S. What is an ontology? In: Handbook on Ontologies (Staab S, Studer R, eds.). Berlin: Springer-Verlag, 2009. pp. 1-17.
4. Jaiswal P, Ware D, Ni J, Chang K, Zhao W, Schmidt S, et al. Gramene: development and integration of trait and gene ontologies for rice. Comp Funct Genomics 2002;3:132-136.
5. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000;25:25-29.
6. Gene Ontology. Bethesda: Gene Ontology, 2020. Accessed 2020 Mar 21. Available from: http://geneontology.org/.
7. Cooper L, Jaiswal P. The plant ontology: a tool for plant genomics. In: Plant Bioinformatics: Methods and Protocols (Edwards D, ed.). New York: Humana Press, 2016. pp. 89-114.
8. Plant Ontology. Cambridgeshire: EMBL-EBI, 2020. Accessed 2020 Mar 21. Available from: https://www.ebi.ac.uk/ols/ontologies/po.
9. Avraham S, Tung CW, Ilic K, Jaiswal P, Kellogg EA, McCouch S, et al. The Plant Ontology Database: a community resource for plant structure and developmental stages controlled vocabulary and annotations. Nucleic Acids Res 2008;36:D449-D454.
10. Plant Trait Ontology. OBO Technical WG, 2020. Accessed 2020 Mar 21. Available from: http://www.obofoundry.org/ontology/to.html.
11. Cooper L, Meier A, Laporte MA, Elser JL, Mungall C, Sinn BT, et al. The Planteome database: an integrated resource for reference ontologies, plant genomics and phenomics. Nucleic Acids Res 2018;46:D1168-D1180.
12. Shrestha R, Matteis L, Skofic M, Portugal A, McLaren G, Hyman G, et al. Bridging the phenotypic and genetic data useful for integrated breeding through a data annotation using the Crop Ontology developed by the crop communities of practice. Front Physiol 2012;3:326.
13. Crop Ontology Curation Tool. Crop Ontology, 2020. Accessed 2020 Mar 21. Available from: http://cropontology.org.
14. Van Landeghem S, De Bodt S, Drebert ZJ, Inze D, Van de Peer Y. The potential of text mining in data integration and network biology for plant research: a case study on Arabidopsis. Plant Cell 2013;25:794-807.
15. Nedellec C, Bossy R, Valsamou D, Ranoux M, Golik W, Sourdille P. Information extraction from bibliography for marker assisted selection in wheat. In: Metadata and Semantics Research. MTSR 2014. Communications in Computer and Information Science,

https://doi.org/10.5808/GI.2020.18.2.e14

Vol. 478 (Closs S, Studer R, Garoufallou E, Sicilia MA, eds.). Cham: Springer, 2014. pp. 301-313.
16. Rebholz-Schuhmann D, Oellrich A, Hoehndorf R. Text-mining solutions for biomedical research: enabling integrative biology. Nat Rev Genet 2012;13:829-839.
17. Harper L, Campbell J, Cannon EKS, Jung S, Poelchau M, Walls R, et al. AgBioData consortium recommendations for sustainable genomics and genetics databases for agriculture. Database (Oxford) 2018;2018:bay088.
18. Arighi CN, Carterette B, Cohen KB, Krallinger M, Wilbur WJ, Fey P, et al. An overview of the BioCreative 2012 Workshop Track III: interactive text mining task. Database (Oxford) 2013;2013:bas056.
19. Chaix E, 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. In: Proceedings of the BioNLP Shared Task 2016 Workshop, 2016 Aug, Berlin, Germany. Stroudsburg: Association for Computational Linguistics, 2016. pp. 1-11.
20. Spasic I, Ananiadou S, McNaught J, Kumar A. Text mining and ontologies in biomedicine: making sense of raw text. Brief Bioinform 2005;6:239-251.
21. Limsopatham N, Collier N. Normalising medical concepts in social media texts by learning semantic representation. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). Stroudsburg: Association for Computational Linguistics, 2016. pp. 1014-1023.
22. Duveiller E, Singh RP, Nicol JM. The challenges of maintaining wheat productivity: pests, diseases, and potential epidemics. Euphytica 2007;157:417-430.
23. Suarez-Figueroa MC, Gomez-Perez A, Fernandez-Lopez M. The NeOn methodology for ontology engineering. In: Ontology Engineering in a Networked World (Suarez-Figueroa MC, Gomez-Perez A, Motta E, Gangemi A, eds.). Berlin: Springer, 2012. pp. 9-34.
24. Ranoux M, Nedellec C, Cariou-Pham E, Bossy R, de Vallavieille-Pope C, Leconte M, et al. Validation of markers linked to genes of interest with a view to establishing a database for assisted selection in common wheat. In: Synthèse des programmes de Recherche FSOV (Fond de Soutien à l'Obtention Végétale): Actes de la Rencontre Scientifique, 2015 Jan, Paris, France. Groupement National Interprofessionnel des Semences et Plants, 2015. pp. 1-10.
25. McIntosh RA, Dubcovsky J, Rogers WJ, Morris C, Apels R, Xia XC. Catalog of gene symbols for wheat: 2009 supplement. SHEGEN, 2009. Accessed 2020 Mar 21. Available from: https://shigen.nig.ac.jp/wheat/komugi/genes/macgene/supplement2009.pdf.
26. GrainGenes: a database for Triticeae and Avena. Albany: GrainGenes, 2020. Accessed 2020 Mar 21. Available from: https://wheat.pw.usda.gov/GG3/.
27. Nedellec C, Golik W, Aubin S, Bossy R. Building large lexicalized ontologies from text: a use case in indexing biotechnology patents. In: International Conference on Knowledge Engineering and Knowledge Management (EKAW 2010), Volume 6317 of the series Lecture Notes in Computer Science (Cimiano P, Pinto HS, eds.). Lisbon: Springer Verlag, 2010. pp. 514-523.
28. Aubin S, Hamon T. Improving term extraction with terminological resources. In: International Conference on Knowledge Engineering and Knowledge Management (EKAW 2010), Volume 6317 of the series Lecture Notes in Computer Science (Cimiano P, Pinto HS, eds.). Lisbon: Springer Verlag, 2010. pp. 514-523.
29. Golik W, Bossy R, Ratkovic Z, Nedellec C. Improving term extraction with linguistic analysis in the biomedical domain. In: Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'13), Special Issue of the journal Research in Computing Science, Vol. 70 (Gelbukh A, ed.). Zacatenco: Instituto Politecnico Nacional, Centro de Investigacion en Computacion, 2013. pp. 157-172.
30. Genebank Project, National Agriculture and Food Research Organization. Wheat descriptors (PDF) for characterization and evaluation on plant genetic resources. Tsukuba: Genetic Resources Center, 1997-2020. Accessed 2020 Mar 21. Available from: https://www.gene.affrc.go.jp/manuals-plant_characterization_en.php.
31. Wheat INRA Phenotype Ontology (WIPO). Versailles: INRA-URGI, 2020. Accessed 2020 Mar 21. Available from: https://urgi-git.versailles.inra.fr/urgi-is/ontologies/tree/develop/Wheat.
32. OpenMinTeD platform. Brussels: European Commission, 2020. Accessed 2020 Mar 21. Available from: https:// services.openminted.eu/home.
33. Ehrig M. Ontology alignment: bridging the semantic gap (Vol. 4). New York: Spring-Verlag, 2006.
34. Dzale Yeumo E, Alaux M, Arnaud E, Aubin S, Baumann U, Buche P, et al. Developing data interoperability using standards: a wheat community use case. F1000Res 2017;6:1843.
35. Krajewski P, Chen D, Cwiek H, van Dijk AD, Fiorani F, Kersey P, et al. Towards recommendations for metadata and data handling in plant phenotyping. J Exp Bot 2015;66:5417-5427.

G&I
Genomics & Informatics

Original article

eISSN 2234-0742
Genomics Inform 2020;18(2):e15
https://doi.org/10.5808/GI.2020.18.2.e15

Received: March 15, 2020
Revised: May 6, 2020
Accepted: May 22, 2020

*Corresponding author:
E-mail: jlever@stanford.edu

Extending TextAE for annotation of non-contiguous entities

Jake Lever[1*], Russ Altman[1], Jin-Dong Kim[2]

[1]Department of Bioengineering, Stanford University, Stanford, CA 94305, USA
[2]Database Center for Life Science, Research Organization of Information and Systems, Kashiwa 277-0871, Japan

Named entity recognition tools are used to identify mentions of biomedical entities in free text and are essential components of high-quality information retrieval and extraction systems. Without good entity recognition, methods will mislabel searched text and will miss important information or identify spurious text that will frustrate users. Most tools do not capture non-contiguous entities which are separate spans of text that together refer to an entity, e.g., the entity "type 1 diabetes" in the phrase "type 1 and type 2 diabetes." This type is commonly found in biomedical texts, especially in lists, where multiple biomedical entities are named in shortened form to avoid repeating words. Most text annotation systems, that enable users to view and edit entity annotations, do not support non-contiguous entities. Therefore, experts cannot even visualize non-contiguous entities, let alone annotate them to build valuable datasets for machine learning methods. To combat this problem and as part of the BLAH6 hackathon, we extended the TextAE platform to allow visualization and annotation of non-contiguous entities. This enables users to add new subspans to existing entities by selecting additional text. We integrate this new functionality with TextAE's existing editing functionality to allow easy changes to entity annotation and editing of relation annotations involving non-contiguous entities, with importing and exporting to the PubAnnotation format. Finally, we roughly quantify the problem across the entire accessible biomedical literature to highlight that there are a substantial number of non-contiguous entities that appear in lists that would be missed by most text mining systems.

Keywords: editor, text annotation, text mining, visualization

Introduction

Information extraction and retrieval methods are essential tools to enable scientists to find and read the appropriate papers to enable discoveries. Many of these methods require identifying mentions of specific biomedical entities in the text and make use of named entity recognition (NER) tools for this task. Most entities are represented by a single span of text, e.g., the name of a drug. However, some entities are represented by multiple spans of text that are separated by other words and together identify the entity, for example, the separate words "skin" and "cancer" in "skin and lung cancer." These are known as non-contiguous, or discontiguous entities. Table 1 illustrates several more examples from public text mining resources. It should be noted that non-contiguous entities are different from anaphora or coreference resolution, in which multiple spans refer to the same entity separately and do not work together to identify the entity.

Robust annotation tools that are capable of annotating non-contiguous entities are important so that valuable entity information is not missed. These tools are needed to create

Table 1. Examples of non-contiguous entities from different public text mining datasets

Source	PubMed ID	Snippet	Non-contiguous entity
BioNLP 2019 Bacteria Biotope Task	23224222	Both **French** and German **cheeses** have previously been reported to contain M. psychrotolerans	French cheeses
	19622846	...and used API tests to identify S. aureus and E-tests to determine **methicil-lin**/oxacillin **resistance**	Methicillin resistance
CancerMine	19855840	It is suggested that DLC1 is a candidate tumour suppressor gene for human liver cancer, as well as for **prostate**, lung, colorectal and breast **cancers**	Prostate cancers
	19734946	LARG at chromosome 11q23 has functional characteristics of a tumor suppressor in human **breast** and colorectal **cancer**	Breast cancer
PGxMine	23385314	In vitro analysis and quantitative prediction of efavirenz inhibition of eight cytochrome P450 (CYP) enzymes: major effects on **CYP**s 2B6, 2C8, 2C9 and **2C19**	CYP 2C19

The examples from CancerMine [8] and PGxMine [9] are not currently captured by the corresponding method and are false negatives.

Table 2. An analysis of the annotation tools reviewed in Neves and Seva's study [4] for their capabilities to annotate non-contiguous entities

Tool	URL	Can run?	Supports entity annotation?	Support non-contiguous entities?
BioQRator	http://www.bioqrator.org	Y	Y	N
brat	http://brat.nlplab.org	Y	Y	Y
Catma	https://catma.de	Y	Y	Y
Djangology	https://sourceforge.net/projects/djangology	Y	Y	N
ezTag	https://eztag.bioqrator.org	Y	Y	N
FLAT	https://github.com/proycon/flat	Y	Y	N
LightTag	https://www.lighttag.io	Y	Y	N
MAT	http://mat-annotation.sourceforge.net	Y	Y	N
MyMiner	http://myminer.armi.monash.edu.au	Y	Y	N
PDFAnno	https://github.com/paperai/pdfanno	N	-	-
prodigy	https://prodi.gy/	Y	Y	N
tagtog	https://www.tagtog.net/	Y	Y	N
WAT-SL	https://github.com/webis-de/wat	Y	N	-
WebAnno	https://webanno.github.io	Y	Y	N

corpora with non-contiguous entities that can be used as training data for machine learning-based NER methods and also evaluate all NER methods fairly. The leading NER methods frequently use machine-learning methods such as conditional random fields (CRF) that are incapable of capturing non-contiguous entities without additional postprocessing. Popular tools such as BANNER [1], tmChem [2], and DNorm [3] do not support non-contiguous entities.

Many annotation tools have been developed for manual tagging of entities within a document for the biological domain and other domains. A detailed recent review of the strengths and weaknesses of different methods can be found in Neves and Seva's study [4]. To gauge the support for non-contiguous entities, we manually tested the 15 tools selected in that review with an overview shown in Table 2. We were able to run all but one, PDFAnno which displayed an error message that others have reported on Github. We found that only 2 support non-contiguous entities, BRAT [5], and Catma. Furthermore the AlvisAE [6] tool that was not included in the review also supports non-contiguous entity annotation. We suggest that more tools need to provide support for non-contiguous entities.

To that goal, we describe the addition of non-contiguous entity support to TextAE. TextAE is an annotation platform that forms part of the PubAnnotation system for storing and editing text annotations [7]. It is a Node.js web component that accepts text annotations in PubAnnotation JSON format. The PubAnnotation format currently has support for non-contiguous entities but are converted to an alternative representation when edited using the current release of TextAE, known as the chaining representation. This representation converts an entity that contains multiple spans to multiple entities and links them with a relation with type "_lexicallyChainedTo." This representation is time-consuming to edit and visualizes poorly. Fig. 1 illustrates the current representation of three non-contiguous entities within a sentence using the chaining

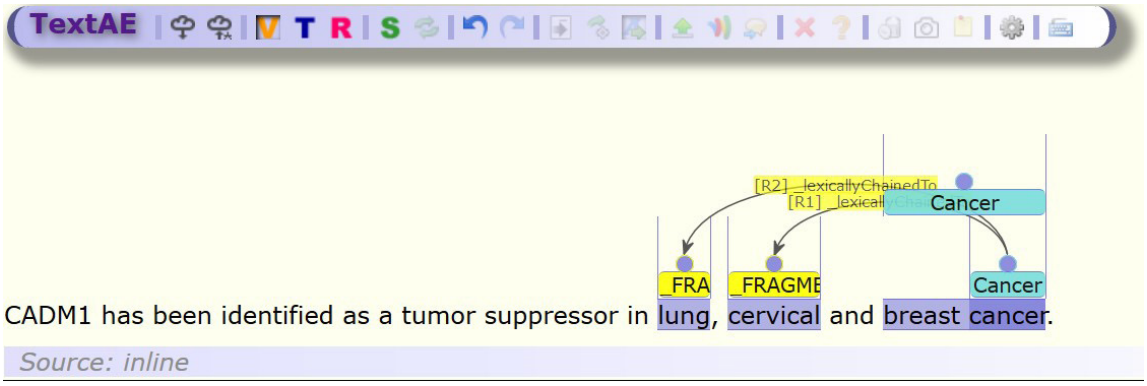

Fig. 1. Illustration of three entity annotations including two non-contiguous represented using the older chaining model which is cumbersome to annotate and visually cluttered.

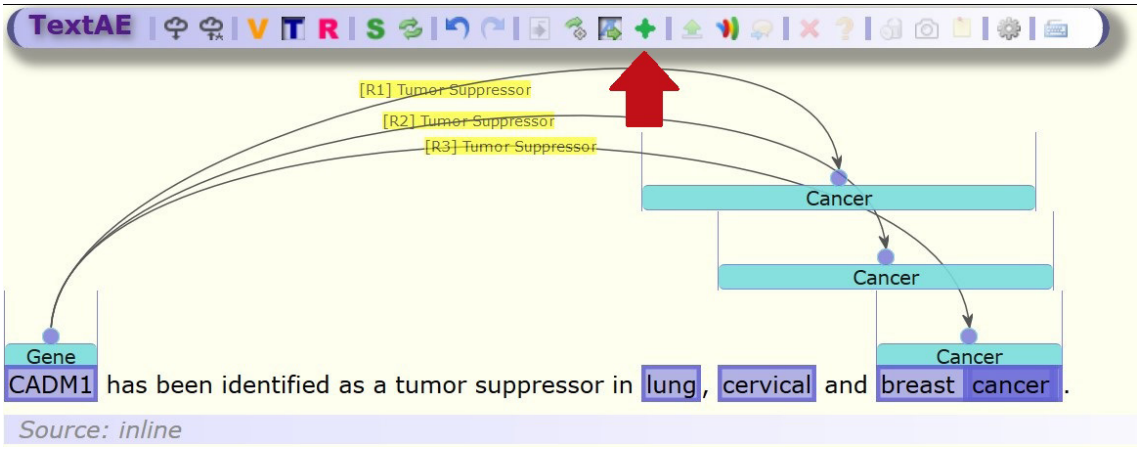

Fig. 2. The user interface with the new Add subspan button in the toolbar (highlighted by red arrow) and non-contiguous entities annotated as part of three relationships.

method. With the current TextAE interface, it is time-consuming to annotate each entity. Assuming TextAE has been set up with appropriate entity and relation types, for each non-contiguous entity, it requires creating two entities (2 mouse clicks), designating one entity with the type "_FRAGMENT" (2 clicks), switching to the relation mode (1 click), creating a relation between the two entities (2 clicks), changing its type to "_lexicallyChainedTo" (2 clicks) and switching back to Term Mode to continue entity annotation (1 click). Even for a TextAE power user, ten clicks for each non-contiguous entity is time-consuming for a large-scale annotation and produces an unwieldy result which is not visually clear.

In this paper, we describe our solution of an extension to the existing TextAE annotation platform to provide seamless support for annotating non-contiguous entities. Finally, we provide evidence of the widespread nature of non-contiguous entities in the biomedical literature using a rule-based extraction system to roughly quantify the scale of non-contiguous entities across all PubMed abstracts and accessible PubMed Central full-text papers.

Methods

To develop improved methods to capture non-contiguous entities, well-annotated data needs to be prepared and examined that contain non-contiguous entities. We extend the TextAE annotation platform that is part of the PubAnnotation system [7]. This enables annotation of entities with multiple spans as shown in Fig. 2 with a new subspan mode. The user can select new spans of text that will be added to an existing entity and displayed clearly.

The first task for implementation was changing the underlying span model in TextAE so that all spans are represented as a list of subspans. We dynamically check the input annotation data (in PubAnnotation format) to check if an entity has a single span, or a list of spans, and convert all entities to contain lists of spans, even for single spans. Previously, spans were rendered using a single

https://doi.org/10.5808/GI.2020.18.2.e19

HTML span tag around the section with appropriate CSS styling to identify the span as an entity. To visualize the new subspans, we removed the styling from the span class, and create subspans for each part of the span and transferred the stylings to the subspans. TextAE implements an Undo/Redo system so changes were required across the codebase to enable the existing functionality to work with the new underlying data structure and allow entities to be manipulated as before.

A new toggleable button (Add subspan) was added to the toolbar. When this button is toggled, any new spans that are selected by the user are added to the previously selected entity. This requires checking that new subspans were compatible with the structure that is enforced by the HTML page. This means that spans and subspans cannot intersect unless one is contained within the other entirely. This means that in the snippet: "breast cancer gene", it would not be possible for "breast cancer" and "cancer gene" to be tagged as entities. However "breast cancer" and "cancer" could be tagged as "cancer" is fully contained within "breast cancer." We have not come across use-cases where this functionality is currently needed but cannot discount the potential of this limitation. Fig. 2 illustrates the user interface with an example of non-contiguous entities.

TextAE has several user interface shortcuts to enable fast annotation and correction of entities. Users can extend an entity annotation by highlighting text that begins within an entity annotation and goes beyond the entity. Inversely, users can also shorten entity annotations by highlighting text that begins outside an entity span and finishes within an entity, thereby removing the selected text from the entity annotation. We extended this functionality to work for the new subspan system so that it would extend the appropriate first or last subspan in an entity outwards, or would shorten or even remove subspans that are highlighted. We further added user interface tweaks so that when a user selected a subspan, it would select all the subspans for the entity. Finally, we implemented export functionality so that the new subspans would be correctly stored in the PubAnnotation format with a list of spans for those entities with multiple subspans.

The code for this paper is available at https://github.com/jakelever/textae.

Results

We first tested to check that all existing functionality of TextAE remained operational. We confirmed that the new subspan model was able to load data containing non-contiguous entities and annotations with non-contiguous entities could be saved correctly to the PubAnnotation format. Furthermore, we tested that all existing functionality, including relation annotation, worked correctly with non-contiguous entity annotations.

We quantified the user interactions required to annotate non-contiguous entities. With this new interface, the user needs to annotate a single span (1 click), enable the Add subspan mode (1 click), add a new subspan (1 click), and disable the Add subspan mode (1 click). With only four clicks, we have drastically reduced the user effort, compared to the 10 clicks required previously, and no longer require the user to switch annotation modes within TextAE. Furthermore, the output is visually clearer. This performance is similar to the Catma tool, which requires four clicks to annotate a non-contiguous entity (1 to activate the discontinuous mode, 2 to select the two spans and 1 to select the entity type). And it is marginally easier than BRAT which takes five clicks (2 to annotate the first entity, 1 to edit the entity, 1 to select Add Frag and 1 to select the new span).

Discussion

While non-contiguous entities initially seem like a limited problem for text annotation, we note that two other BLAH 6 hackathon projects requested this functionality during the event: a project working on annotations from the recent BioNLP Shared Task [10] and a project focused on Medical Device Indication annotation. To understand the scale of this problem, we quantified the number of non-contiguous entities that appear in lists, as shown in the CancerMine examples in Table 1. We focussed on this format as these can be extracted using a modified dictionary matching method.

We used the PubTator Central resource [11] as it provides text-level entity annotations of a very large set of biomedical publications and also a rough set of synonyms for different entity types. The annotations provide locations of biomedical entities that may be the final element in a list. For example, the phrase "prostate, skin and lung cancer" would only likely be tagged for "lung cancer" in PubTator. We aimed to retrieve other entities from these lists using the set of synonyms from

PubTator Central, so that "prostate cancer" and "skin cancer" would be extracted from the example phrase. We used a simple rule-based system that identified candidate lists by searching for tagged biomedical entities that follow the word "and." We then searched the preceding words in the candidate list and attempted word substitutions with the final term to find terms that were in the lexicon.

Across the 30,044,935 abstracts and 2,485,641 full-text papers that were minable, we find 3,269,632 potential mentions of non-contiguous entities in the example list format. We manually

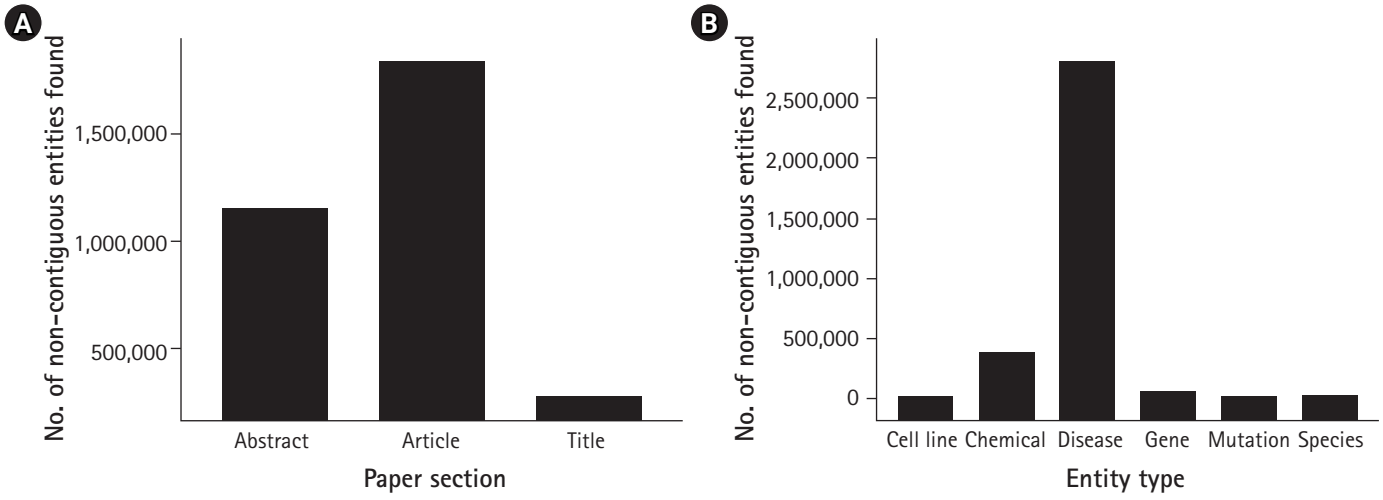

Fig. 3. The non-contiguous entities grouped the section of the paper (A) and the entity type (B).

reviewed 100 of them to understand the error profile and found that 42% were true positives. The main errors were caused by spurious mistakes in the lexicon and a more conservative lexicon would likely improve precision but may affect overall recall. Nevertheless, this initial result suggests that many biomedical entities are described in the list form that would be missed with most current methods. While there are considerable false-positive dues to the dictionary matching method, we would argue that this will only be a fraction of non-contiguous entities across the biomedical literature as we examine only one type of linguistic structure that could contain non-contiguous entities.

Fig. 3 shows an overview of the results from the literature analysis. Lists appear more in full-text papers than in abstracts even when taking account of the substantially larger number of abstracts than full-text articles in the corpus. They can even appear in the article title. Furthermore, disease has substantially more non-contiguous entities, likely due to the larger number of multiple word terms in that lexicon (837,390 compared to 103,427 for genes for example).

This analysis strongly suggests that non-contiguous are a substantial problem in biomedical text mining and that methods that ignore them will be missing large amounts of potential extracted information. We hope our contribution to an annotation tool that could help visualize and annotate these problematic entities may take a step towards new methods to identify them.

ORCID

Jake Lever: https://orcid.org/0000-0001-8198-2939
Russ Altman: https://orcid.org/0000-0003-3859-2905
Jin-Dong Kim: https://orcid.org/0000-0002-8877-3248

Authors' Contribution

Conceptualization: JL, JDK. Data curation: JL. Formal analysis: JL. Funding acquisition: RA, JDK. Methodology: JL. Writing – original draft: JL, RA, JDK. Writing – review & editing: JL, RA, JDK.

Conflicts of Interest

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

Acknowledgments

The authors would like to thank the funders of the Biomedical Linked Annotation Hackathon series.

References

1. Leaman R, Gonzalez G. BANNER: an executable survey of advances in biomedical named entity recognition. Pac Symp Biocomput 2008:652-633.
2. Leaman R, Wei CH, Lu Z. tmChem: a high performance approach for chemical named entity recognition and normalization. J Cheminform 2015;7:S3.
3. Leaman R, Islamaj Dogan R, Lu Z. DNorm: disease name normalization with pairwise learning to rank. Bioinformatics

https://doi.org/10.5808/GI.2020.18.2.e19

2013;29:2909-2917.
4. Neves M, Seva J. An extensive review of tools for manual annotation of documents. Brief Bioinform 2019 Dec 15 [Epub]. https://doi.org/10.1093/bib/bbz130.
5. Stenetorp P, Pyysalo S, Topic G, Ohta T, Ananiadou S, Tsujii J. BRAT: a web-based tool for NLP-assisted text annotation. In: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics (Segond F, ed.), 2012 Apr 23-27, Avignon, France. Stroudsburg: Association for Computational Linguistics, 2012. pp. 102-107.
6. Papazian F, Bossy R, Nedellec C. AlvisAE: a collaborative Web text annotation editor for knowledge acquisition. In: Proceedings of the Sixth Linguistic Annotation Workshop (Ide N, Xia F, eds.), 2012 Jul 12-13, Jeju, Korea. Stroudsburg: Association for Computational Linguistics, 2012. pp. 149-152.
7. 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, Tshujii J, Pestian J, eds.), 2012 Jun 3-8, Montreal, Canada. Stroudsburg: Association for Computational Linguistics, 2012. pp. 202-205.
8. Lever J, Zhao EY, Grewal J, Jones MR, Jones SJ. CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer. Nat Methods 2019;16:505-507.
9. Lever J, Barbarino JM, Gong L, Huddart R, Sangkuhl K, Whaley R, et al. PGxMine: text mining for curation of PharmGKB. Pac Symp Biocomput 2020;25:611-622.
10. Bossy R, Deleger L, Chaix E, Ba M, Nedellec C. Bacteria Biotope at BioNLP Open Shared Tasks 2019. In: Proceedings of The 5th Workshop on BioNLP Open Shared Tasks (Kim JD, Nedellec C, Bossy R, Deleger L, eds.), 2019 Nov 4, Hong Kong. Stroudsburg: Association for Computational Linguistics, 2019. pp. 121-131.
11. Wei CH, Allot A, Leaman R, Lu Z. PubTator central: automated concept annotation for biomedical full text articles. Nucleic Acids Res 2019;47:W587-W593.

G&I
Genomics & Informatics

Application note

eISSN 2234-0742
Genomics Inform 2020;18(2):e16
https://doi.org/10.5808/GI.2020.18.2.e16

Received: March 16, 2020
Revised: May 6, 2020
Accepted: May 22, 2020

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

Social Media Mining Toolkit (SMMT)

Ramya Tekumalla, Juan M. Banda*

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

There has been a dramatic increase in the popularity of utilizing social media data for research purposes within the biomedical community. In PubMed alone, there have been nearly 2,500 publication entries since 2014 that deal with analyzing social media data from Twitter and Reddit. However, the vast majority of those works do not share their code or data for replicating their studies. With minimal exceptions, the few that do, place the burden on the researcher to figure out how to fetch the data, how to best format their data, and how to create automatic and manual annotations on the acquired data. In order to address this pressing issue, we introduce the Social Media Mining Toolkit (SMMT), a suite of tools aimed to encapsulate the cumbersome details of acquiring, preprocessing, annotating and standardizing social media data. The purpose of our toolkit is for researchers to focus on answering research questions, and not the technical aspects of using social media data. By using a standard toolkit, researchers will be able to acquire, use, and release data in a consistent way that is transparent for everybody using the toolkit, hence, simplifying research reproducibility and accessibility in the social media domain.

Keywords: data mining, information storage and retrieval, machine learning, social media
Availability: All code described in this paper is fully available at: https://github.com/thepanacealab/SMMT.

Introduction

Only in the last six years, there has been a great influx of research works that describe different types of research works using Twitter and Reddit data, nearly 2,500 papers are found in PubMed [1]. These works encompass countless applications, such as the usage of opioids [2], the flu [3], eating disorder [4] networks analyses, depression symptoms detection [5], and diabetes interventions [6], etc. While all the listed studies use data from Twitter and Reddit, we can only find code available for one of them. Additionally, the data acquisition methodology is different on each study and seldomly reported, a crucial step towards reproducibility of any of their analyses. When it comes to using Twitter data for drug identification and pharmacovigilance tasks, authors of works like [7-9] have been consistently releasing publicly available datasets, software tools, and complete Natural Language Processing (NLP) systems with their works. In an attempt to shift the biomedical community into better practices for research transparency and reproducibility, we introduce the Social Media Mining Toolkit (SMMT), a suite of tools aimed to encapsulate the cumbersome details of acquiring, preprocessing, annotating, and standardizing social media data. The need for a toolkit like SMMT arose from our work using Twitter data for the characterization of disease transmission during natural disasters [10] and mining large-scale repositories for drug usage related tweets for pharmacovigilance purposes [11]. We originally wanted to use other researcher's tools and surprisingly we found very little code available with the majority outdated and non-functioning. Going one step

back, we did find rudimentary Python libraries to interact with the Twitter API, but some of their learning curves are steep and not overly documented. We then decided to clean and integrate our code into a toolkit that would help us provide a comprehensive resource for other researchers/users to replicate our work and to use in their own analyses, hence SMMT was born. Parallel to tools like SMMT, there are other research groups that are outlining frameworks to streamline the mining of social media like Sarker et al. [12], which are complementary to the use and need of this tool.

Methods

Programmed using Python version 3 and the latest Twitter API interfaces, the functionality of SMMT is divided into three separate sets of tools: data acquisition tools, data preprocessing tools, and data annotation and standardization tools. The particular versions of the additional Python libraries used by SMMT are available at the github documentation [13] since they are constantly updated and refreshed. Besides extensive usage documentation, the tool also provides two end-to-end usage examples, as well as additional Google Colaboratory [14] interactive Python notebooks with data usage examples. Note than in order to use most of the functionality of SMMT, users need to sign-up to acquire access to the Twitter Application Program Interface API. Once approved, users will be provided a set of API credential keys, more information can be found in [15]. Fig. 1 shows all current components of SMMT. In the following sections we provide additional details of each category of available tools.

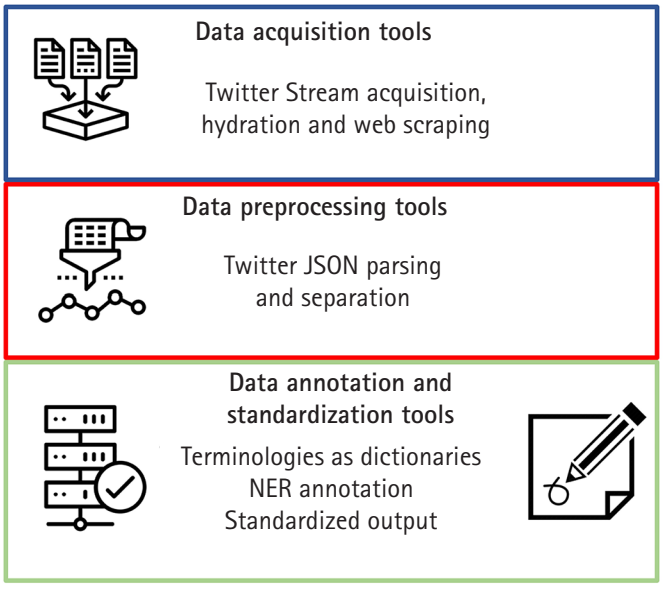

Fig. 1. SMMT tools categories and purpose outline.

Data acquisition tools

The tools in this category are used to gather data from social media sites, namely Twitter for this initial release of SMMT. The most common way of acquiring Tweets is to use the Twitter streaming API [16]. Our toolkit provides two separate utilities to capture streaming data, one will gather all available tweets and will continue running until terminated (*streaming.py*), and the other will take a list of search keywords and number of desired tweets and will pull those from the current tweet stream (*search_generic.py*). Details on how to use these utilities can be found on the READ.ME file.

The most common and permitted way of sharing Twitter data publicly is by only sharing the tweet id number. This number then needs to be 'hydrated', which means that the Twitter API needs to be used to fetch all the complete tweet and additional meta-data fields. This is a vital step for most users trying to replicate other studies or analyses. We provide a utility called *get_metadata.py* which reads a list of tweet ids and hydrates them automatically.

One of the major drawbacks of the Twitter API is the fact that unless having paid access to it, researchers cannot extract all historical tweets for any given Twitter user. Also, extracting all tweets from a given time range is not always easily and efficiently possible with the API. For these purposes we provide a utility called *scrape.py* which, once given a list of Twitter handles and corresponding date ranges, will automatically scrape the Twitter page and pull the tweet ids of the desired user and date range. These tweet ids then need to be 'hydrated' to be able to fully use them.

Data preprocessing tools

After having acquired enough data for research purposes from the Twitter stream, or identified and 'hydrated' a publicly available dataset, there is a need to subset the tweets and process the tweets JSON files to extract the fields of interest. While seemingly trivial, most biomedical researchers do not want to work with JSON objects, and since around 70% of the JSON fields are not populated, precise preprocessing steps need to be carried out to clean the data and render it useful in friendlier formats.

SMMT contains the *parse_json_lite.py* tool which takes a relatively small file (less than 1 Gigabyte in size) of Twitter JSON objects and separates these objects into a tab delimited file with each JSON field converted to a column and each tweet into a data row. With over 170 fields of meta-data, researchers are usually not interested in the vast majority of them. This tool can be configured to select which fields are of interest to be parsed and only process those into the tab delimited format. If the size of the tweets JSON objects file is larger than 1 Gigabyte, we provide an additional tool, *parse_json_heavy.py*, which can handle Terabyte sized files sequen-

tially rather than reading them all in memory for speed.

Once all the tweets are processed into the cleaner tab delimited format, which can even be read in Excel, there might be a need to further subset the tweets based on a given list of terms, or dictionary. For this purpose, we have included the *separate_tweet_tsv.py* file, which takes a term list in a format specified in the READ.ME file of SMMT and will return only the tweets that contain the provided terms.

Data annotation and standardization tools

After preprocessing the acquired social media data, researchers have the capabilities of standardizing their tweets' text with our set of tools. Taking advantage of OntoGene Bio Term Hub [17] and their harmonization of biomedical terminologies into a unified format, we provide a tool, *create_dictionary.py*, that converts their downloads into SMMT-compatible dictionaries. To avoid creating a complicated and cumbersome format for our tool, we opted for simplicity and only rely on having a tab delimited file with an identifier column and a term name column. Other dictionaries that we have made available will standardize any annotations using the Observational Health Data Sciences and Informatics (OHDSI) vocabulary [18]. We are testing functionality to also convert our dictionaries to the PubDictionaries [19] format for the next release, allowing researchers to use their functionality and online REST services.

One of the most important tools of SMMT is the Spacy [20] NER annotator, *SMMT_NER_basic.py*, this tool will take the tab delimited tweets, a dictionary file, and the name of the output file for the annotations. In order to extend the usability of our tool, we provide the resulting annotations in a traditional format: document, span, term format; as well as pre- formatted outputs compatible with the brat annotation tool [21] and PubAnnotation and its viewer TextAE [22] as shown in Fig. 2.

Discussion

While all the tools have their own documentation, in order to ease the adoption of the tools available in SMMT, we have included an end-to-end example in the examples folder that performs the following tasks:

(1) Download 300 tweets from the Twitter API stream for each of the following keywords: donald trump,coronavirus,cricket

(2) We then preprocess those tweets to extract Tweet Id and their text into tab delimited files.

(3) Using a Google Colab Notebook, we use these preprocessed files and then use the TF-IDF vectorizer on the text of the tweets to create a test and train set and build a Multi Nomial Naive Bayes Classifier to separate tweets based on their label. All details and steps of this process are outlined in the Colab Notebook.

(4) We then test our trained model on the test set and generate a confusion matrix heat-map (Fig. 3) of the classification task, and show the model performance metrics.

The whole process of this example takes less than 30 minutes to complete and is heavily documented for SMMT users to overcome the learning curve of acquiring and preprocessing tweets. While our example is simple in nature, users can build upon it and modify it to better suit their needs.

The tools part of SMMT allow users to simplify their research workflows and to focus on determining which data they want to use and the analyses they want to perform, rather than deciphering how to acquire the data. While most cutting-edge and near real-time research will be done pulling tweets from the Twitter API stream, there are countless datasets available for historical research, from large general purpose databases like the Internet Archive's Twitter Stream Grab dataset [23], which consists of data from 2014 to 2019, to more specialized and pre-curated datasets for

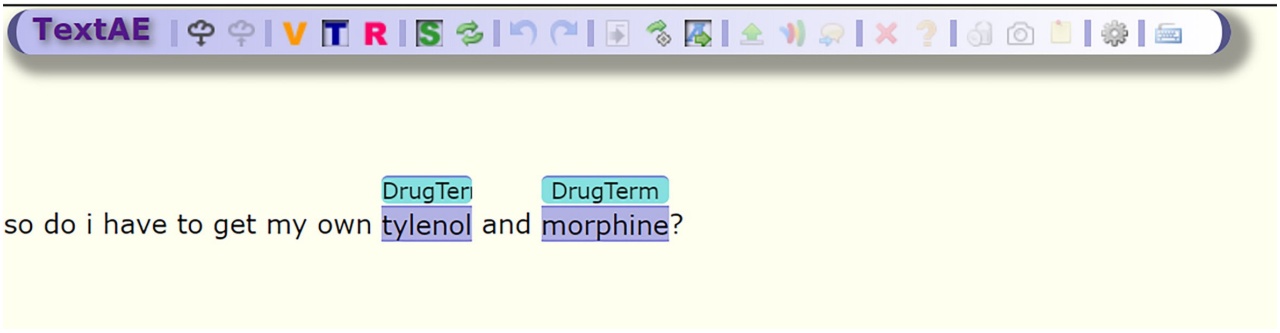

Fig. 2. Sample SMMT_NER_basic annotation using an RxNorm-based dictionary and displayed on TextAE.

https://doi.org/10.5808/GI.2020.18.2.e16

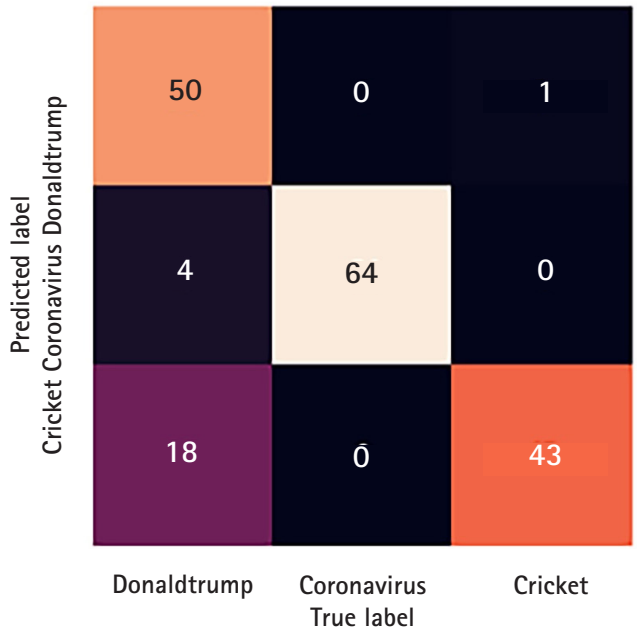

Fig. 3. Confusion Matrix Heat-Map for the classification of tweet labels.

uses like Pharmacovigilance [11] among others. This initial version release of SMMT will continue growing with additional tools being developed for platforms like Reddit, Dark Web forums, and other social media data sources.

ORCID

Ramya Tekumalla: https://orcid.org/0000-0002-1606-4856
Juan M. Banda: https://orcid.org/0000-0001-8499-824X

Authors' Contribution

Conceptualization: RT, JMB. Formal analysis: RT, JMB. Methodology: RT, JMB. Writing – original draft: RT, JMB. Writing – review & editing: RT, JMB.

Conflicts of Interest

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

Acknowledgments

We would like to acknowledge Jin-Dong Kim, DBCLS, and ROIS for making our participation in BLAH6 (Biomedical Linked Annotation Hackathon) possible. We also like to thank Javad Rafiei Asl, Kevin Bretonnel Cohen, Núria Queralt Rosinach, Yue Wang and Atsuko Yamaguchi for the help and input during the Biomedical Linked Annotation Hackathon 6.

References

1. PubMed search: social media. Bethesda: National Library of Medicine, 2020. Accessed 2020 Dec 3. Available from: https://www.ncbi.nlm.nih.gov/pubmed/?term=(Social+media+OR+social+network)+AND+(twitter+OR+reddit).
2. Jain P, Zaher Z, Mazid I. Opioids on Twitter: a content analysis of conversations regarding prescription drugs on social media and implications for message design. J Health Commun 2020;25:74-81.
3. Yun GW, Morin D, Park S, Joa CY, Labbe B, Lim J, et al. Social media and flu: Media Twitter accounts as agenda setters. Int J Med Inform 2016;91:67-73.
4. Moessner M, Feldhege J, Wolf M, Bauer S. Analyzing big data in social media: text and network analyses of an eating disorder forum. Int J Eat Disord 2018;51:656-667.
5. Jeri-Yabar A, Sanchez-Carbonel A, Tito K, Ramirez-delCastillo J, Torres-Alcantara A, Denegri D, et al. Association between social media use (Twitter, Instagram, Facebook) and depressive symptoms: are Twitter users at higher risk? Int J Soc Psychiatry 2019;65:14-19.
6. Gabarron E, Bradway M, Fernandez-Luque L, Chomutare T, Hansen AH, Wynn R, et al. Social media for health promotion in diabetes: study protocol for a participatory public health intervention design. BMC Health Serv Res 2018;18:414.
7. O'Connor K, Pimpalkhute P, Nikfarjam A, Ginn R, Smith KL, Gonzalez G. Pharmacovigilance on twitter? Mining tweets for adverse drug reactions. AMIA Annu Symp Proc 2014;2014:924-933.
8. Sarker A, Ginn R, Nikfarjam A, O'Connor K, Smith K, Jayaraman S, et al. Utilizing social media data for pharmacovigilance: a review. J Biomed Inform 2015;54:202-212.
9. 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.
10. Chowell G, Mizumoto K, Banda JM, Poccia S, Perrings C. Assessing the potential impact of vector-borne disease transmission following heavy rainfall events: a mathematical framework. Philos Trans R Soc Lond B Biol Sci 2019;374:20180272.
11. Tekumalla R, Asl JR, Banda JM. Mining Archive.org's twitter stream grab for pharmacovigilance research gold. Preprint at http://doi.org/10.1101/859611 (2019).

12. Sarker A, DeRoos A, Perrone J. Mining social media for prescription medication abuse monitoring: a review and proposal for a data-centric framework. J Am Med Inform Assoc 2020;27:315-329.
13. SMMT. San Francisco: GitHub, 2020. Accessed 2020 Dec 3. Available from: https://github.com/thepanacealab/SMMT.
14. Bisong E. Google colaboratory. In: Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners (Bisong E, ed.). Berkeley: Apress, 2019. pp. 59-64.
15. Apply for access. Twitter developers. San Francisco: Twitter Inc., 2020. Accessed 2020 Mar 12. Available from: https://developer.twitter.com/en/apply-for-access.
16. Standard stream parameters. San Francisco: Twitter Inc., 2020. Accessed 2020 Mar 12. Available from: https://developer.twitter.com/en/docs/tweets/filter-realtime/guides/basic-stream-parameters.
17. Ellendorff TR, van der Lek A, Furrer L, Rinaldi F. A combined resource of biomedical terminology and its statistics. In: Proceedings of the International Conference Terminology and Artificial Intelligence (Poibeau T, Faber P, eds.), 2015 Nov 4-6, Granada, Spain. Spanish Terminology Association, 2015. pp. 39-49.
18. 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.
19. PubDictionaries. Accessed 2020 Mar 12. Available from: http://pubdictionaries.org/.
20. spaCy. Industrial-strength natural language processing in Python. Explosion AI, 2017. Accessed 2020 Mar 12. Available from: https://spacy.io/.
21. Stenetorp P, Pyysalo S, Topic G, Ohta T, Ananiadou S, Tsujii J. BRAT: a web-based tool for NLP-assisted text annotation. Stroudsburg: Association for Computational Linguistics, 2012. Accessed 2020 Mar 12. Available from: https://www.aclweb.org/anthology/E12-2021.pdf.
22. Kim JD, Wang Y. PubAnnotation: a persistent and sharable corpus and annotation repository. Stroudsburg: Association for Computational Linguistics, 2012. Accessed 2020 Mar 12. Available from: https://dl.acm.org/doi/10.5555/2391123.2391150.
23. Archive Team: the Twitter Stream grab. San Francisco: Internet Archive, 2020. Accessed 2020 Mar 12. Available from: https://archive.org/details/twitterstream.

https://doi.org/10.5808/GI.2020.18.2.e16

G&I
Genomics & Informatics

Application note

eISSN 2234-0742
Genomics Inform 2020;18(2):e17
https://doi.org/10.5808/GI.2020.18.2.e17

Received: March 18, 2020
Accepted: June 18, 2020

*Corresponding author:
atsuko@tcu.ac.jp

A proof-of-concept study of extracting patient histories for rare/intractable diseases from social media

Atsuko Yamaguchi[1*], Núria Queralt-Rosinach[2]

[1]Tokyo City University, Setagaya, Tokyo 157-0087, Japan
[2]Leiden University Medical Center, Leiden, 2333 ZA, The Netherlands

The amount of content on social media platforms such as Twitter is expanding rapidly. Simultaneously, the lack of patient information seriously hinders the diagnosis and treatment of rare/intractable diseases. However, these patient communities are especially active on social media. Data from social media could serve as a source of patient-centric knowledge for these diseases complementary to the information collected in clinical settings and patient registries, and may also have potential for research use. To explore this question, we attempted to extract patient-centric knowledge from social media as a task for the 3-day Biomedical Linked Annotation Hackathon 6 (BLAH6). We selected amyotrophic lateral sclerosis and multiple sclerosis as use cases of rare and intractable diseases, respectively, and we extracted patient histories related to these health conditions from Twitter. Four diagnosed patients for each disease were selected. From the user timelines of these eight patients, we extracted tweets that might be related to health conditions. Based on our experiment, we show that our approach has considerable potential, although we identified problems that should be addressed in future attempts to mine information about rare/intractable diseases from Twitter.

Keywords: intractable diseases, rare diseases, social media mining
Availability: In this paper, we used Twitter timelines and the Human Phenotype Ontology. We obtained user timelines from Twitter (https://twitter.com) using Python code (https://github.com/acopom/smm4rd) with Tweepy (https:// www.tweepy.org/), which is a Python library for accessing the Twitter API (https://developer.twitter.com/). The Human Phenotype Ontology is available at https://hpo.jax.org/app/download/ontology.

Introduction

Social media has become a data source that is making a major contribution to big data. Recent scientific research has started to use and evaluate social media in the context of healthcare [1-4]. Svenstrup et al. [5] highlighted the potential of social media platforms dedicated to healthcare specialists as a means of knowledge-sharing for rare disease (RD) diagnoses. Schumacher et al. [6] introduced a case of online research and analysis of respondents using social media for the study of RDs. The role of social media was as a "participation caption" for recruiting a patient cohort and collecting clinical information. The authors concluded that the methodology and response patterns can be used for RD research. However, in those studies, social media platforms were used only from the viewpoint of healthcare specialists (e.g., medical doctors), even though a much broader range of people, including patients, are contributing to social media data. In particular, commu-

nities of patients suffering from RDs are very active on social media platforms. By definition, RDs affect small percentages of the population (https://ec.europa.eu/info/research-and-innovation/research-area/health-research-and-innovation/rare-diseases_en). These RD patient communities are small and patients are geographically scattered. Even though there are more than 8,000 RDs, only 5% have treatment. The lack of patient information available for research seriously hinders the diagnosis and treatment of rare and intractable diseases [7]. In general, RD patients suffer from very severe and heterogeneous symptoms and remain undiagnosed for several years [8]. Consequently, these disease communities use social media platforms to try to find other patients with similar health problems or expertise about their rare condition, sharing manifold types of information—including symptoms, treatments, side effects, and other diseases and activities—that go beyond what is normally captured in a clinical setting or patient registry [9]. Recently, Klein et al.[10] mined Twitter to collect data on rare health-related events reported by patients, and showed that this social media platform was useful for gathering patient-centric information that could be used for future epidemiological analyses. Our hypothesis was that data from RD patient histories posted on social media would capture patients' perspectives of their health status, which may be valuable for research into ways of helping undiagnosed patients by accelerating the timeline to diagnosis and treatment.

The special theme of the Biomedical Linked Annotation Hackathon 6 (BLAH6) was "social media mining." Therefore, we attempted to extract patient-centric knowledge from social media as a task for the 3-day hackathon. In this paper, we present our work that we conceived, designed, and developed during BLAH6 to explore the potential of social media data as a source of patient-centric knowledge. For this project, we focused on rare and intractable diseases and selected Twitter to obtain patients' timelines, as this platform may contain descriptions of the history of their health conditions. By focusing on the date of diagnosis, we intended to obtain histories of their health conditions before and after diagnosis.

Methods

Due to the time constraints of the hackathon, we selected one RD and one intractable disease. Then, we searched for patients with the two diseases and obtained their timelines. We also tried to extract tweets related to the disease and symptoms from each timeline.

First, we selected a RD that is adult-onset and not too rare to facilitate the extraction of a proper amount of data for analysis. To do so, we used information on the number of patients diagnosed with rare and intractable diseases in Japan, provided by Japan Intractable Disease Information Center (https://www.nanbyou.or.jp/). Based on this information, we selected amyotrophic lateral sclerosis (ALS) as an RD, and for similar reasons, we selected multiple sclerosis (MS) as an intractable disease. Second, we obtained a list of Twitter users who were diagnosed with ALS or MS using the search terms "I was diagnosed" and the disease name. Then, we selected users diagnosed during the last 5 years who had more than 100 tweets, excluding retweets and replies. This resulted in four users for each disease. By using Tweepy with a Python script (https://github.com/acopom/smm4rd), we obtained the timestamp and the text of the Twitter timelines, including 6088 tweets without retweets and replies for the eight users.

To extract tweets dealing with a user's health conditions, we used all terms in the Human Phenotype Ontology (HPO) [11] except for three ("all," "left," and "right"). All tweets that included HPO terms in the text were extracted. We then removed some tweets by manual search inspection because they described the health condition of someone else, such as the user's child. Through this process, we obtained a set of tweets that were related to the user's health condition. We called this set of tweets "tweets by HPO" for a user u and denoted it as $H(u)$.

Additionally, we extracted tweets dealing with health conditions using common words, such as "cold." However, many tweets extracted in this way were not related to health conditions, for example, "It's cold today." Consequently, we manually removed many tweets from this extracted tweet set. We called this set of tweets "tweets by manual" for a user u and denoted it as $M(u)$.

We called $H(u) \cup M(u)$ "tweets about the disease" and denoted this set as $D(u)$. As each tweet in $D(u)$ may contain sensitive information from the viewpoint of user protection, a short summary of each tweet to conceal details was made manually.

Results and Discussion

To conceal the identity of the users with ALS and MS, we used ALS1, ALS2, ALS3, and ALS4 to refer to the ALS patients and MS1, MS2, MS3, and MS4 to refer to the MS patients instead of their Twitter user names. Table 1 shows the numbers of tweets, the number of tweets in $H(u)$, and the number of tweets in $M(u)$ for each user u. Of note, all tweets about ALS were posted after the users were diagnosed, whereas all tweets about MS, except for one, were posted before the diagnosis.

We next constructed a patient history for each user u using tweets in $D(u)$. For example, ALS1 had two tweets in H(ALS1) extracted by the HPO term "pain" (HP:0012531). M(ALS1) in-

Table 1. Summary of the eight users analyzed in this experiment

User	#Tweets	#H	#M
ALS1	2135	2	3
ALS2	1295	0	0
ALS3	213	1	1
ALS4	182	7	5
MS1	777	3	1
MS2	348	1	0
MS3	572	0	2
MS4	566	2	1
Total	6088	16	13

#Tweets, #H, and #M show the total numbers of tweets, the number of tweets in H, and the number of tweets in M, respectively.

cluded three tweets that were extracted manually. From these five tweets, we obtained four events related to health conditions because two of the tweets in $H(\mathrm{ALS1})$ indicated one event. Fig. 1 shows the patient history of ALS1, who had four events after diagnosis. We set the date of diagnosis as a reference point. We presented short summaries such as "can talk" instead of showing real tweets because the extracted tweets may contain sensitive information from the viewpoint of user protection. At 270 days after the date of diagnosis, we can see that ALS1 could work, walk, and talk. However, ALS1 could no longer walk 644 days after the date of diagnosis.

Similarly, Fig. 2 shows the patient history of MS1, who had three

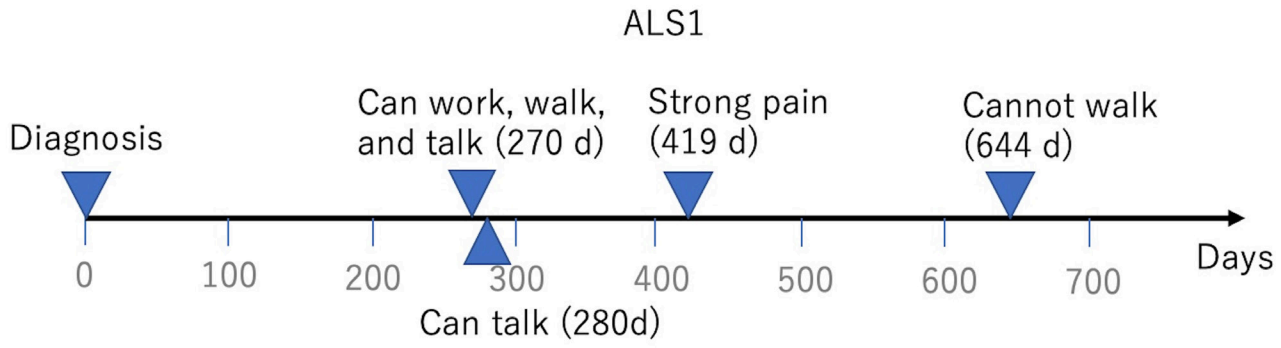

Fig. 1. Patient history with four events constructed by five tweets in D(ALS1). ALS1, amyotrophic lateral sclerosis 1.

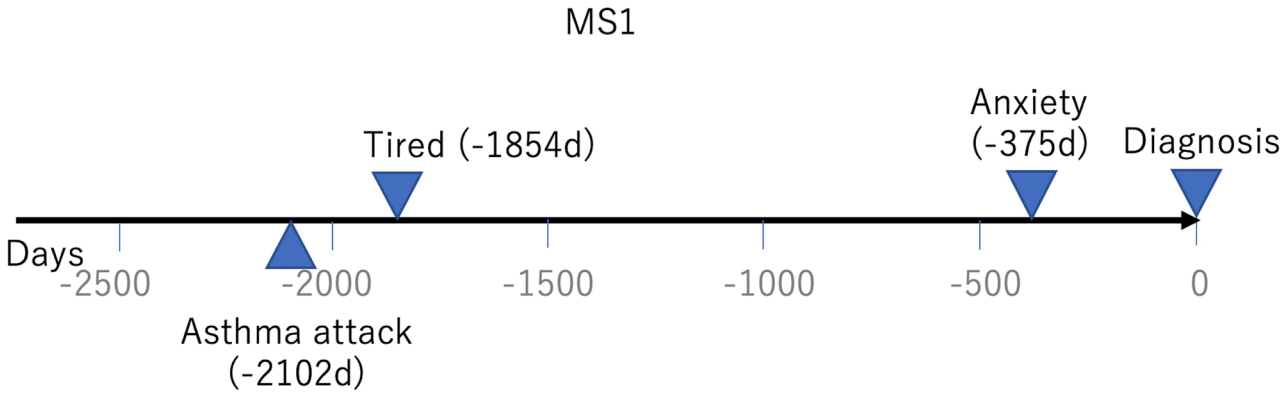

Fig. 2. Patient history with three events constructed by four tweets in D(MS1) MS, multiple sclerosis.

events as constructed by four tweets in $D(MS1)$. MS1 had an asthma attack 2,102 days before the diagnosis, and experienced anxiety and received a drug for it 375 days before the diagnosis.

This experiment showed the potential of Twitter data as a source of patient-centric knowledge, by extracting tweets related to health conditions and constructing a patient history from each user's timeline. However, we found that the typical method of scientific data extraction did not work well for mining tweets. As shown in Table 1, we obtained a very small number of tweets related to health conditions. To address this limitation, the development of a dictionary for the healthcare domain specialized for social media data is vitally necessary to leverage and better understand the scientific value of data from social media for rare and intractable diseases.

ORCID

Atsuko Yamaguchi: https://orcid.org/0000-0001-7538-5337
Núria Queralt Rosinach: https://orcid.org/0000-0003-0169-8159

Authors' Contribution

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

Conflicts of Interest

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

Acknowledgments

The authors would like to express their gratitude to the participants of BLAH6, especially to members of Social Media Mining Group for their valuable comments. This work was supported by the National Bioscience Database Center (NBDC) of the Japan Science and Technology Agency (JST).

References

1. Hays R, Daker-White G. The care.data consensus? A qualitative analysis of opinions expressed on Twitter. BMC Public Health 2015;15:838.
2. Allen CG, Andersen B, Chambers DA, Groshek J, Roberts MC. Twitter use at the 2016 Conference on the Science of Dissemination and Implementation in Health: analyzing #DIScience16. Implement Sci 2018;13:34.
3. Pemmaraju N, Utengen A, Gupta V, Kiladjian JJ, Mesa R, Thompson MA. Rare cancers and social media: analysis of Twitter metrics in the first 2 years of a rare-disease community for myeloproliferative neoplasms on social media-#MPNSM. Curr Hematol Malig Rep 2017;12:598-604.
4. Pemmaraju N, Utengen A, Gupta V, Thompson MA, Lane AA. Analysis of first-year Twitter metrics of a rare disease community for blastic plasmacytoid dendritic cell neoplasm (BPDCN) on social media: #BPDCN. Curr Hematol Malig Rep 2017;12:592-597.
5. Svenstrup D, Jorgensen HL, Winther O. Rare disease diagnosis: a review of web search, social media and large-scale data-mining approaches. Rare Dis 2015;3:e1083145.
6. Schumacher KR, Stringer KA, Donohue JE, Yu S, Shaver A, Caruthers RL, et al. Social media methods for studying rare diseases. Pediatrics 2014;133:e1345-e1353.
7. Kaufmann P, Pariser AR, Austin C. From scientific discovery to treatments for rare diseases: the view from the National Center for Advancing Translational Sciences - Office of Rare Diseases Research. Orphanet J Rare Dis 2018;13:196.
8. Kerr K, McAneney H, Smyth LJ, Bailie C, McKee S, McKnight AJ. A scoping review and proposed workflow for multi-omic rare disease research. Orphanet J Rare Dis 2020;15:107.
9. Subirats L, Reguera N, Banon AM, Gomez-Zuniga B, Minguillon J, Armayones M. Mining Facebook data of people with rare diseases: a content-based and temporal analysis. Int J Environ Res Public Health 2018;15:1877.
10. Klein AZ, Sarker A, Cai H, Weissenbacher D, Gonzalez-Hernandez G. Social media mining for birth defects research: a rule-based, bootstrapping approach to collecting data for rare health-related events on Twitter. J Biomed Inform 2018;87:68-78.
11. Kohler S, Carmody L, Vasilevsky N, Jacobsen JO, Danis D, Gourdine JP, et al. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Nucleic Acids Res 2019;47:D1018-D1027.

https://doi.org/10.5808/GI.2020.18.2.e17

G&I
Genomics & Informatics

Application note

eISSN 2234-0742
Genomics Inform 2020;18(2):e18
https://doi.org/10.5808/GI.2020.18.2.e18

Received: March 13, 2020
Revised: May 28, 2020
Accepted: May 28, 2020

*Corresponding author:
E-mail: mariana.lara-neves@bfr.bund.de

Integration of the PubAnnotation ecosystem in the development of a web-based search tool for alternative methods

Mariana Neves*

German Federal Institute for Risk Assessment (BfR), German Centre for the Protection of Laboratory Animals (Bf3R), 12277 Berlin, Germany

Finding publications that propose alternative methods to animal experiments is an important but time-consuming task since researchers need to perform various queries to literature databases and screen many articles to assess two important aspects: the relevance of the article to the research question, and whether the article's proposed approach qualifies to being an alternative method. We are currently developing a Web application to support finding alternative methods to animal experiments. The current (under development) version of the application utilizes external tools and resources for document processing, and relies on the PubAnnotation ecosystem for annotation querying, annotation storage, dictionary-based tagging of cell lines, and annotation visualization. Currently, our two PubAnnotation repositories for discourse elements contain annotations for more than 110k PubMed documents. Further, we created an annotator for cell lines that contain more than 196k terms from Cellosaurus. Finally, we are experimenting with TextAE for annotation visualization and for user feedback.

Keywords: animal testing alternatives, information storage and retrieval, text mining
Availability: Our resources and tools are available at: https://github.com/mariananeves/pubannotation-integration.

Introduction

According to the Directive 2010/63/EU (https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32010L0063), researchers who plan to carry out animal experimentation are required to examine whether alternative methods not entailing the use of a live animal are already available for the planned research purpose (replacement). In addition, the chosen method should ensure that the animal number is reduced to a minimum without comprising the objective (reduction), and to reduce the possible pain, distress, and suffering (refinement). These measures are known as the 3R principles.

When searching for alternative methods to animal experiments, researchers have to carry out various queries to bibliographic databases, e.g., PubMed (https://www.ncbi.nlm.nih.gov/pubmed/), and carefully analyze the potential candidate articles. For each of these potential articles, the researcher should check whether it addresses two important issues: (1) a method for replacement, and (2) the planned research question. To assist researchers in their search for alternative methods, we are currently developing a Web application that addresses these two aspects. We rank the potential candidate articles based on the similarity of the research question (with regard to an input article) and iden-

tify the proposed methods in each of the articles.

For the implementation of the Web application, we rely on various tools, such as document classification, named-entity recognition, and annotation storage, among others. In the scope of the BLAH6 Hackathon (https://blah6.linkedannotation.org/), we integrated the PubAnnotation ecosystem [1] into the backend of our application. PubAnnotation contains three main components that can support our application for some of those tasks: the PubAnnotation repository, PubDictionaries, and the TextAE annotation tool.

Here we describe the integration of these tools into our application. We start by introducing our application, followed by how the components are being integrated into it. The Web application is still under development and not yet available for the final user. However, the resources that we created in the PubAnnotation platform are already available to the research community.

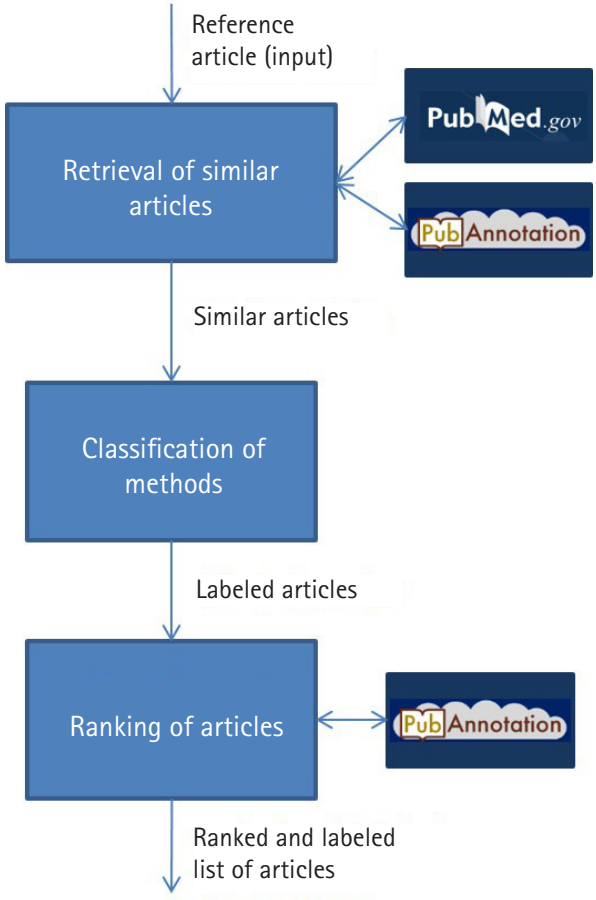

Fig. 1. Workflow of the interaction between the application with PubMed and PubAnnotation.

The Web Application

Fig. 1 shows an overview of the real-time interaction of the Web application with PubMed and PubAnnotation. Given a reference article as input, our application retrieves the so-called similar articles from PubMed (https://www.ncbi.nlm.nih.gov/pubmed), i.e., the ones that were pre-compiled by PubMed [2]. The tool performs two processing tasks based on the title and abstracts of these similar articles: (1) classification of the proposed methods; and (2) ranking of the retrieved similar articles according to the similarity of their research questions to the one in the reference article.

The classification of the proposed methods will utilize machine learning algorithms to be trained based on manually annotated articles (abstracts), which are currently being manually labeled. These labels cover the various types of methods that are relevant for our domain, such as whether the experiments have been carried out *in vivo* (e.g., vertebrates, invertebrates) or *in vitro* (e.g., cell lines or organs). Further, we are also experimenting with named-entity recognition tools to support this task. We focus on entities which are still not well supported by the existing tools, e.g., cell lines, and on a dictionary-based approach that relies on the comprehensive Cellosaurus resource [3].

For the ranking task, the application calculates the text similarity between the reference article and each of the PubMed similar articles, and the resulting scores are used to rank these articles. For the text similarity, we utilize the TextFlow tool [4]. However, instead of relying on the whole abstract of the articles, we utilize only the most relevant discourse categories (or zones), such as "introduction" and "results". This is due to the fact that only some parts of the abstract potentially describe the research question.

We recently published a study in which we compared four tools for the extraction of the zones and evaluated them on seven use studies (https://github.com/mariananeves/scientific-elements-text-similarity) [5]. Our study also demonstrated that using pre-selected zones, instead of the whole abstracts, yields better performance in the ranking task.

The zones can be manually annotated or automatically detected. The manually annotated zones are the original ones included in the structured abstracts in Pubmed (https://www.nlm.nih.gov/bsd/policy/structured_abstracts.html). However, given that not all abstracts in PubMed are structured, we automatically extract the zones for the remaining articles using the ArguminSci tool [6]. This was the best performing tool according to our study [5].

Finally, our Web application will contain visual components to display the abstract of the articles involved in the search. We currently consider two scenarios. The first is the visualization of the reference article in order to obtain feedback from the user, e.g., the research question in mind, by asking the user to highlight this in-

formation on the text. The second scenario is a side-by-side display of the reference article and each one of the retrieved similar articles in order to compare two articles and further gather user feedback.

Integration with the PubAnnotation Ecosystem

We are currently integrating the PubAnnotaton ecosystem in various components of our Web application in order to support various tasks, namely, storage, alignment, named-entity recognition, and visualization of annotations. Here we describe how each of the tools is being integrated in our application.

PubAnnotation database

We utilize the PubAnnotatin database to allow easy storage and retrieval of PubMed titles, abstracts, and their annotations. We store the zones coming from the reference article and its similar articles into one of the two repositories that we created in PubAnnotation, depending on the origin of these zones: (1) the PubMed_Structured_Abstracts repository (http://pubannotation.org/projects/PubMed_Structured_Abstracts) for the original zones available in the structured abstracts in PubMed; and (2) the PubMed_ArguminSci repository (http://pubannotation.org/projects/PubMed_ArguminSci) for the zones automatically extracted by the ArguminSci tool [6]. Both repositories are public and the annotations (zones) can be retrieved using the PubAnnotation API (http://www.pubannotation.org/docs/intro/).

For each article processed by our application, both reference articles or similar articles, we first check whether the article is already included in the PubAnnotation (cf. "fetch article" in Fig. 2), i.e., in any of its repositories. The output is either a message that the article is inexistent or a JSON object that includes the article's title, abstract, and its annotations, which may come from various repositories in PubAnnotaton. We check whether annotations already exist from any of our two repositories described above (cf. "get zones" in Fig. 2). If any zones could be found, these are returned to be further processed by the Web application.

In case that no zones have been stored for the article in none of our two repositories, we first check whether the article contains a structured abstract. This information is contained in the data retrieved from PubMed. If the article contains a structured abstract, its zones are simply stored into PubAnnotation (cf. "store zones" in Fig. 2) and will be available for future queries. Otherwise, we extract the zones using the ArguminSci tool, followed by their stor-

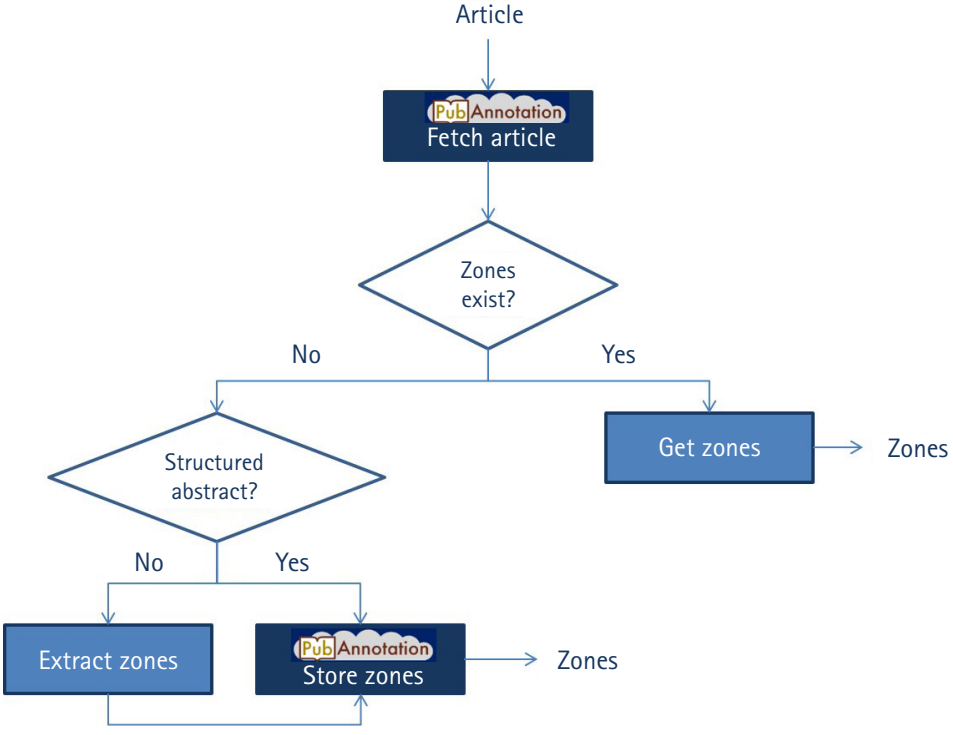

Fig. 2. Workflow of the interaction of the Web application with the PubAnnotation API. The components that perform calls to API are identified with the logo of PubAnnotation.

age into PubAnnotation. For any of the two situations, we store the zones in PubAnnotation in a two-steps procedure: (1) we add the article into the corresponding repository (either PubMed_Structured_Abstracts or PubMed_ArguminSci), and (2) we add the annotations into the same repository. It is not possible to perform the second step if the article was not previously included in the repository.

As of May 2020, the PubMed_Structured_Abstracts repository contains more than 31k documents while the PubMed_ArguminSci repository holds almost 80k documents. Therefore, we can state that less than 30% of the documents processed by our application included a structured abstract, while we had to perform predictions for zones for more than 70% of them. These documents consist of reference articles and the corresponding similar articles derived from the various queries that we made to our application in the last months, but also from the machine learning experiments that we carried out for the document classification step. Currently, we do not plan to include zoning annotations for all articles in PubMed, but just for those that happen to be processed by our application during our various experiments, and later, from the queries made by the users. Therefore, the repositories should incrementally grow with the time.

Another interesting feature in PubAnnotation is the annotation alignment. We deal with annotations retrieved from two sources, i.e., ArguminSci and PubMed Structured Abstracts, whose annotations might have been derived from a slightly different version of the article's abstract, or the corresponding text somehow altered by the tool during processing. The annotation alignment function in PubAnnotation automatically converts the offsets of these annotations to the article's abstract that is stored in PubAnnotation. Therefore, this function relieves us from writing customized scripts for dealing with the annotations returned by the various resources or tools.

Currently, no storage in PubAnnotation is being carried out for annotations coming from the classification task. This is due to a couple of reasons. First, the performance of our algorithms is not yet satisfactory. Further, document-level annotations are currently not supported by the JSON format of PubAnnotation. However, we plan to store them in PubAnnotation in the near future.

TextAE (Text Annotation Editor)

Besides using the PubAnnotation ecosystem for annotation storage, we also plan to rely on other tools of the plattform in our Web application. For instance, we are currently experimenting with the TextAE tool (http://textae.pubannotation.org/) for displaying articles and annotations to the user. TextAE can be embedded into a HTML page to display the text and annotations that are passed in the JSON format. TextAE can be used in both of our visualization scenarios, i.e., either for displaying single articles or a side-by-side comparison. For the first scenario, an editable version of TextAE can potentially be used for collecting user feedback on the reference article, i.e., through text highlighting. For the second scenario (side-by-side comparison), we currently display annotations for species, disease, and chemicals from PubTator Central [7] using its annotator (http://pubannotation.org/annotators/PubTator) that is currently available in PubAnnotation. We also envisage relying on a side-by-side comparison to gather feedback from the user about the similarity of both research questions.

PubDictionaries

We are also experimenting with PubDictionaries in the PubAnnotation ecosystem. Given a dictionary composed of terms (i.e., sets of identifiers and names), it is possible to perform a dictionary-based named-entity recognition by matching the terms in the dictionary to the title and abstract of articles in PubAnnotation. We are evaluating this functionality for the task of identifying cell lines, which might support our classification task, in addition to the machine learning approach.

For this purpose, we created the Cellosaurus_v33 dictionary (http://pubdictionaries.org/dictionaries/Cellosaurus_v33) that includes cell lines released in version 33 of Cellosaurus [3]. Based on this dictionary, we created a corresponding annotator in PubAnnotation (http://pubannotation.org/annotators/Cellosaurus_v33), which is a Web service that can be applied to any article in PubAnnotation for real-time annotation. We are currently developing a pre-processing script to filter out cell line names that match to a list of stopwords. Further, a post-processing script will also be applied to filter out mentions that match entities returned by PubTator Central, which are potential false positives.

Conclusion

We presented the integration of the PubAnnotation ecosystem in our planned Web application which aims to mine alternative methods to animal experiments. We use all main functionalities of the ecosystem, namely, the PubAnnotation repository for the storage and alignment of annotations, PubDictionaries for dictionary matching of cell lines, and the TextAE annotation tool for the visualization of articles and annotations. Two repositories for annotations of discourse elements were created and are available to the research community. Further, these two repositories are being frequently and automatically updated by our Web application. Finally, we also released a cell line dictionary and its corresponding annotator.

https://doi.org/10.5808/GI.2020.18.2.e18

ORCID

Mariana Neves: https://orcid.org/0000-0002-6488-2394

Conflicts of Interest

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

References

1. Kim JD, Wang Y, Fujiwara T, Okuda S, Callahan TJ, Cohen KB. Open Agile text mining for bioinformatics: the PubAnnotation ecosystem. Bioinformatics 2019;35:4372-4380.
2. Lin J, Wilbur WJ. PubMed related articles: a probabilistic topic-based model for content similarity. BMC Bioinformatics 2007;8:423.
3. Bairoch A. The Cellosaurus, a cell-line knowledge resource. J Biomol Tech 2018;29:25-38.
4. Mrabet Y, Kilicoglu H, Demner-Fushman D. TextFlow: a text similarity measure based on continuous sequences. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Vol. 1, Long Papers) (Barzilay R, Kan MY, eds.), 2017 Jul 30-Aug 4, Vancouver, Canada. Stroudsburg: Association for Computational Linguistics, 2017. pp. 763-772.
5. Neves M, Butzke D, Grune B. Evaluation of scientific elements for text similarity in biomedical publications. In: Proceedings of the 6th Workshop on Argument Mining (Stein B, Wachsmuth H, eds.), 2019 Aug 1, Florence, Italy. Stroudsburg: Association for Computational Linguistics, 2019. pp. 124-135.
6. Lauscher A, Glavas G, Eckert K. ArguminSci: a tool for analyzing argumentation and rhetorical aspects in scientific writing. In: Proceedings of the 5th Workshop on Argument Mining, 2018 Nov, Brussels, Belgium. Stroudsburg: Association for Computational Linguistics, 2018. pp. 22-28.
7. Wei CH, Allot A, Leaman R, Lu Z. PubTator central: automated concept annotation for biomedical full text articles. Nucleic Acids Res 2019;47:W587-W593.

G&I
Genomics & Informatics

Application note

eISSN 2234-0742
Genomics Inform 2020;18(2):e19
https://doi.org/10.5808/GI.2020.18.2.e19

Received: March 18, 2020
Revised: May 26, 2020
Accepted: May 26, 2020

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

Enabling a fast annotation process with the Table2Annotation tool

Pierre Larmande[1,2*], Kazim Muhammed Jibril[2]

[1]DIADE, Univ. Montpellier, IRD, Montpellier 34398, France
[2]ICTLab, USTH, Hanoi 10000, Vietnam

In semantic annotation, semantic concepts are linked to natural language. Semantic annotation helps in boosting the ability to search and access resources and can be used in information retrieval systems to augment the queries from the user. In the research described in this paper, we aimed to identify ontological concepts in scientific text contained in spreadsheets. We developed a tool that can handle various types of spreadsheets. Furthermore, we used the NCBO Annotator API provided by BioPortal to enhance the semantic annotation functionality to cover spreadsheet data. Table2Annotation has strengths in certain criteria such as speed, error handling, and complex concept matching.

Keywords: bioinformatics, ontologies, semantic annotation
Availability: GitHub: https://github.com/pierrelarmande/Table2Annotation.

Introduction

Semantic annotation has been defined in various ways by various authors, but these definitions are all similar and reflect a single clear purpose. For instance, Oliveira and Rocha [1] defined semantic annotation as the process in which semantic concepts are linked to natural language. Liao et al. [2] defined semantic annotation as methods of describing resources (texts, images ...) with metadata where the meaning has been specified in an ontology. According to Oliveira and Rocha [1] semantic annotation can be seen as a methodology of adding metadata—comprising classes, properties, relations, and instances (i.e., the concepts of an ontology)—to web resources to be able to give or allocate semantics. Summarizing all of these definitions, we can simply state that semantic annotation is a way of matching resources to ontologies.

To make this point clearer, take this example of the text *"..days to flowering…"* With the help of semantic annotation, we would be able to match this text to the ontology concept "*days to flowering trait*" from the Trait Ontology [3], which has the concept ID of TO:0000344.

Semantic annotation helps to boost the ability to search and access resources. It is also a step towards data FAIRification [4]. According to Jovanovic and Bagheri [5], semantic annotation can be used in information retrieval to expand the queries from the user with some ontology terms and also to provide a grouping of documents retrieved based on specific content. Biomedical resources contain numerous abbreviations in the texts, sometimes with different meanings, which makes it hard to perform comprehensive searches. Semantic annotation helps to disambiguate these abbreviated terms based on the way they appear in a certain context.

In this paper, we sought to identify ontological concepts in scientific texts. This could

be seen as an ontology matching process, in which natural language texts are matched with concepts. There are already some existing web services and tools that use semantic annotation for ontology matching, as have been evaluated by Oliveira and Rocha [1]. However, few of these tools handle spreadsheet data as text input. We developed the Table2Annotation tool with that purpose because there is a need for such a tool in the life sciences community, which produces extensive experimental data in spreadsheets. Semantic annotations will facilitate more complex analyses across several datasets.

The paper is organized as follows. Section 2 defines the challenges of semantic annotation. Section 3 presents an overview of Table2Annotation. Section 4 analyzes the results of semantic annotation through some examples. Section 5 concludes the manuscript.

Semantic Annotation Challenges

Semantic annotation has some benefits, but some challenges are also faced during the annotation of biological texts or other resources. Some of these challenges, as also described in previous research [6-8], are follows:

- Word sense disambiguation: It is necessary to determine the correct meaning of a word as used in a sentence when a word has multiple meanings.
- Spelling/grammatical error identification: Correcting spelling or grammar in biomedical texts is very important. Spelling and grammar errors cause ambiguity in already sparse text.
- Discontinuous entities: Entities can be composed of multiple words in a discontinuous span. For example, "*drought and salinity tolerance*" means "*drought tolerance and salinity tolerance,*" but in this case we might only have matching for "salinity tolerance."
- Gene/protein disambiguation: In the biomedical context, all proteins have associated genes, often with the same name, making it difficult to annotate texts dealing with genes and proteins.
- Detection of name variants: Variations of entity naming can take many forms, thereby complicating annotations. For example, abbreviations and shorthand texts are difficult to normalize with ontological concepts.

The challenges faced in annotation can be tackled by two approaches, which can be also combined. The first one is the term-to-concept matching method, which involve matching some parts of the provided text to structured knowledge databases, dictionaries, or vocabularies. However, it is difficult to maintain comprehensive lexicons to be used for annotation. The second approach is machine learning, which involves creating annotators for specific purposes and usage instead of more general ones [5].

Of particular note, the third challenge (discontinuous entities) can be tackled by creating algorithms that can transform texts with conjunctions like "*and*" or "*or.*" Thus, in the example of "*drought and salinity tolerance,*" an algorithm could transform this phrase to "*drought tolerance and salinity tolerance*" before the annotation process.

Although these are good solutions to tackle some challenges, some drawbacks remain. For instance, a drawback of term-to-concept matching is its inability to disambiguate terms, so annotators that inherit this method usually match terms with several possibilities. This drawback is encountered in the use of the NCBO annotator [9], and one way to solve this problem is to have several algorithms that use knowledge-based dictionaries to transform ambiguous terms into meanings that are clear for the annotator. These algorithms should also be able to correct incorrect grammar usage and wrong spellings by matching dictionary terms with similar spellings or phrases.

Challenges of semantic annotation tools

Diverse tools are used in semantic annotation [1]. These tools also encounter some challenges, a few of which are listed below:

- Speed: This is one of the most common challenges. Annotations performed on huge datasets can take a lot of time to process.
- Language specificity: Most annotators are in English, which makes it difficult to apply semantic annotation in other languages.
- Document genre genericity: Annotators that support document input can face the problem of having to annotate different document formats, and not supporting a particular format could be a challenge.
- Text variation: According to Jovanovic and Bagheri [5], challenges are faced also due to the fact that there are different kinds of biomedical texts and variations in texts, for example between biomedical and clinical texts.
- Entity disambiguation: Entities mentioned in biomedical texts sometimes do not have enough context to disambiguate them.

These challenges and others are been studied, and many experts have tried to figure out ways to tackle them in newly developed systems. It may not be possible to fully resolve these challenges, but they can be reduced, and the following section shows how we tackled some of these challenges in the system developed for this project.

Overview of the Table2Annotation Tool

In this section, we describe the proposed solution to build an ontology matching system. Our solution uses the NCBO annotator

web service API for primary information retrieval.

The NCBO annotator annotates data with the MGrep term-to-concept matching tool and retrieves sets of annotations that are later expanded using various methods of semantic matching, meaning that this annotator goes through two stages. This annotator is unique because of the method it uses to associate concepts, instead of looking for the concept that best matches the provided context. This annotator uses BioPortal [10] and although it does not support disambiguation of terms, it is suitable for real-time processing. This annotator is available for free and is implemented through web services. This annotator is currently used in AgroPortal [11] and BioPortal.

The flow of Table2Annotation is quite simple and understandable. The system starts by taking an input dataset (CSV, Excel, etc.) and then processes the file by reading the data and fetching the necessary data to be annotated. It takes the necessary data and calls an external API provided by AgroPortal to annotate the data. The results returned from this process are processed by taking the Uniform Resource Identifier (URI), concept ID, and the matched words. Finally the annotated terms are saved and written to an output file for the user to access.

The operation of the matching system is described diagrammatically in Supplementary Fig. 1. In building this Table2Annotation tool we decided to use the NCBO annotator (AgroPortal API) to support the annotation of terms.

Important algorithms

As discussed in the challenges section, there are several problems that must be dealt with. Thus, we developed specific algorithms to handle some of them.

Threading

First of all, the system was created in a functional independent approach where the major functions are independent. For example, obtaining inputs and annotation are independent. This allows us to better handle the slower part of the system. The function that slows down the system is the one that deals with iterating through the cells, taking the cell data, and then annotating this data. To reduce the problem of speed, we decided to create an algorithm to speed up the process. The algorithm uses the concept of multi-threading, allowing the function to be run by several processors (threads) concurrently.

Permutation

As discussed above regarding the problem of discontinuous entities, although this issue has not been fully resolved and future enhancements remain to be made, the problem of conjunctions can be reduced by creating an algorithm to handle this case.

Multiple dataset formats

The problem of document genre genericity was reduced by creating an algorithm to detect the format of the file being input by the user and then handling the process depending on the file format.

Running Table2Annotation

Table2Annotation is a Java-based program that is currently executed through the command line interface. The user must have a dataset that he or she wants to annotate first. Table2Annotation is compiled after the code and all the functions explained in the previous section have been fully implemented. The compilation of Table2Annotation is done with all the necessary libraries included in the Java project. To run the system the user needs to input the following parameters: *input file* (mandatory), *column* (mandatory), *suggestions* (optional), *slice* (optional), *separator* (optional), and *sheet* (optional).

First, the user provides the path to the input file (dataset) and then provides the name of the column to be annotated. These two parameters are mandatory and the others are optional.

The other functions that can be passed as parameters are as follows: (1) suggestions (recommendations) of ontologies, allowing the user to specify which AgroPortal (or BioPortal) ontologies to use for the annotation process; (2) the slice (grouping), allowing the user to define which slice to use for the annotation process (slices can be compared to an instance of AgroPortal or BioPortal for a defined subset of ontologies); (3) the separator, if the file is a separated file type, allowing the user to define the type of separator used to split the cells; and (4) the sheet number if the file is an Excel file with multiple sheets.

After the command is executed, the system starts processing and stops when the process is completed. The results of this operation are output to a file in same format as the input file and given to the user.

Results

In this section, we describe the results obtained from the Table2Annotation tool. We also describe the context of obtaining the results and an evaluation of the system.

First of all, we needed a dataset to test the tool, as shown in Fig. 1. The dataset that we used was quite small, as using a small dataset better demonstrates how the results are obtained, but the same principles are applied when using a large dataset. The dataset contains a "*PROPERTY*" column, which contains the terms to be annotated.

https://doi.org/10.5808/GI.2020.18.2.e19

	PROPERTY	PROPERTY_id	PROPERTY_id_uri	PROPERTY_id_match
2	days to flowering	TO:0000344, CO_335:0000021	http://purl.obolibrary.org/obo/TO_0000344, http://www	days to flowering, days to flowering
3	Plant height	CO_341:0000021, CO_320:0000076, CO_321:000002(	http://www.cropontology.org/rdf/CO_341:0000021, http:	plant height, plant height, plant heigh
4	Flowering date	CO_320:0000498, SOY:0001353, C7190, 301315	http://www.cropontology.org/rdf/CO_320:0000498, http:	flowering date, flowering date, flowe
5	Grain yield 2	CO_320:0000073, CO_321:0000013, CO_322:000000	http://www.cropontology.org/rdf/CO_320:0000073, http:	grain yield, grain yield, grain yield, gra
6	harvested hill	3966, CO_324:0000165, ENVO:00000264, ENVO:000(	http://www.eionet.europa.eu/gemet/concept/3966, htt	hill, hill, hill, hill, hill, hill
7	moisture content	moisture:content, 18883, CO_336:0000195, p1524,	http://opendata.inra.fr/resources/Durum_Wheat#moist	moisture content, moisture content, n
8	PLANT SELECTION	8912	http://www.eionet.europa.eu/gemet/concept/8912	plant selection
9	Maturity	CO_336:0000021, c:4656, CO_322:0000032, PATO:00	http://www.cropontology.org/rdf/CO_336:0000021, http:	maturity, maturity, maturity, maturity,
10	moisture factor	C2786, 5332, p1524, factor	http://id.agrisemantics.org/gacs/C2786, http://www.eio	moisture, moisture, moisture, factor
11	Number of tillers	CO_320:0000963	http://www.cropontology.org/rdf/CO_320:0000963	number of tillers
12	Total hill	c:1405700091939, 3966, CO_324:0000165, ENVO:000	http://opendata.inra.fr/anaeeThes/c_1405700091939, htt	total, hill, hill, hill, hill, hill, hill
13	Panicle length	CO_320:0000033, TO:0000040, CO_324:0000094, CO	http://www.cropontology.org/rdf/CO_320:0000033, http:	panicle length, panicle length, panicle
14	Bacterial blight	CO_320:0000173, SOY:0001427, SOY:0001312, SOY:0	http://www.cropontology.org/rdf/CO_320:0000173, http:	bacterial blight, bacterial blight, bacte
15	Leaf blast	CO_325:0000068, PO:0025034, CO_336:0000188, SP:	http://www.cropontology.org/rdf/CO_325:0000068, http:	leaf, leaf, leaf, leaf, leaf, leaf, leaf, lea

Fig. 1. Dataset result without using suggestion and slice parameters. The column PROPERTY_ID shows all the matching ontologies.

	PROPERTY	PROPERTY_id	PROPERTY_id_uri	PROPERTY_id_match
2	days to flowering	TO:0000344	://purl.obolibrary.org/obo/TO_	days to flowering
3	Plant height	CO_320:0000076, CO_321:0(	http://www.cropontology.org/rdf/	plant height, plant height, plant height
4	Flowering date	CO_320:0000498	http://www.cropontology.org/rdf/	flowering date
5	Grain yield 2	CO_320:0000073, CO_321:0(	http://www.cropontology.org/rdf/	grain yield, grain yield
6	harvested hill			
7	moisture content	PATO:0000025	http://purl.obolibrary.org/obo/PA	content
8	PLANT SELECTION	T002	http://purl.bioontology.org/ontolc	plant
9	Maturity	PATO:0000261, CO_320:000(	http://purl.obolibrary.org/obo/PA	maturity, maturity
10	moisture factor			
11	Number of tillers	CO_320:0000963	http://www.cropontology.org/rdf/	number of tillers
12	Total hill			
13	Panicle length	CO_320:0000033, TO:000004	http://www.cropontology.org/rdf/	panicle length, panicle length
14	Bacterial blight	CO_320:0000173	http://www.cropontology.org/rdf/	bacterial blight
15	Leaf blast	PO:0025034, CO_320:00010(	http://purl.obolibrary.org/obo/PO	leaf, blast, blast, blast, blast

Fig. 2. Dataset result using the slice parameter which filtered out some ontologies. The column named PROPERTY_ID shows less ontology matching than the one in Fig. 1.

Test without recommendations or slices

In this test, we ran the system without giving recommendations or slice options (i.e., an ontology list to map on provided by AgroPortal), and the results are shown in Fig. 1. In the results obtained by processing, we can see that there are three new columns: "*PROPERTY_id*", "*PROPERTY_id_uri,*" and "*PROPERTY_id_match.*" The first added column contains the concept IDs obtained from the annotation, the second added column contains the URIs of the concepts, and the third added column contains the matching of the terms with the concept.

Test with a slice

In this test, we test-ran the system by giving it a slice called "*agrold*", which contains ontology groups for agronomy. The results of the test are shown in Fig. 2. In the results, we can see three terms (highlighted in yellow) that do not match with any concept, because they do not have ontologies belonging to the "*agrold*" group.

Test with recommendations

In this test, we ran the system with three suggestion parameters: "*PO (Plant Ontology)*", "*TO (Plant Trait Ontology)*", and "*PATO (Phenotypic Quality Ontology).*" The results of the test are shown in Fig. 3. In the results, we can see that six terms (highlighted in pink) had no matching concepts, because we filtered the annotation to the three ontologies given in the suggestions.

Test with a permutation algorithm

In this section, we tried to show the effect of having an algorithm to solve the problem of conjunctions in terms, which was mentioned earlier. We annotated the term "*drought and salinity tolerance*" and Fig. 4 shows the results. Fig. 4 (A, dataset result without algorithm) shows the results from the operation without the algorithm, and we can see that there are only matches for "*drought*" and "*salinity tolerance.*" Fig. 4 (B, dataset result with algorithm) shows the results from the operation with the algorithm, which yields

PROPERTY	PROPERTY_id	PROPERTY_id_uri	PROPERTY_id_match
days to flowering	TO:0000344	http://purl.obolibrary.org/obo/TO_0000344	days to flowering
Plant height	TO:0000207	http://purl.obolibrary.org/obo/TO_0000207	plant height
Flowering date			
Grain yield 2	PATO:0000170	http://purl.obolibrary.org/obo/PATO_0000170	yield
harvested hill			
moisture content	PATO:0000025	http://purl.obolibrary.org/obo/PATO_0000025	content
PLANT SELECTION			
Maturity	PATO:0000261	http://purl.obolibrary.org/obo/PATO_0000261	maturity
moisture factor			
Number of tillers	PATO:0001555	http://purl.obolibrary.org/obo/PATO_0001555	number of
Total hill			
Panicle length	TO:0000040	http://purl.obolibrary.org/obo/TO_0000040	panicle length
Bacterial blight			
Leaf blast	PO:0025034	http://purl.obolibrary.org/obo/PO_0025034	leaf

Fig. 3. Dataset result using the recommendation parameter and keeping the three best ontologies annotating the dataset. The column named PROPERTY_ID shows less ontology matching than the one in Figs. 1 and 2.

PROPERTY	PROPERTY_id	PROPERTY_id_uri	PROPERTY_id_match
A:Dataset Result without Algorithm			
drought and salinity tolerance	CO_331:0000712 CO_321:0000132	http://www.cropontology.org/rdf/CO_321:0000132 http://www.cropontology.org/rdf/CO_331:0000712	drought, salinity tolerance
B:Dataset Result with Algorithm			
drought and salinity tolerance	CO_331:0000712 CO_321:0000132 CO_346:0000036	http://www.cropontology.org/rdf/CO_321:0000132 http://www.cropontology.org/rdf/CO_331:0000712 http://www.cropontology.org/rdf/CO_346:0000036	drought, salinity tolerance, drought tolerance

Fig. 4. Comparison between splitting algorithm and without.

matches for three terms: "*drought*", "*salinity tolerance*", and "*drought tolerance.*"

Conclusion

In conclusion, Table2Annotation has strengths in certain criteria such as speed, error handling, and concept matching. First, we use a multi-threading algorithm that runs the process very effectively and efficiently. Second, it handles errors and exceptions by ignoring them whenever they occur. If there is an error while matching one term, it skips the term with an error and continues to the next one. If there is a general error, it still completes the matching process, but returns empty results. This method of error handling allows the user to run the process while multitasking and return to obtain the results without having to worry about system process terminations. Last, the matching results are good, and we see that cases of conjunctions are handled appropriately, so that the results contain more matches. The filters (slice and suggestions) also help to tailor the results to match the user's expectations.

The system has strengths, but also has some weaknesses, such as relying on an internet connection and being dependent on the API. The system uses an external API, which can cause problems. Firstly, the system cannot work offline as it needs internet access to call the external API, which could be seen as a weakness. Secondly, if the external API is down for some reason, the system cannot be used. These weaknesses can be solved by building a full annotation system that does not depend on the availability of any external annotation API.

In the future, we think that we can improve algorithms to handle grammar problems and disambiguation. These algorithms should use language dictionaries to be able to transform terms without meaning (short forms) to something understandable to improve the process of matching to concepts. For example, when an abbreviated term is encountered, there should be a dictionary to look up

https://doi.org/10.5808/GI.2020.18.2.e19

the term and return the full meaning. This will further help to reduce the problems of spelling and grammar mentioned earlier.

ORCID

Pierre Larmande: https://orcid.org/0000-0002-2923-9790
Kazim Muhammed Jibril: https://orcid.org/0000-0002-0493-4973

Authors' Contribution

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

Conflicts of Interest

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

Acknowledgments

The authors thank the ICTLab USTH for their support.

Supplementary Materials

Supplementary data can be found with this article online at http://www.genominfo.org.

References

1. Oliveira P, Rocha J. Semantic annotation tools survey. In: 2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM), 2013 Apr 16-19, Singapore. New York: Institute of Electrical and Electronics Engineers, 2013. pp. 301-307.
2. Liao Y, Lezoche M, Panetto H, Boudjlida N. Why, where and how to use semantic annotation for systems interoperability. In: 1st UNITE Doctoral Symposium, 2011 Jun, Bucarest, Romania. pp. 71-78.
3. Cooper L, Meier A, Laporte MA, Elser JL, Mungall C, Sinn BT, et al. The Planteome database: an integrated resource for reference ontologies, plant genomics and phenomics. Nucleic Acids Res 2018;46:D1168-D1180.
4. 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.
5. Jovanovic J, Bagheri E. Semantic annotation in biomedicine: the current landscape. J Biomed Semantics 2017;8:44.
6. Bossy R, Golik W, Ratkovic Z, Bessieres P, Nedellec C. BioNLP shared Task 2013: an overview of the bacteria biotope task. In: Proceedings of the BioNLP Shared Task 2013 Workshop, 2013 Aug 9, Sofia, Bulgaria. Stroudsburg: Association for Computational Linguistics, 2013. pp. 161-169.
7. Bossy R, Deleger L, Chaix E, Ba M, Nedellec C. Bacteria Biotope at BioNLP Open Shared Tasks 2019. In: Proceedings of the 5th Workshop on BioNLP Open Shared Tasks, 2019 Nov 4, Hong Kong, China. Stroudsburg: Association for Computational Linguistics, 2019. pp. 121-131.
8. Baumgartner W, Bada M, Pyysalo S, Ciosici MR, Hailu N, Pielke-Lombardo H, et al. CRAFT Shared Tasks 2019 overview: integrated structure, semantics, and coreference. In: Proceedings of the 5th Workshop on BioNLP Open Shared Tasks, 2019 Nov 4, Hong Kong, China. Stroudsburg: Association for Computational Linguistics, 2019. pp. 174-184.
9. Jonquet C, Shah NH, Musen MA. The open biomedical annotator. Summit Transl Bioinform 2009;2009:56-60.
10. Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, et al. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res 2009;37:W170-W173.
11. Jonquet C, Toulet A, Arnaud E, Aubin S, Dzale Yeumo E, et al. AgroPortal: a vocabulary and ontology repository for agronomy. Comput Electron Agric 2018;144:126-143.

G&I
Genomics & Informatics

Application note

eISSN 2234-0742
Genomics Inform 2020;18(2):e20
https://doi.org/10.5808/GI.2020.18.2.e20

Received: March 17, 2020
Revised: May 26, 2020
Accepted: May 26, 2020

*Corresponding author:
E-mail: dfsousa@lasige.di.fc.ul.pt

Improving accessibility and distinction between negative results in biomedical relation extraction

Diana Sousa*, Andre Lamurias, Francisco M. Couto

LASIGE, Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal

Accessible negative results are relevant for researchers and clinicians not only to limit their search space but also to prevent the costly re-exploration of research hypotheses. However, most biomedical relation extraction datasets do not seek to distinguish between a false and a negative relation among two biomedical entities. Furthermore, datasets created using distant supervision techniques also have some false negative relations that constitute undocumented/ unknown relations (missing from a knowledge base). We propose to improve the distinction between these concepts, by revising a subset of the relations marked as false on the phenotype-gene relations corpus and give the first steps to automatically distinguish between the false (F), negative (N), and unknown (U) results. Our work resulted in a sample of 127 manually annotated FNU relations and a weighted-F1 of 0.5609 for their automatic distinction. This work was developed during the 6th Biomedical Linked Annotation Hackathon (BLAH6).

Keywords: biomedical research, knowledge base, negative results, relation extraction
Availability: The code supporting our work and the sample of 127 manually annotated FNU relations of the PGR dataset is publicly available at https://github.com/lasigeBioTM/blah6.

Introduction

Researchers and clinicians need to have access not only to known relations between biomedical entities but also to relations that were already disproven. Accessible negative results limit their search space and prevent the costly re-exploration of research hypotheses. However, most biomedical relation extraction datasets do not seek to distinguish between a false and a negative relation among two biomedical entities, and few knowledge bases hold negative examples. Some domain-specific exceptions are worth noticing, such as the Negatome database [1] for protein-protein interactions, and the phenotype-disease relations annotation file made available by the Human Phenotype Ontology (HPO) organization [2] that contains both positive and negative relations.

A false relation should express a context where the entities are not related. In contrast, a negative relation should express a context where there is an affirmation of no association between the two entities. Furthermore, datasets created using distant supervision techniques also have some false negative relations that constitute undocumented/unknown relations [3]. These relations are not marked true because they are not described in a knowledge base at the moment of the dataset creation, even though upon reading the context of these relations within their respective sentences one can support a true relation. Unknown relations are good examples of hypotheses to be further explored by re-

searchers and clinicians and can be of use to effectively populate the biomedical relations knowledge bases.

We propose to improve the distinction between false, negative, and unknown (FNU) relations, by:

- Revising a subset of the relations marked as false on the phenotype-gene relations (PGR) corpus [4] to create a sample dataset of FNU relations (made available on PubAnnotation platform (http://pubannotation.org/collections/Annotation%20of%20Human%20Phenotype-Gene%20Relations%20-%20Identification%20of%20Negative,%20False,%20and%20Unknown%20Relations) [5])

- Implementing the first steps (using regular expressions and a neural network) to automatically distinguish between the FNU relations, using the previous sample FNU dataset as a test set.

Methodology

The PGR corpus consists of 1,712 abstracts, 5,676 human phenotype annotations, 13,835 gene annotations, and 4,283 relations [4]. This automatically annotated corpus distinguishes between false and true relations but fails to identify different types of FNU relations. Fig. 1 illustrates the levels that we considered to represent true PGR relations (true, positive, and known), and false PGR relations (false, negative, and unknown).

Previously, our team had an expert curating a subset of the PGR corpus (around 30%). These annotations were initially divided into true, and false, for a different scope out of the reach of this work.

Table 1. Distribution of each type of FNU relation: false, negative, unknown, and the total number of relations

	False	Negative	Unknown	Total
No.	73	11	43	127

Nonetheless, for this work, we used the 127 false annotations curated by our domain expert in that subset to make the distinction between false (F), negative (N), and unknown (U) relations. The distribution of each type of relation is displayed in Table 1.

Some concrete examples of what sentences constitute each type of relation are presented in Fig. 2.

The manual annotations allowed for the assessment of common patterns for the false and negative types of relations:

- False relations are often enumerations or an explanation of protocol that does not imply any type of relation.

- Negative relations are more regular, with words that imply the negation of association, such as *non, no, dissociation*, and *not*, frequently combined with *associated*, and *involved*.

Contrarily, unknown relations follow intractable patterns and are the most heterogeneous.

The first approach towards catching false and negative examples that follow the specified patterns was using regular expressions by:

- Analyzing the list of detected negative expressions and of detected false expressions and possible equivalences (for instance, for the negative expressions list, *not associated*).

- Introducing patterns that use those expressions, such as *'('+gene_entity+'|'+phenotype_entity+')(.*?)'+negative_expression+'(.*?)('+gene_entity+'|'+phenotype_entity+')'* that translates to *gene or phenotype followed by negative expression followed by gene or phenotype* (for negative examples).

- Evaluating using the manually curated dataset of 127 FNU relations (gold standard dataset) if those patterns are able to correctly classify the FNU relations.

Using regular expressions based on the annotation process can and probably will introduce a bias towards the relations that we annotated. Further applications of these regular expressions should be explored for the approach to be fully validated. Nevertheless, the creation of the regular expressions was done posterior-

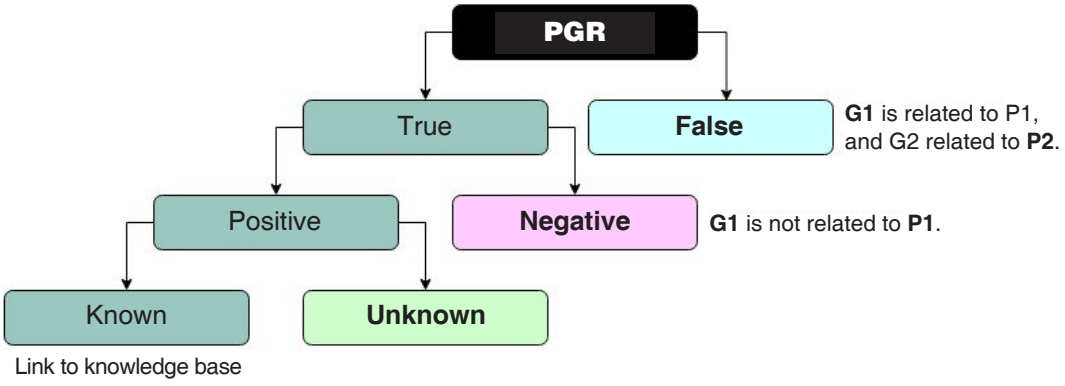

Fig. 1. Illustration of the levels that correspond to the true phenotype-gene relations (PGR) relations (true, positive, and known), and false PGR relations (false, negative, and unknown). Also, some generic sentences that elucidate the distinction between false and negative relations, and the distinction between known and unknown relations, according to the authors.

PMID:25343988

In humans, mutations in several genes involved in the Notch pathway are associated with SDV, with both **autosomal recessive** (MESP2, DLL3, LFNG, HES7) and autosomal dominant **(TBX6)** inheritance.

HP:0000007 — **false** — 6911

PMID:16960806

To date, **CRYBA4** was the only gene in this cluster not associated with either human or murine **cataracts**.

1413 — **negative** — HP:0000518

PMID:28698647

FBXL4 potentially controls **cancer** metastasis through regulation of ERLEC1 levels.

26235 — **unknown** — HP:0002664

Fig. 2. Example sentences for each type of false, negative, and unknown (FNU) relation: false (PMID:25343988), negative (PMID:16960806), and unknown (PMID:28698647). Also, the identified entities for each sentence, and their identifiers in the National Center for Biotechnology Information (NCBI) (for genes) and HPO (for human phenotypes).

ly to the annotation process, solely based on the patterns described above, with the goal of generalizing as much as possible to avoid overfitting.

As a second approach, we briefly tried to apply a neural network using the Keras library (without any tuning, due to time constraints). For this purpose, we divided the FNU dataset into a training set (70%, 89 FNU relations) and a test set (30%, 38 FNU relations).

Results and Discussion

The application of a small subset of regular expressions to catch false and negative examples that follow the previously mentioned patterns had some promising results. We opted for the unknown relation as our default label since this type of sentences are more heterogeneous with irregular patterns that are difficult to capture by the use of regular expressions. Testing against the gold standard dataset shows a weighted-F1 of 0.5609. Other relevant metrics are displayed in Table 2.

The use of the neural network produced poor results (0.2308 accuracy) mainly due to the lack of tuning and the small size of our FNU dataset.

These preliminary results show that it is possible to capture common patterns of false and negative relations with high precision, but also shows the need for more work and possible exploration of machine learning techniques in order to capture more instances of those types of relations. More manual work, building regular expressions, should boost these preliminary results. Using syntax and dependency parsing to capture complex enumerations can also boost performance (e.g., enumerations where a group of genes is associated with a phenotype A and another group of genes is related to phenotype B).

Table 2. The evaluation metrics (precision, recall, and f-measure) for the false, negative, and unknown relations, and the weighted-F1 for all classes

Type of relation	Precision	Recall	F-measure	Weighted-F1
False	0.8438	0.3699	0.5143	0.5609
Negative	0.8333	0.4545	0.5882	
Unknown	0.427	0.8837	0.5758	

Conclusions and Future Work

This work demonstrated that regular expressions are a feasible way of capturing differences between FNU relations, at least at a preliminary stage. The false and negative types of relations follow distinctive patterns that should be further explored to boost the weighted-F1 of 0.5609. Preliminary work with neural networks showed poor results (due to time constraints), but tuning the training and a larger dataset should boost these early results.

Future work could be revising all the false relations within the PGR corpus, and also of other datasets. Negative relations in manually annotated datasets should be easier to detect since the unknown relations would not be present. All of this will allow us to further explore machine learning approaches to tackle this problem more effectively.

https://doi.org/10.5808/GI.2020.18.2.e20

ORCID

Diana Sousa: https://orcid.org/0000-0003-0597-9273
Andre Lamurias: https://orcid.org/0000-0001-7965-6536
Francisco M. Couto: https://orcid.org/0000-0003-0627-1496

Authors' Contribution

Conceptualization: DS, FMC. Data curation: DS, AL. Formal analysis: DS. Funding acquisition: FMC. Methodology: DS, AL, FMC. Writing – original draft: DS. Writing – review & editing: DS, AL, FMC.

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 BLAH6. This work was also supported by FCT through funding of DeST: Deep Semantic Tagger project, ref. PTDC/CCI-BIO/28685/2017 (http://dest.rd.ciencias.ulisboa.pt/), LASIGE Research Unit, ref. UIDB/00408/2020, and PhD Scholarship, ref. SFRH/BD/145221/2019.

References

1. Blohm P, Frishman G, Smialowski P, Goebels F, Wachinger B, Ruepp A, et al. Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis. Nucleic Acids Res 2014;42:D396-D400.
2. Kohler S, Carmody L, Vasilevsky N, Jacobsen JOB, Danis D, Gourdine JP, et al. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Nucleic Acids Res 2019;47:D1018-D1027.
3. Lamurias A, Clarke LA, Couto FM. Extracting microRNA-gene relations from biomedical literature using distant supervision. PLoS One 2017;12:e0171929.
4. Sousa D, Lamurias A, Couto FM. A silver standard corpus of human phenotype-gene relations. In: The 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019 Jun 2-7, Minneapolis, MN, USA. Stroudsburg: Association for Computational Linguistics, 2019. pp. 1487-1492.
5. Kim JD, Wang Y, Fujiwara T, Okuda S, Callahan TJ, Cohen KB. Open Agile text mining for bioinformatics: the PubAnnotation ecosystem. Bioinformatics 2019;35:4372-4380.

G&I
Genomics & Informatics

Application note

eISSN 2234-0742
Genomics Inform 2020;18(2):e21
https://doi.org/10.5808/GI.2020.18.2.e21

Received: March 21, 2020
Revised: May 26, 2020
Accepted: May 26, 2020

*Corresponding author:
E-mail: fs@felipesoares.net

SciBabel: a system for crowd-sourced validation of automatic translations of scientific texts

Felipe Soares[1*], Rozane Rebechi[2], Mark Stevenson[1]

[1]Computer Science Department, University of Sheffield, Sheffield S38RA, UK
[2]Instituto de Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre 91540-000, Brazil

Scientific research is mostly published in English, regardless of the researcher's nationality. However, this growing practice impairs or hinders the comprehension of professionals who depend on the results of these studies to provide adequate care for their patients. We suggest that machine translation (MT) can be used as a way of providing useful translation for biomedical articles, even though the translation itself may not be fluent. To tackle possible mistranslation that can harm a patient, we resort to crowd-sourced validation of translations. We developed a prototype of MT validation and edition, where users can vote for that translation as valid, or suggest modifications (i.e., post-editing the MT). A glossary match system is also included, aiming at terminology consistency.

Keywords: crowdsourcing, linguistics, machine translation, medical informatics applications, PubMed
Availability: Available online under the MIT license at https://github.com/soares-f/scibabel.

Introduction

Research in the biomedical domain, particularly about treatments and procedures for humans, can help improve the patient care offered by physicians. Evidence-based medicine is based on the premise that physicians give the best care possible when they base their treatments on reliable scientific evidence. But, although in practice this access is possible, there is a limitation that makes evidence-based medicine out of the reach for many physicians: almost all of its contents are written in English.

During the first half of the 20th century, scientific research was published in a variety of languages. But, as Gordin [1] described in detail, a complex set of factors led to English becoming the language of most scientific publications following the Second World War. Researchers tend to publish in English regardless of their native language. But, while academic researchers are often proficient in English, this may not be true for physicians in non-English speaking countries.

Translation of documents into the languages with which physicians are familiar seems like an obvious way to make the world's scientific production accessible to them. But new research is produced so quickly and its results are published so rapidly that translating the information manually would be impractical. For example, in 2019 alone, more than 10,000 new articles were published in PubMed (PubMed Query: ((("2019"[Date - Publication] : "3000"[Date - Publication])) AND (treatment[Title/Abstract])) AND (procedure[Title/Abstract])) containing the keywords "treatment" and "procedure"—exactly

the kind of articles that would be of interest to physicians. However, there is a technology that could potentially do this translation automatically: machine translation (MT).

MT is a technology to render texts written in one language to another language. Modern MT research began just after the Second World War with the automatic translation of Russian scientific texts to English [2] as part of the scientific response to the Cold War (e.g., see Hutchins[3]). Machine translation research fell into decline soon thereafter due to considerable skepticism about whether practical MT systems were possible within the research community [4], but MT resurged in the 1990s with the advent of more powerful computers and alternative approaches. The field of MT experienced explosive growth after the September 2001 terrorist attacks and is an active area of scientific research [5-8]. This effort has led to a substantial improvement in the quality of translations produced by MT systems [9].

The earliest work on MT for scientific content concentrated on the physical sciences, however the focus of current research is shifting towards biomedical texts, especially due to shared tasks. This difference is important because, while users of translations in other scientific fields can tolerate some amount of error, as they do not have such a strict vocabulary and are not dealing directly with human beings, even a small mistranslation in this domain (e.g., a drug name being incorrectly translated, or a negation being ignored) could lead to disastrous consequences to patients. For example, consider Supplementary Table 1 which shows examples of a simple medical instruction (i.e., "Take two pills orally every day unless you feel dizzy or lightheaded") usually found in drugs prescriptions translated into Finnish, Korean, Portuguese, Italian, Spanish, Japanese, French, German, Russian, Chinese (simplified) and Ukrainian by Google Translate. The third column contains their translations back into English by an educated native speaker (a common method of evaluating MT, similar to an approach known as back-translation) [10]. Contraindications that have been incorrectly translated are highlighted in bold font and it can be seen that these occur in six of the 11 translations. This demonstrates the need for automatic translations to be manually checked for critical mistranslations. However, this process is time-consuming and unlikely to scale well. Therefore, we propose a crowdsourced approach to validate automatic translations of biomedical articles and develop a prototype to facilitate such task.

In the proposed system volunteers who are able to read biomedical articles in English and also in another language would check MT output for critical mistranslations and vocabulary adequacy. The purpose of this system is to guarantee that the message in the source text is correctly conveyed in the translation, even though the translated text may lack fluency. Volunteers would accept the proposed translations if they are correct and be able to make editions when appropriate (e.g., incorrect terminology). We expect that our system, named SciBabel, would allow physicians and medical staff not proficient in English to access the most recent advances in medicine, enabling them to provide their patients with better treatment. The source code is available at https://github.com/soares-f/scibabel.

Background

An illustration of the recent improvements in MT can be seen from the performance of systems reported in the biomedical track of the Conference on Machine Translation (WMT), which focuses on the translation of PubMed abstracts. Translation quality increased by around 51% (or 16 percentage points) from 2016 to 2019 for English to Spanish. In Table 1 [11-16], we show the MT performance for some language pairs for biomedical texts with dates ranging from 2013 to 2019 for selected language pairs. Note that translation quality is measured automatically using the BLEU

Table 1. Machine translation performance in biomedical article abstract translation and Cochrane reviews

Reference	Language pair	Score (%)
Neveol et al. (2013) [11]	English → French Cochrane Reviews	BLEU: 40
Neves et al. (2016) [12]	English → Portuguese	BLEU: 33.37
	English → Spanish	BLEU: 31.11
Soares and Becker (2018) [13]	English → Portuguese	BLEU: 48.51
	English → Spanish	BLEU: 37.93
Saunders et al. (2019) [14]	English → Spanish	BLEU: 48.93
Soares and Krallinger (2019) [15]	English → Portuguese	BLEU: 49.51
	English → Spanish	BLEU: 47.01
Peng et al. (2019) [16]	English → German	BLEU: 35.26
	English → French	BLEU: 38.29
	English → Chinese	BLEU: 37.09

For years 2018 and 2019, metrics refer to the WMT challenge of the respective years.

score, a common MT metric that relies on the overlapping portions of the generated translations and the manually translated text [17].

In the two most recent WMT conferences (2018 and 2019) interesting results were reported for the English/Portuguese and English/Spanish language pairs. For instance, for the English to Spanish, the number of MT-generated sentences judged by humans as better than human translations was larger than the number of human sentences judged better than MT ones.

When combining the number of times that the best MT was equally good or better than human translation for WMT19, we get an average of 73% of correct translations according to human judgment, with surprising 90% for EN/ES and 82.09% for ZH/EN. This strengthens our point that MT can indeed be used to aid dissemination of biomedical scientific content.

However, as shown in Supplementary Table 1, MT systems can make critical mistakes when considering the usage of a medicine, for instance. It has been shown in literature that even human translation is prone to errors [18]. That is why the translation and localization industry usually has a two-step (or even more) process for translation. That is, at least one additional human is involved in checking the translation already carried out (also called proof-reading) [19].

Crowdsourcing of intensive tasks is not new in science. One example can be the Folding@Home initiative [20], which was popular in the first decade of the years 2000'. This initiative consisted of crowdsourcing computational power from regular end-users (that signed to the initiative) to simulate protein folding, drug design, and molecular dynamics. Similarly, Seti@Home [21] tried to follow the same path to search for extraterrestrial life.

The crowdsourcing of manual annotation (or evaluation) was already explored by different authors [22,23]. For instance, the information retrieval (IR) shared tasks can be seen as the pioneers of human distributed annotation. Participants of IR shared tasks would blindly evaluate the participants' automatic predictions. Another example of distributed annotation is the Amazon Mechanical Turk, which pays users to manually annotate tasks. Some authors developed games [24-26] or mobile apps [27] to gather human annotation.

Regarding crowdsourcing of translations, Zaidan and Callison-Burch [28] state that collecting translations by crowdsourcing using non-professionals may lead to low-quality results. They propose the use of distance among translations and LM perplexity to score collected translations to discriminate between "good" and "bad" translations.

Ambati et al. [29] explored the challenges involved in crowdsourcing translation based on their experiments with Amazon Mechanical Turk. Their main findings regarding challenges are related to the large label space, that is, even though there is a finite number of possible translations for a single translations, there is a much larger space of acceptable sentences in the target space, but that may not be adequate or not style compliant. The second one is the small number of bilingual speakers for low-resourced languages. The third one is low quality, as most of the crowd-sourced translators are not professional linguists. Given this scenario, they proposed a framework based on phases to enhance the final quality of crowd-sourced translations. The first step of the translation is done by weak bilingual translators, translations which are revised by bilingual translators and the final step is done by monolinguals of the target language or bilinguals whose mother tongue is the target one. Considering the potential of crowd-sourced annotation, we aimed at developing a prototype of a system to enable the manual evaluation of automatic translations tailored to biomedical texts and post-edition. Our goal was to produce a simple yet usable interface to annotate translations as valid in the target language, while enabling users to make adjustments in the translation to correct possible mistakes.

Design

When idealizing such a tool, we envisioned not to provide perfect and fluent translations, since that would require a considerable effort from users. We are rather interested in finding gross and dangerous MT mistakes, the ones that could completely hinder the interpretation of the article. That is, we are interested in assuring that the translated text conveys the same original message, even though it may not sound completely fluent for a native speaker.

We can see as an example the sentence "*Nehmen Sie jeden Tag zwei Tabletten ein, es sei denn, Ihnen ist schwindelig oder benommen*" in German. The direct translation, as seen in Supplementary Table 1, is "Take two pills every day by mouth unless you feel dizzy or lightheaded." This may not sound natural, but it conveys the message that the dosage is two pills with a daily frequency and the contra-indication is if the person feels dizzy or lightheaded.

Functionalities

The following functionalities were implemented:

- Parallel visualization of the original text and the machine translated version.
- A "voting" system that allows users to flag a particular translation as correct (similar to a "like" in social media).
- An option to edit a suggested translation, allowing users to correct possible mistranslations.
- Only the last translation is available, since this is deemed to be

https://doi.org/10.5808/GI.2020.18.2.e21

the one with best quality.

- When editing a translation, a terminology lookup is available. That is, for each matched string in the source text, the suggested translation is shown.

Technical details

In our prototype we aim at providing a simple and easily upgradable interface for document validation and modification. The prototype is coded in Python 3 using the Flask microframework. Our choice of Flask is due to its simplicity regarding back-end and frontend, while being able to scale if required.

For the interface, we opted for the Bootstrap library (https://getbootstrap.com/), since it provides responsive mobile-ready frontend components. The functionalities were expanded using JQuery and Javascript.

As for the backend, we took advantage of the SQLAlchemy toolkit (https://www.sqlalchemy.org/), which is an ORM (Object Relational Mapper) that abstracts database operations. By using SQLAlchemy, we were able to make the app database agnostic. That is, the user can easily switch among the RDBMS supported by the package without needing to change several parts in the code.

Regarding the translation system behind the prototype, we used an in-house model developed with OpenNMT (https:// opennmt.net/) which is decoupled from the interface. We do not think that at this point it is extremely relevant to have an online translation system, since new articles can be batch translated overnight, for instance.

For the dictionary, we encourage the usage of UMLS, since it is a very comprehensive asset, already standardized and is available in many languages. Users can also make use of SNOMED CT available in more than one language, when compatible with licensing.

Results

We implemented our prototype following the design specified in Section 3. For such, we first created a simple interface to visualize the translated content in the source language (e.g., English in this case) and target language (e.g., French). In this first screen, bilingual users can check the translation, which is shown in column format. We also introduced a feature that allows users to hover over the source or target sentence and check which sentence it refers to on the other column of the parallel text. After checking the translation, bilingual users can flag (i.e., Like) the translation as good, or perform modifications (editing).

In Fig. 1, we show a screenshot of the article validation step. We have already included placeholders in the top bar to allow inclusion of alternative MT models as well as access to an Administrator backend which is under development.

In Fig. 2, we included a screenshot of the edition mode for the translated contents. In this view, the text is shown by sentences, with translations displayed as text boxes, such that users can perform post-edition on the suggested text. In addition, we included a glossary functionality, which can help users to guarantee terminology consistency. For this, a dictionary has to be supplied beforehand, and then a simple string matching is used to show the suggested translation. For instance, for the term "estrogen receptors", the suggested translation in French is "Récepteur des œstrogènes", while the automatic translation is "récepteurs aux œstrogènes". Although the automatic translation is not wrong, the suggested term "Récepteur des œstrogènes" is flagged in UMLS (https://www.nlm.nih.gov/research/umls/index.html, Unified Medical Language System) as preferred.

Conclusion and Further Steps

In this article, we pointed out the importance of making biomedical literature accessible to all healthcare professionals, despite the language they speak. As scientific publication, especially in biomedical sciences, has been fastly growing, manual translation of articles is an untractable approach to make such information multilingual. Thus, we argue that MT can be an alternative to alleviate such bottleneck.

However, despite the increasing performance of MT systems, some critical errors may occur when texts are translated, which can ultimately hinder patient safety. Thus, manual validation/evaluation of translations should be performed to mitigate potential risks. To enable validation to scale to several languages, we point out that crowdsourcing the effort may be a solution. Therefore, we developed a prototype of a system that can allow an easy translation validation and possible edition.

The prototype was developed using Python 3 and Flask (https://flask.palletsprojects.com/en/1.1.x/), with Bootstrap for the visual interface. A visualization and edition interface was created, and an Administrator interface is currently under development. We included visual features to help users when doing the validation or editing the text.

As future steps, we envision some important upgrades:

- Ability to export translations into TMX and TXML formats, since they are standard in the localization industry;

- Ability to flag different unit of measurements in translation (e.g., pounds to kilograms), since the numbers need to be converted accordingly;

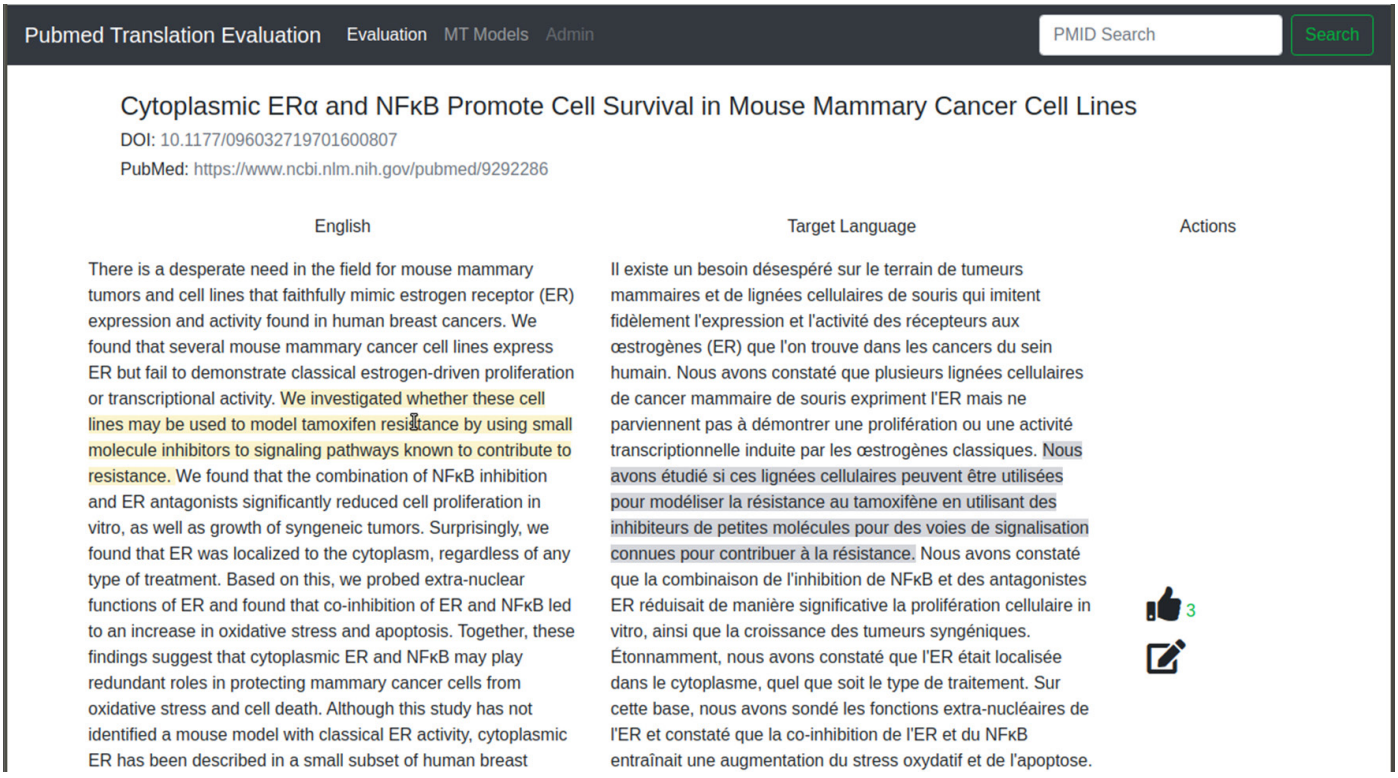

Fig. 1. Interface for translation evaluation. Users can flag the translation as adequate (i.e., Like) or edit the proposed translation using the links in the Actions column.

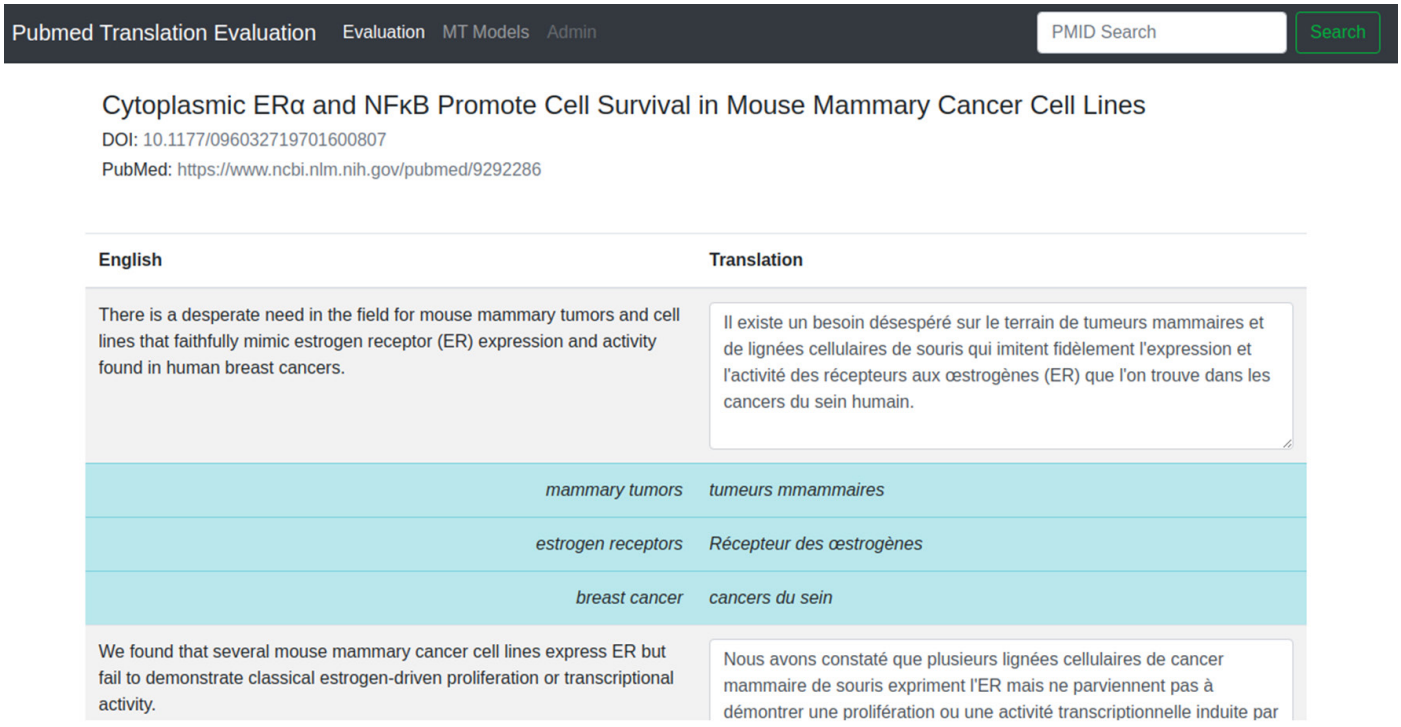

Fig. 2. Interface for translation editions Users can edit the proposed translation to make corrections on mistranslations or terminology adequacy. The prototype also shows suggested translations from terms matched in a dictionary, aiming at providing terminology consistency.

https://doi.org/10.5808/GI.2020.18.2.e21

- Include a voting scheme for rollback of manual edits and a "annotation" weight according to the mother tongue of the annotator. In addition, a similar approach for quality assurance as proposed by [29] could be used, by establishing a score for annotators as well as for annotations;

- Develop an additional view to allow annotation transfer between source and target languages.

The last upgrade, related to annotation transfer, can be extremely helpful to create multilingual annotated datasets by leveraging existing annotations in one language. For instance, one could use annotations already made in a document in English to transfer those annotations to a translated text, making annotation quicker and less expensive.

ORCID

Felipe Soares: https://orcid.org/0000-0002-2837-1853
Rozane Rebechi: https://orcid.org/0000-0002-1878-7548
Mark Stevenson: https://orcid.org/0000-0002-9483-6006

Authors' Contribution

Conceptualization: FS, RR. Data curation: RR. Formal analysis: FS. Funding acquisition: MS. Methodology: FS. Writing – original draft: FS, RR, MS. Writing – review & editing: MS.

Conflicts of Interest

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

Acknowledgments

The authors would like to acknowledge the University of Sheffield (Computer Science department) for the PhD scholarship of Felipe Soares and the ROIS-DS and DBCLS for the travel support provided.

Supplementary Materials

Supplementary data including one table can be found with this article online at http://www.genominfo.org.

References

1. Gordin MD. Scientific Babel: How Science Was Done before and after Global English. Chicago: University of Chicago Press, 2015.
2. Hutchins JW. Machine translation: a concise history. Comput Aided Transl Theor Pract 2007;13:11.
3. Hutchins JW. Early Years in Machine Translation: Memoirs and Biographies of Pioneers. Amsterdam: John Benjamins Publishing, 2000.
4. Poibeau T. The 1966 ALPAC report and its consequences. In: Machine Translation (Poibeau T, ed.). Cambridge: MIT Press, 2017. pp. 75-89.
5. Garg A, Agarwal M. Machine translation: a literature review. Preprint at http://arxiv.org/abs/1901.01122 (2018).
6. Okpor MD. Machine translation approaches: issues and challenges. Int J Comput Sci Issues 2014;11:159-165.
7. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. Preprint at http://arxiv.org/abs/1409.0473 (2014).
8. Cheng Y. Joint Training for Neural Machine Translation. Singapore: Springer Singapore, 2019. pp. 25-40.
9. Apter E. Translation-9/11: terrorism, immigration, and the world of global language politics. Global South 2007;1:69-80.
10. Rapp R. The back-translation score: automatic MT evaluation at the sentence level without reference translations. In: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers (Su KY, Su J, Wiebe J, Li H, eds.), 2009 Aug 4, Singapore. Stroudsburg: Association for Computational Linguistics, 2009. pp. 133-136.
11. Neveol A, Zweigenbaum P, Max A, Yvon F, Ivanishcheva Y, Ravaud P. Statistical machine translation of systematic reviews into French. Training 2013;15:366K.
12. Neves M, Yepes AJ, Neveol A. The scielo corpus: a parallel corpus of scientific publications for biomedicine. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 2016 May, Portoroz, Slovenia. Paris: European Language Resources Association, 2016. pp. 2942-2948.
13. Soares F, Becker K. UFRGS participation on the WMT Biomedical Translation Shared Task. Preprint at http://arxiv.org/abs/1905.01855 (2018).
14. Saunders D, Stahlberg F, Byrne B. UCAM biomedical translation at WMT19: transfer learning multi-domain ensembles. Preprint at http://arxiv.org/abs/1906.05786 (2019).
15. Soares F, Krallinger M. BSC participation in the WMT translation of biomedical abstracts. In: Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2) (Bojar O, Chatterjee R, Federmann C, Fishel M, Graham Y, Haddow B, et al., eds.), 2019 Aug 1-2, Florence, Italy. Stroudsburg: Association for Computational Linguistics, 2019. pp. 175-178.
16. Peng W, Liu J, Li L, Liu Q. Huawei's NMT systems for the WMT 2019 Biomedical Translation Task. In: Proceedings of the Fourth

Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2) (Bojar O, Chatterjee R, Federmann C, Fishel M, Graham Y, Haddow B, et al., eds.), 2019 Aug 1-2, Florence, Italy. Stroudsburg: Association for Computational Linguistics, 2019. pp. 164-168.
17. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 2002 Jul 6-12, Philadelphia, PA. Stroudsburg: Association for Computational Linguistics, 2002. pp. 311-318.
18. Daems J, Macken L, Vandepitte S. Quality as the sum of its parts: a two-step approach for the identification of translation problems and translation quality assessment for HT and MT+PE. In: Proceedings of MT Summit XIV Workshop on Post-Editing Technology and Practice (O'Brien S, Simard M, Specia L, eds.), 2013 Sep 2, Nice, France. Allschwil: European Association for Machin Translation, 2013. pp. 63-71.
19. Esselink B. A Practical Guide to Localization: Language International World Directory. Amsterdam: John Benjamins Publishing Company, 2000.
20. Beberg AL, Ensign DL, Jayachandran G, Khaliq S, Pande VS. Folding@home: lessons from eight years of volunteer distributed computing. In: 2009 IEEE International Symposium on Parallel & Distributed Processing, 2009 May 23-29, Rome, Italy. New York: Institute of Electrical and Electronics Engineers, 2009.
21. Anderson DP, Cobb J, Korpela E, Lebofsky M, Werthimer D. SETI@home: an experiment in public-resource computing. Commun ACM 2002;45:56-61.
22. Sabou M, Bontcheva K, Derczynski L, Scharl A. Corpus annotation through crowdsourcing: towards best practice guidelines. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), 2014 May, Reykjavik, Iceland. Paris: European Language Resources Association, 2014. pp. 859-866.
23. Bontcheva K, Derczynski L, Roberts I. Crowdsourcing named entity recognition and entity linking corpora. In: Handbook of Linguistic Annotation (Ide N, Pustejovsky J, eds.). Dordrecht: Springer Netherlands, 2017. pp. 449-464.
24. Jurgens D, Navigli R. It's all fun and games until someone annotates: video games with a purpose for linguistic annotation. Trans Assoc Comput Linguist 2014;2:449-464.
25. Rokicki M, Chelaru S, Zerr S, Siersdorfer S. Competitive game designs for improving the cost effectiveness of crowdsourcing. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM '14), 2014 Nov, Sanghai, China. New York: Association for Computing Machinery, 2014. pp. 1469-1478.
26. Munezero M, Kakkonen T, Sedano CI, Sutinen E, Montero CS. EmotionExpert: Facebook game for crowdsourcing annotations for emotion detection. In: 2013 IEEE International Games Innovation Conference (IGIC), 2013 Sep 23-25, Vancouver, BC, Canada. New York: Institute of Electrical and Electronics Engineers, 2013.
27. Chen N, Hoi SC, Li S, Xiao X. Mobile app tagging. In: Proceedings of the 9th ACM International Conference on Web Search and Data Mining (WSDM '16), 2016 Feb 22-25, San Francisco, CA, USA. New York: Association for Computing Machinery, 2016. pp. 63-72.
28. Zaidan OF, Callison-Burch C. Crowdsourcing translation: professional quality from non-professionals. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Vol. 1, 2011 Jun 19-24, Portland, OR, USA. Stroudsburg: Association for Computational Linguistics, 2011. pp. 1220-1229.
29. Ambati V, Vogel S, Carbonell J. Collaborative workflow for crowdsourcing translation. In: Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, 2012 Feb 11-15, Seattle, WA, USA. New York: Association for Computing Machinery, 2012. pp. 1191-1194.

https://doi.org/10.5808/GI.2020.18.2.e21

G&I
Genomics & Informatics

Application note

eISSN 2234-0742
Genomics Inform 2020;18(2):e22
https://doi.org/10.5808/GI.2020.18.2.e22

Received: March 24, 2020
Revised: May 26, 2020
Accepted: May 26, 2020

*Corresponding author:
tateisi@biosciencedbc.jp

open-japanese-mesh: assigning MeSH UIDs to Japanese medical terms via open Japanese-English glossaries

Ryota Yamada[1], Yuka Tatieisi[2*]

[1]Fuku Inc., Tokyo 113-0033, Japan
[2]National Bioscience Database Center, Japan Science and Technology Agency, Tokyo 102-8666, Japan

The Medical Subject Headings (MeSH) thesaurus is a controlled vocabulary for indexing biomedical documents that is used for document retrieval and other natural language processing purposes. However, although the oariginal English MeSH is freely available, its Japanese translation has a restricted license. We attempted to create an open alternative, and for this purpose we made a script for assigning MeSH UIDs to Japanese medical terms using Japanese-English glossaries. From the MeSpEn glossary and MEDUTX dictionary, we generated a 12,457-word Japanese-MeSH dictionary.

Keywords: dictionaries, Medical Subject Headings, natural language processing
Availability: The script is available from before the URL https://github.com/roy29fuku/open-japanese-mesh.

Introduction

The Medical Subject Headings (MeSH) [1] thesaurus is a controlled vocabulary developed and maintained by the United National Library of Medicine (NLM) that is used for indexing biomedical articles in PubMed.

MeSH is primarily used for indexing and searching the PubMed database, but it can also be used as a reliable dictionary of technical terms in the biomedical domain, as its headings and entry terms are representations of biomedical concepts approved by the NLM. Thus, MeSH is a valuable resource for natural language processing (NLP) applications. The metathesaurus in the Unified Medical Language Systems (UMLS) [2] includes translations of MeSH to several languages including Japanese.

However, although the original MeSH in English can be freely downloaded and used, the translations of MeSH in the UMLS are provided with "category 3" restrictions, which means that they cannot be incorporated into applications available outside the institution of the licensee. According to a mini-survey conducted in the 5th Biomedical Linked Annotation Hackathon (BLAH5) [3], although there are web-based dictionaries/thesauri that are freely consulted for finding MeSH UIDs or tree numbers by human readers, no dictionaries that are completely free for NLP applications are available.

Methods

MeSH consists of three types of records: descriptors (main headings), qualifiers, and supplementary concept records.

Descriptors are terms that characterize the subject matter. They are organized in a hierarchical structure based on broader/ narrower relations of concepts. Qualifiers are used with descriptors and describe an aspect of a subject denoted by the descriptor. Supplementary concept records are names of chemicals, drugs, and new concepts. Supplementary concept records are not hierarchically ordered. Instead, each supplementary concept is linked to one or more descriptors. Descriptors and supplementary concepts have a heading (representative term) and entry terms (synonyms). Each record in MeSH is accompanied by an identifier (UID).

In order to link Japanese medical terms with medical concepts in MeSH, we used two open Japanese-English bilingual glossaries. The MeSpEn English-Japanese glossary is part of the MeSpEn multilingual medical glossary developed by the Text Mining Unit (TEMU) of the Barcelona Computing Center and available under a Creative Commons Attribution 4.0 International License [4]. The MEDUTX dictionary was developed by Kitasato University aAttribution 3.0 International License [5]. The MeSpEn English-Japanese glossary has 16,756 unique Japanese terms and 10,738 unique English terms (27,668 unique pairs). The MEDUTX dictionary has 21,821 unique Japanese terms and 22,276 unique English terms (27,122 unique pairs). Merging the two dictionaries yielded a resource with 35,903 unique Japanese terms and 30,853 unique English terms (54,790 unique pairs).

We used the 2020 MeSH ASCII files for descriptors (d2020.bin) and supplementary concepts (c2020.bin) downloaded from the FTP site of the NLM on February 5, 2020. The descriptors file contained 242,205 terms (headings and entry terms) that were mapped to 29,640 concepts (UIDs) and the supplementary concepts file had 649,322 terms that were mapped onto 268,825 UIDs.

Since the Japanese-English dictionaries we used were much smaller than the MeSH vocabulary, we developed a Python script that can be applied to any Japanese-English glossary (in the form of tab-separated list of Japanese terms and corresponding English terms) and assigned the UIDs to Japanese terms where applicable, in order to be able to expand the output dictionary when more Japanese-English resources are available.

We mapped Japanese medical terms to UIDs in the process illustrated in Fig. 1. First, a Japanese term was mapped to English term(s) with Japanese-English dictionary. The English terms were normalized as follows: they were placed in lowercase, *zenkaku* (full-width, non-ASCII) characters were converted to their *hankaku* (half-width, ASCII) counterparts, Greek characters were spelled out, and Roman numerals were converted into Arabic numerals. The jaconv library [6] was used for *zenkaku-to-hankaku* normalization. The MeSH terms were also normalized, and the normalized English terms from the dictionary were matched against the normalized MeSH terms.

The Python class for Japanese-English dictionaries, MeSH data, and normalization rules were defined in order to easily incorporate new dictionaries and new normalization rules. We also investigated the effect of each type of normalization.

Results and Discussion

Without normalization of English terms, 2,838 out of 35,903 Japanese terms were mapped onto MeSH concepts (UIDs). With normalization of English terms, 12,457 Japanese terms out of 35,903 (about 34.7%) were mapped to UIDs. The contributions of each type of normalization are summarized in Table 1. The results show that case matching of the alphabet was the most effective normalization step, and the contributions of other types of normalization were small.

At least one Japanese term was assigned to 7,346 out of 298,465

Table 1. Number of terms successfully assigned MeSH UIDs according to normalization

Normalization	Example	Mapped Japanese terms
None		2,838
Lowercasing	A → a	12,406
Zenkaku-to-hankaku	A (\uFF21) →A (\u0041)	2,839
Greek-to-English	α→ alpha	2,857
Roman numerals-to-Arabic	VIII → 8	2,838
All		12,457

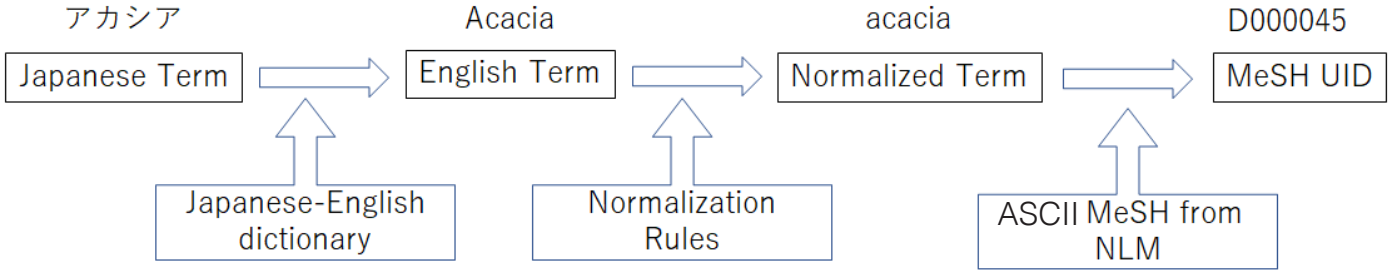

Fig. 1. The UID assignment process.

MeSH concepts (UIDs), of which 6,185 were descriptors and 1,161 were supplementary concepts. This means that Japanese terms were assigned to about 20.9% (6,185/29,640) of descriptors and 0.4% (1161/268,825) of supplementary concepts.

Considering the size of the Japanese-English dictionary (about 3% of the MeSH vocabulary) this result seems reasonable. For improving its coverage, a list of translations of names of chemicals, drugs, and other named entities regarded as supplementary concepts in MeSH should be obtained.

Conclusion

We made a script for assigning MeSH UIDs to Japanese medical terms using Japanese-English glossaries. From the MeSpEn glossary and MEDUTX dictionary, we obtained a 12,457-word Japanese-MeSH dictionary. This dictionary could be enhanced by using additional Japanese-English dictionaries. The script is available from https://github.com/roy29fuku/open-japanese-mesh under the Creative Commons Attribution 4.0 International License. Our future work includes a comparison with the Japanese translations in the UMLS metathesaurus.

ORCID

Ryota Yamada: https://orcid.org/0000-0003-2237-5025
Yuka Tatieisi: https://orcid.org/0000-0002-3813-5782

Authors' Contribution

Conceptualization: YT. Formal analysis: YT, RY. Methodology: RY. Writing – original draft: YT. Writing – review & editing: RY, YT.

Conflicts of Interest

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

References

1. Postell WD. Medicines for the Union Army. Bull Med Lib Assoc 1963;51:144-146.
2. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004;32:D267-D270.
3. Tateisi Y. Resources for assigning MeSH IDs to Japanese medical terms. Genomics Inform 2019;17:e16.
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: Proceedings of the LREC 2018 Workship "MultilingualBIO: Multilingual Biomedical Text Processing" (Melero M, Krallinger M, Gonzalez-Agirre A, eds.), 2018 May 8, Miyazaki, Japan. Paris: European Language Resources Association, 2018. pp. 32-39.
5. Asia-Pacific Association for Machine Translation. UTX glossaries. Kyoto: Asia Pacific Machine Translation Association, 2017. Accessed 2020 Mar 21. Available from: https://aamt.info/english/download/.
6. Ikegami Y. jaconv: Pure-Python Japanese character interconverter for Hiragana, Katakana, Hankaku and Zenkaku. San Francisco: GitHub Inc., 2020. Accessed 2020 Mar 21. Available from: https://github.com/ikegami-yukino/jaconv.

G&I
Genomics & Informatics

Application note

eISSN 2234-0742
Genomics Inform 2020;18(2):e23
https://doi.org/10.5808/GI.2020.18.2.e23

Received: March 16, 2020
Revised: May 6, 2020
Accepted: May 22, 2020

*Corresponding author:
kninomiya-mope@g.ecc.u-tokyo.ac.jp

Choosing preferable labels for the Japanese translation of the Human Phenotype Ontology

Kota Ninomiya[1,2*], Terue Takatsuki[3], Tatsuya Kushida[4,5], Yasunori Yamamoto[3], Soichi Ogishima[6]

[1]National Institute of Public Health, Wako 351-0197, Japan
[2]Social Cooperation Program of IT Healthcare, Graduate School of Pharmaceutical Sciences, The University of Tokyo, Tokyo 113-0033, Japan
[3]Database Center for Life Science, Research Organization of Information and Systems, Kashiwa 277-0871, Japan
[4]BioResource Research Center, RIKEN, Tsukuba 305-0074, Japan
[5]National Bioscience Database Center, Japan Science and Technology Agency, Tokyo 102-8666, Japan
[6]Advanced Research Center for Innovations in Next-Generation Medicine, Tohoku University, Sendai 980-8573, Japan

The Human Phenotype Ontology (HPO) is the de facto standard ontology to describe human phenotypes in detail, and it is actively used, particularly in the field of rare disease diagnoses. For clinicians who are not fluent in English, the HPO has been translated into many languages, and there have been four initiatives to develop Japanese translations. At the Biomedical Linked Annotation Hackathon 6 (BLAH6), a rule-based approach was attempted to determine the preferable Japanese translation for each HPO term among the candidates developed by the four approaches. The relationship between the HPO and Mammalian Phenotype translations was also investigated, with the eventual goal of harmonizing the two translations to facilitate phenotype-based comparisons of species in Japanese through cross-species phenotype matching. In order to deal with the increase in the number of HPO terms and the need for manual curation, it would be useful to have a dictionary containing word-by-word correspondences and fixed translation phrases for English word order. These considerations seem applicable to HPO localization into other languages.

Keywords: biological ontologies, natural language processing, phenotype, rare diseases, translations
Availability: As these new translations are still in progress, only an old version of Japanese translations can be obtained from https://github.com/ogishima/HPO-japanese.

Introduction

The Human Phenotype Ontology (HPO) [1] is the de facto standard ontology to describe human phenotypes. Increasingly, many researchers have been using it for accurate phenotype-driven diagnoses and translational research. As the HPO has been used by an increasingly diverse group of researchers since it was first released in 2008, its content has continued to expand. This new content includes the translation of HPO terms from English into several languages [2]. As translation into Japanese has been conducted since the BioHackathon 15 (BH15) [3], there are several Japanese translations for each English

HPO term. At the Biomedical Linked Annotation Hackathon 6 (BLAH6), an attempt was made to select preferable unique Japanese terms.

The HPO has mainly been used in the field of rare diseases as the most comprehensive resource for deep phenotyping, which is defined as "the precise and comprehensive analysis of phenotypic abnormalities in which the individual components of the phenotype are observed and described" [4]. As approximately 80% of rare diseases, the number of which is estimated to be between 5,000 and 8,000, are thought to be genetic [5,6], they may occur anywhere.

For individuals with rare diseases, delays in diagnoses and frequent misdiagnoses lead to irreversible disease progression, and mistreatment based on a misdiagnosis can even harm patients in some circumstances. This problematic journey faced by patients with rare diseases is sometimes called the "diagnostic odyssey." It has reported that it takes 5–7 years on average for patients with rare diseases to be diagnosed correctly in the UK and USA, and that patients received incorrect diagnoses two or three times [7]. In Japan, the average time to be diagnosed correctly with Fabry disease was also found to be about 20 years [8].

Therefore, HPO localization is expected to help clinicians who are not fluent in English make early diagnoses based on medical records containing standardized and detailed phenotypic information. HPO terms are being translated into Japanese, French, German, Russian, Turkish, Spanish, Italian, Dutch, Portuguese, and Chinese.

In order to understand the pathology of a specific disease, researchers often use model animals that present the same symptoms or have the same genetic abnormalities. When they choose the appropriate model animals, standardized phenotyping can be a critical clue. In Exomiser [9], phenotypic data from several species, such as mice and zebrafish, are also used for functional annotation of genetic variants from human whole-genome sequencing data. The standardized description of phenotypes by the HPO and other phenotype ontologies has enabled a phenotype-based comparison of species through cross-species phenotype matching. Harmonization of translations is also expected to make it possible for researchers to search for bio-resources for human beings or other species only using the same terms in Japanese.

At BH15, which was held in 2015, HPO terms started to be translated into Japanese. As a result of the hackathon and subsequent efforts, each HPO term had four Japanese equivalent terms, which were translated using different English-Japanese dictionaries, and the translations have been made available to the public [10].

One of the four translations is based on the Life Science Dictionary (LSD) [11], which is an English-Japanese dictionary for the life sciences; this translation is updated by researchers at Kyoto University. The second translation is based on the Japanese translation of the Mammalian Phenotype (MP) ontology [12], and was created by Riken BioResource Research Center. The third translation was created by Kenji Naritomi, a medical expert who has translated many materials about genetic diseases into Japanese. He translated the HPO terms to the extent that he could. The last translation is an automatic translation using Google Translate.

At BLAH6, a unique Japanese translation for each English term in the four translations was selected through trial and error based on the criterion that translated terms should not sound anomalous or unnatural in Japanese. Translations were prepared for the 10,668 HPO terms as of October 2017.

As the HPO describes the phenotypes of human beings and the MP describes those of mammals, they have many concepts in common. The equivalence of their concepts has already been explored by Mungall [13]. At BLAH6, the relationship between the Japanese translation of the HPO and that of the MP was examined with the goal of harmonizing them so that researchers could easily search for biological resources, using the same expression for the same phenotype. In this comparison, the Japanese translations made by Kenji Naritomi were adopted as the counterparts of the MP Japanese terms.

Methods

First, a rule-based method was used to choose the most appropriate translated terms in the following order.

1. If two or more translated terms were the same among the four translations, they were chosen as a unique Japanese term. If there were two sets of words, such that two of the four translations were the same, and the other two were the same, they were labeled as "two appropriate candidates determined by a majority." The rest of these cases were labeled as "a unique Japanese term" and "determined by a majority."

2. For the rest of the HPO terms, a morphological analysis was conducted using Mecab [14], with the MANBYO dictionary [15] as a user dictionary, for all Japanese translation candidates except those based on the LSD. Then, candidates for preferable labels were automatically chosen based on whether the morphological analysis indicated that the terms included anomalous features, defined as below. The MANBYO dictionary contains a large number of medical terms in Japanese. As some of the terms derived from the LSD are combinations of translated words, they were excluded from this analysis.

As no consensus necessarily exists regarding the precise defini-

tion of "anomalous" features, the terms were separately labeled with each feature to make it possible to change the criteria used to identify anomalous terms. The features of anomalous terms were as follows:

(1) Terms including verbs or ending with a non-noun word (e.g., 出生時にみられ時間とともに真っすぐなる大腿骨湾曲). These features seem anomalous because HPO terms are supposed to be nouns, and it is preferable for combinations to only involve nouns.

(2) Terms including particles or adjectival verbs (e.g., 尺骨の有力な茎状突起), for the same reason as (1).

(3) Terms including adjectives (e.g., 幅広い長管骨), also for the same reason as (1).

(4) Terms including Japanese commas, which appear much more unusual than English commas when they are used in terms (e.g., 異所性心臓、心臓転位).

(5) Terms including untranslated English words (e.g., 角膜 stromal 浮腫).

(6) Cases where English terms were not translated at all for unknown reasons, and the translated terms were blank.

3. In this analysis, all the anomalous features were adopted. Candidate terms were ranked in the following order:

(1) If a translated term included some strange features, it was excluded from consideration.

(2) If only one term was left after the exclusion of anomalous terms, it was chosen as the most appropriate one. Such terms were labeled as "a unique Japanese term" and "determined by an exclusion process."

(3) If more than two terms were left, it was difficult to choose which was better, and such cases were labeled as "multiple appropriate candidates determined by an exclusion process."

(4) If all of the terms were excluded, the item was labeled as "no appropriate candidates determined by an exclusion process." If all the translated terms were initially blank, they were labeled as "BLANK."

Second, an attempt was made to find out how equivalent concepts between HPO and MP are described in English and Japanese to promote the consistency of translations between these resources. As the equivalence data only contain the IDs of concepts, the English and Japanese terms were collected using the Japanese translation of the HPO [10], with HPO data as of August 2015 and July 2016, and the relationship between HPO and MP was assessed based on the MP data as of October 2012.

Results

The results of labeling all the HPO terms are shown below (Table 1).

In the second phase, the relationships between HPO and MP concepts in Japanese and English were explored, and ways to harmonize the translations were examined. A flow chart is shown below (Fig. 1).

All the HPO and MP terms referring to the same concepts were divided into the four categories described in the flow chart. As the equivalence data were created after the first translation attempt in 2015, some terms had no Japanese translation candidates. The results of a character-string comparison between them are as follows (Table 2).

Discussion

In this trial, about half of the HPO terms were found to have a unique Japanese translation. However, there are three points to consider regarding these labels.

First, those labeled as "determined by a majority" sometimes included anomalous Japanese expressions, as terms were not excluded based on anomalous features if they were identical in a majority of sources (50 percent and more). Therefore, the order of assigning labels should perhaps be reconsidered.

Table 1. Summary table of the labels assigned to all the HPO terms

Label	No.
All the HPO terms	10,668
A unique Japanese term	5,678
Determined by a majority	3,096
Determined by an exclusion process	2,687
Two appropriate candidates determined by a majority	105
Multiple appropriate candidates determined by an exclusion process	2,165
No appropriate candidates determined by an exclusion process	2,720
BLANK	5

HPO, Human Phenotype Ontology.

Table 2. Summary table of the categories assigned to HPO/MP terms with the same concepts

Category	No.
All pairs of HPO/MP terms with the same concepts	1,442
Both words are the same in both languages	219
Only the English words are the same	420
Only the Japanese words are the same	115
Both words are different in both languages	688
Japanese translation candidates do not exist yet	128

HPO, Human Phenotype Ontology; MP, Mammalian Phenotype.

https://doi.org/10.5808/GI.2020.18.2.e23

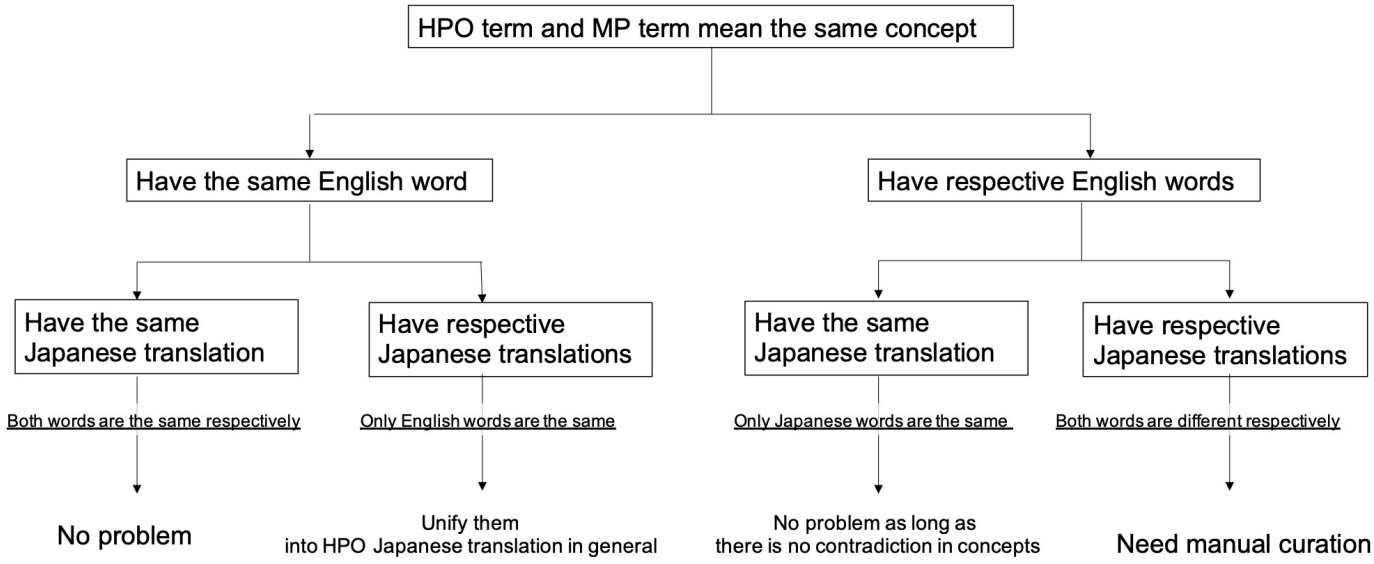

Fig. 1. Analysis of the relationships between the Human Phenotype Ontology (HPO) and Mammalian Phenotype (MP) and suggestions regarding how to harmonize these resources.

Second, HPO terms that had two appropriate candidates determined by a majority were divided into three groups, although they had the same problem as those labeled as "determined by a majority." The first group included terms with only slight differences, such as whether or not they included "症", which means "syndrome" in Japanese (e.g., 不眠|不眠症). Therefore, such typical and almost meaningless characters or words should be omitted as stop words in the next matching trial. The second group contained translations that had entirely different meanings (e.g., 硬化症|第1中足骨硬化症). In this case, one of the options must be a mistranslation. A possible reason for this is that some words in the terms were ignored in translation because the translation systems did not contain them in their dictionaries and could not recognize them properly. The last group required manual curation because the order and the selection of translated words were different (e.g., 髄様甲状腺癌|甲状腺髄様癌).

Finally, problems in Japanese translation labeling related to the exclusion process are mainly caused by the definition of anomalous features and the accuracy of the morphological analysis. Therefore, the definitions need to be made more sophisticated in future trials by adding or removing exclusion criteria. It is also important to choose an appropriate morphological analyzer for dealing with medical expressions, such as Juman++ [16] or Sudachi [17].

The relationship between the HPO and MP translations was classified into four categories according to character-string comparisons. First, if the English and Japanese terms are both the same, there is nothing to change. Second, if only the English terms are the same, the HPO translations take precedence over the MP translations, and the latter is unified to follow the former, as the former already seems to be used for more diverse purposes and to be more widespread. There are two reasons for inconsistencies in Japanese translations. One is the same as encountered for Japanese localized terms assigned the label "two appropriate candidates determined by a majority." The other is that the same terms, especially those that refer to morphological abnormalities of external body parts (instead of abnormal internal situations), are sometimes translated differently depending on the species. For example, the words "male" and "female" are "男性" and "女性" for human beings, respectively. However, for non-human mammals such as mice and rats, these terms are written as "オス" and "メス", respectively. Therefore, the principle of assigning precedence to the HPO translations is acceptable only in a general sense. Third, if only the Japanese terms are the same, there is no need to change the translation as long as the concepts are similar between the HPO and MP terms. Finally, if both the English and Japanese terms are different, there is no option other than manual curation. Since applying these principles led to the finding that roughly half of the terms need manual curation to be harmonized, another way needs to be found to decrease the necessity for manual curation in further research.

As the HPO includes technical terms, orthodox translations that are generally accepted among health professionals should be adopted. An excellent approach would seem to be to map these terms to other dictionaries for translation and to adapt their translations if doing so is permissible because other dictionaries are thought to be edited according to the same policy. This approach

seems to contribute to external consistency among dictionaries and to reinforce the stability of orthodox translations. Nonetheless, the MP translations can be candidates for replacing the HPO translations, as they sometimes contains better expressions, and a comparison between them enables harmonization and cross-species matching or searching. If translations of the terms cannot be found in other resources, or there are several translation candidates, experts need to translate them manually. Although this task requires extensive work and costs, it is ultimately unavoidable.

To deal with the increase of the number of HPO terms and the excessive dependence on manual curation—despite its inevitability in principle—it may be a good idea to develop a dictionary that contains word-by-word correspondences based on the temporarily completed translations of the HPO and MP. Such a dictionary would enable the generation of translation candidates for new terms consistent with the fixed HPO and MP translations created previously. As some word orders are common in English terms, it is possible to establish fixed Japanese phrases for each of these frequent word orders. Therefore, dictionaries and lists of fixed phrases can reduce the task of manual curation by changing it from translation of terms from scratch to only selection of the most appropriate candidates. These approaches seem to be applicable to HPO localization into other languages.

Conclusion

In this study, an attempt was made to determine a single unique translation for each term in the HPO in a rule-based way. For about half of the terms, only one appropriate Japanese word was identified, and for the rest, manual curation was needed. However, as this approach yielded insufficient accuracy, further consideration is necessary and will be given in venues such as another future hackathon.

The relationship between the HPO and MP was also investigated to evaluate the task of establishing consistency between them. Based on the analysis, the translations of both ontologies should be harmonized to improve their usability for annotating phenotypes of humans and non-human mammals.

It is possible that the number of HPO terms will continue to increase and that there will be more need for manual curation. An effective approach would seem to be to create a dictionary that contains word-by-word correspondences based on the temporary translations and fixed translation phrases for English terms in word orders that frequently appear. These approaches are most likely applicable to HPO localization into other languages.

ORCID

Kota Ninomiya: https://orcid.org/0000-0002-7381-1643
Terue Takatsuki: https://orcid.org/0000-0003-0011-764X
Tatsuya Kushida: https://orcid.org/0000-0002-0784-4113
Yasunori Yamamoto: https://orcid.org/0000-0002-6943-6887
Soichi Ogishima: https://orcid.org/0000-0001-8613-2562

Authors' Contribution

Conceptualization: KN, TT, TK, YY, SO. Data curation: KN. Formal analysis: KN. Funding acquisition: YY, SO. Methodology: KN, TT, TK. Writing – original draft: KN, TT, TK. Writing – review & editing: KN, TT, TK.

Conflicts of Interest

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

Acknowledgments

I am grateful to Yuka Tateisi, Kevin Bretonnel Cohen, and Felipe Soares for their valuable suggestions.

References

1. Kohler S, Carmody L, Vasilevsky N, Jacobsen JO, Danis D, Gourdine JP, et al. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Nucleic Acids Res 2018;47:D1018-D1027.
2. Progress of localization of Human Phenotype Ontology. Crowdin, 2020. Accessed 2020 Mar 5. Available from: https://crowdin.com/project/hpo-translation.
3. Vos RA, Katayama T, Mishima H, Kawano S, Kawashima S, Kim JD, et al. BioHackathon 2015: semantics of data for life sciences and reproducible research [version 1]. F1000Res 2020;9:136.
4. Robinson PN. Deep phenotyping for precision medicine. Hum Mutat 2012;33:777-780.
5. International Federation of Pharmaceutical Manufacturers and Associations. RARE DISEASES: shaping a future with no-one left behind. Geneva: IFPMA, 2017. Accessed 2020 Jan 25. Available from: https://www.ifpma.org/wp-content/uploads/2018/02IFPMA_Rare_Diseases_Brochure_28Feb2017_FINAL.pdf.
6. European Commission. Policy: rare diseases: what are they? Brussels: European Commission, 2016. Accessed 2020 Jan 25. Available from: http://ec.europa.eu/health/rare_diseases/poli-

https://doi.org/10.5808/GI.2020.18.2.e23

cy/index_en.htm.
7. Shire Human Genetic Technologies. Rare Disease Impact Report: Insights from Patients and the Medical Community. Lexington: Shire Human Genetic Technologies, 2013.
8. Sanofi. Conducted a "patient journey survey" to listen to the daily feelings and worries of patients with Fabry disease. Tokyo: Sanofi, 2018. Accessed 2020 Jan 25. Available from: https://www.sanofi.co.jp/-/media/Project/One-Sanofi-Web/Websites/Asia-Pacific/Sanofi-JP/Home/press-releases/PDF/2018/20181205.pdf?la=ja.
9. Robinson PN, Kohler S, Oellrich A; Sanger Mouse Genetics Project, Wang K, Mungall CJ, et al. Improved exome prioritization of disease genes through cross-species phenotype comparison. Genome Res 2013;24:340-348.
10. Japanese translation of Human Phenotype Ontology (in progress). San Francisco: GitHub, 2020. Accessed 2020 Mar 5. Available from: https://github.com/ogishima/HPO-japanese.
11. 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 and 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.
12. Japanese translation of Mammalian Phenotype. Tokyo: Riken, 2020. Accessed 2020 Mar 5. Available from: https:// ja.brc.riken.jp/lab/bpmp/ontology/ontology_mp_j.html.
13. mp_hp-align-equiv.owl. San Francisco: GitHub, 2020. Available from: https://github.com/obophenotype/upheno/blob/ master/hp-mp/mp_hp-align-equiv.owl.
14. Mecab: yet another part-of-speech and morphological analyzer. Mecab, 2006. Accessed 2020 Mar 5. Available from: http://taku910.github.io/mecab/.
15. MANBYO dictionary. Ikoma: Social Computing Laboratory, NAIST, 2019. Accessed 2020 Mar 5. Available from: http:// sociocom.jp/~data/2018-manbyo/index.html.
16. Morita H, Kawahara D, Kurohashi S. Morphological analysis for unsegmented languages using recurrent neural network language model. In: Proceedings of EMNLP 2015: Conference on Empirical Methods in Natural Language Processing, 2015 Sep 17-21, Lisbon, Portugal. Stroudsburg: Association for Computational Linguistics, 2015. pp. 2292-2297.
17. Takaoka K, Hisamoto S, Kawahara N, Sakamoto M, Uchida Y, Matsumoto Y. Sudachi: a Japanese Tokenizer for Business. Paris: European Language Resources Association, 2018.

G&I
Genomics & Informatics

Opinion

eISSN 2234-0742
Genomics Inform 2020;18(2):e24
https://doi.org/10.5808/GI.2020.18.2.e24

Received: March 29, 2020
Accepted: June 16, 2020

*Corresponding author:
dweissen@pennmedicine.upenn.edu

An empirical evaluation of electronic annotation tools for Twitter data

Davy Weissenbacher[1*], Karen O'Connor[1], Aiko T. Hiraki[2], Jin-Dong Kim[2], Graciela Gonzalez- Hernandez[1]

[1]Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA
[2]Database Center for Life Science, Research Organization of Information and Systems, Kashiwa, Chiba 277-0871, Japan

Despite a growing number of natural language processing shared-tasks dedicated to the use of Twitter data, there is currently no ad-hoc annotation tool for the purpose. During the 6th edition of Biomedical Linked Annotation Hackathon (BLAH), after a short review of 19 generic annotation tools, we adapted GATE and TextAE for annotating Twitter timelines. Although none of the tools reviewed allow the annotation of all information inherent of Twitter timelines, a few may be suitable provided the willingness by annotators to compromise on some functionality.

Keywords: annotation tool, natural language processing, social media mining

Introduction

Twitter is one of the leading social media platforms with more than 126 million daily users [1]. Twitter is now regarded by the natural language processing (NLP) community as a valuable source of information and has been the focus of a significant amount of research this last decade. An increasing number of shared-tasks have been organized utilizing data from this platform. Amongst the shared tasks for Twitter data, named entity recognition is well-represented, including the Named Entity Recognition and Linking Challenge series [2] which ran from 2013 to 2016, or the Workshop on Noisy User-generated Text series [3] which organized shared tasks from 2015 to 2017. Aside from named entity recognition, the community has extended its use of Twitter to broader tasks, such as the SemEval tracks on sentiment, opinion and abusive language classification starting in 2013 [4], or for health research with the Social Media Mining for Health (#SMM4H) running since 2016 [5]. Since more than half of tweets are not written in English, shared tasks are also utilizing corpora in various languages: the conference sur l'Apprentissage Automatique in 2017 in French [6], the Forum for Information Retrieval in 2016 in Indian [7], the Named Entity rEcognition and Linking in 2016 in Italian, a track in Arabic during SemEval 2017 and #SMM4H'20 with a task in French and Russian.

As the foundation for most shared tasks in NLP, and more generally most studies in NLP, the importance of the corpus cannot be overstated. A standardized corpus is essential for the evaluation of the competing systems. The correctness and consistency of the annotations are vital to ensure accurate results and predictions on how the systems will perform on unseen data. Moreover, with the generalization of statistical methods in NLP, annotations are also important for training the systems. Only well-defined, high-quality annotations can ensure that a machine learning-based system will be able to model dis-

criminating patterns and perform correctly on a given task.

Despite the strong interest in Twitter data and the importance of creating high quality annotated corpora, few annotation tools have been developed specifically to handle these data. The only tool that we are aware of is described in Cresci et al.'s study [8] but it is not available to the community. Annotators are therefore forced to annotate Twitter data using third-party tools such as text editors/spreadsheets or adapting generic annotation tools such as GATE [9] or brat [10]. However, Twitter data have their specificities that generic tools do not account for, e.g. tweets are, most often, unrelated and posted over time by a user, making it difficult to annotate all pieces of relevant information needed across different tweets in a user's timeline.

The three days of the Biomedical Linked Annotation Hackathon, BLAH6, was an opportunity for researchers to review existing annotation tools and evaluate their suitability for Twitter data. The four researchers involved in our project were given a real case corpus and they adapted two annotation tools, GATE and TextAE [11], to perform a predefined annotation task. We report in this study their evaluation according to predefined requirements and we discuss the functionalities that are still missing.

Annotating Twitter Data

When registering for a Twitter account, a user is invited to fill a short description and choose other users to follow. The new user is assigned a unique user ID and each tweet posted by the user is identified by a unique tweet ID. In addition, each tweet is described by metadata such as the posting time or the predicted language of the tweet. The collection of all tweets posted by a given user is called the home timeline.

The four researchers participating in our project during the hackathon were provided with 25 timelines of women that had publicly announced their pregnancy on Twitter. These timelines correspond to a total of 74,016 tweets in English, with an average of 3,000 tweets per timeline. We defined 31 annotation types relevant to these pregnancies and manually pre-annotated the 25 timelines for the event.

With no annotation tool designed for Twitter timelines, we had to adapt an existing tool for this type of data. Before the hackathon, we listed a set of requirements a tool should fulfill to be usable with Twitter timelines and we asked our four participants to evaluate two annotation tools according to those specifications. The specifications are detailed in (Table 1).

Adapting Existing Annotation Tools for Twitter

Prior to the publication of an extensive review of 78 annotation tools by Neve and Seva [12], we started a review of annotation tools for Twitter data. The inclusion criteria for our review were the availability and the ease of installation of the tools, or otherwise, a demonstration of the tool online. A tool was not easily installed when dependencies were missing, errors occurred, or external software, such as databases, needed. Among the 19 annotation tools we tested, few met the requirements we needed to perform timeline annotations. We had used the brat annotation tool for a previous project involving the annotation of PubMed Central articles; however, we found several problems with it when trying to use it for timeline annotations. Mainly, brat's user interface was not adapted to annotate adjacent tweets. We reviewed a commercial application, LightTag [13], and though it provided a clean interface and supported many of our requirements, it crashed excessively during use. It also did not allow for subcategories of entity tags and, the tool not being open source, prevented us from modifying it to fit our needs. Other tools tested did not allow for the subcategorization of entity tags, including WebAnno [14], Yedda [15], and Slate [16]. These tools also did not provide support for the normalization of entities extracted. Supplementary Table 1 and 2 summarizes our review of the 19 tools. Our review found three possible annotation tools for our project eHost [17], GATE and TextAE, as they met most of our requirements. We chose the GATE, and TextAE annotation tools for the hackathon because they were actively supported and updated regularly.

Tuning Gate for Twitter Data Annotation

GATE is an open-source toolkit developed for text annotation and automatic text processing. We used the stand-alone version of GATE to annotate Twitter timelines for prior projects [18]. Although a web-based version of GATE is available, GATE teamware [19], we compared TextAE with the stand-alone version of GATE, as we were already familiar with the tool and it was easier to install during the hackathon than the web-based version.

During the hackathon, we imported our 25 timelines and reviewed the tools with respect to our requirements. We imported a timeline as a unique document in GATE, one tweet per line. We inserted the tweet IDs and the posting dates before the text of the tweets to facilitate the annotation process, all items were separated by tabulations. Tweet IDs and dates were pre-annotated with their tags in the document. We named the file with the user ID. We could have added annotations at the timeline level (metadata), such as the gender or the place of residence of the user, by importing them as pre-annotation and inserting them at the beginning of the document in an empty span.

GATE fulfilled many of our specifications. GATE is actively

Table 1. Requirements for an annotation tool dedicated to Twitter data

Requirement	Description
Accessibility	*The annotation tool should be web-based to support for multiple annotators and to enable inter-annotator agreement calculation and disagreement resolution.* Web-based tools, such as GATE teamware or brat, make it easier to manage a team of annotators and compute the inter-annotator agreement.
Set up	*It should be easy to install, to set up the tags and the annotation schema as well as allowing changes to the schema.* Twitter data are used for various research projects, each project mining for different types of information requiring their own annotation schemas (e.g., normalizing adverse drug reaction (ADR), extracting reasons of drug non-persistence, etc.)
Efficiency	*It should load the tweets composing a timeline in less than 2 seconds and load an external dictionary for normalizing an annotation in less than 3 seconds.* A dictionary may be opened several times per tweet to normalize annotations, such as ADRs. A reading time longer than 3 seconds may significantly slow down the annotation of large corpora.
Stability	*It should not present recurrent bugs preventing or modifying the annotation process.* The tool should be actively supported. Active support would ensure the correction of such bugs.
Auto-saving	*It should periodically save the annotated document and save automatically upon closing the document or, in the absence of automatic saving, warn the annotators to save before closing.* When annotating long documents such as timelines, annotators are likely to close a document without saving, losing their annotations.
Import	*It should allow the upload of pre-annotated labels and metadata (e.g., tweet IDs or date of post). The import formats should be standard like XML or JSON. Non-standard formats, such as the XML format used in GATE, required developing conversion scripts to process new corpora.*
Stand-off annotations	*It should store the annotations in a separate file, leaving the original document intact.* Stand-off annotations are preferred because corpora may be used for different projects (e.g., timelines collected to study adverse pregnancy outcomes reused to study topics discussed during pregnancy)
Multi-level annotations	*It should allow for nested and crossing annotations.* Two annotations are nested if the span of one annotation is included in the span of the second annotation; they cross if they share a common span of text.
Annotation spans	*It should allow for annotating various levels of a timeline, the timeline itself, and the network of a Twitter user.* These levels are annotating spans of a tweet (e.g., the name of a drug), the tweet itself (e.g., the sentiment of the tweet), continuous set of tweets, i.e., an annotation spanning over multiple and adjacent tweets (e.g., all tweets posted by a user in May 2016).
Readability	*The interface should present a timeline to the annotator in a way that all annotations are easily distinguishable from each other and from the span annotated. Annotations should appear above the span annotated.* The metadata should be included in the annotation file but not visible in the timeline during annotation. Most research projects involve annotating multiple types of annotations, e.g., annotating a drug name and annotating if the drug was taken. Annotations are likely to overlap, cluttering up the document without a well-designed user interface.
Subcategories	*It should support for defined entity tags to have assignable subcategories.* For example, annotating alcohol intake, subcategories could be: intake, possible_intake, no_intake.
Normalization	*It should support the inclusion of a dictionary or ontology for normalizing the annotated entities to standardized terms.* For example, normalizing the annotated span 'sleepy' by linking it to the MedDRA preferred term 'Somnolence'.
Active learning	*It should provide a default API to plug in an external classifier implementing an active learning algorithm to assist the annotation process.* The classifier could, for example, pre-annotate the sentiments of tweets. Using active learning, it can ask an annotator to correct the labels it assigned with less certainty and retrain its model after the labels are corrected. After some iterations, the classifier should annotate most of the tweets with the correct sentiments, saving manual annotation time compared to manually annotating all tweets [20].
Multi-annotator support	*It should calculate the inter-annotator agreement and provide an interface to help adjudication.*
Export	*It should support the export of the annotations in standard formats such as JSON, TSV, XML, etc.*

supported, with its most recent release occurring on January 17, 2020. Written in Java and well documented, it is easy to setup. Pre-annotations and metadata can be imported provided that they are formatted in an XML file following a format specific to GATE. This XML format has been designed to support both nested and crossing annotations. GATE also supports subcategory annotations. Despite the large number of tweets in a timeline, GATE loads a timeline and its annotations in less than a second. It clearly marks completed annotations in the interface and offers the possibility to hide annotations when appropriate. GATE implements interfaces for active learning, but we did not use the service during the hackathon due to lack of time.

GATE appeared to be a valuable tool for annotating timelines but several drawbacks discourage us from using the tool for long term projects. There are some issues with the stability of the standalone version as GATE would crash occasionally. GATE also allows closing a file without saving the annotations and without warning the annotators. The internal XML format, specific to GATE, was difficult to work with and required the development of scripts to convert pre-annotated timelines in order to import them in GATE as well as to export the timeline annotated for further use in external applications. Whereas annotations at the tweet level were well supported, annotating timelines was only possible as pre-annotation and the built-in GATE User Interface would not allow annotators to edit these annotations. Due to the time constraint, we did not evaluate the diff tool plugin [21] to compute the inter-annotator agreement in GATE. The format of the output also made it difficult to manually perform these two tasks.

https://doi.org/10.5808/GI.2020.18.2.e24

Tuning PubAnnotation/TextAE for Twitter Data Annotation

TextAE is a web-based interface designed for corpus annotation. The interface is integrated with PubAnnotation [22], a public repository for literature annotation. For the hackathon, we chose the public version of PubAnnotation to create a private project, eliminating the need of a local installation and enabling the storage of our data in the cloud. We imported 5 timelines, representing the timelines in the same way as we did in GATE, one timeline per document, one tweet per paragraph, the document named with the user ID, and tweet IDs/posting dates inserted before the texts of the tweets.

The current versions of PubAnnotation/TextAE do not meet some of our requirements and would be too limited for our usage. However, the tools are still under development and, with the improvements scheduled, they could become standard for annotating Twitter data. An annotation project can be set up in PubAnnotation for multiple annotators as a collection, with a project created for each annotator. Annotation tags are created using a JSON configuration file. The TextAE interface allowed each tweet to be loaded as a paragraph. The annotation interface was not intuitive for all users. However, with documentation online, most annotators were readily able to access text and begin annotating within an hour. Some choices in the ergonomic design of the annotation interface were not optimal for our task and added time to the annotation process. The interface displays the annotated texts with the labels appearing on top of the text. The tool supports nested annotations. Although it is possible to add multiple annotations on the same span of text, this functionality was unstable in the version evaluated. Annotating the span in the wrong order resulted in the loss of the top-level annotation. The tool does not support crossing annotations since it uses HTML to display annotations and HTML does not allow crossing tags. TextAE could annotate a continuous set of tweets but this will require minor changes in the tool. TextAE does not currently support timeline annotations, but plans were made during the hackathon to extend the interface to add and edit this level of annotations. TextAE, combined with PubDictionaries, allows subcategories and normalization of annotations to standardized terms. Despite the size of our timelines (3,828 tweets, 418 KB, on average) and the dictionary used for testing the normalization (two million entries, 132 MB), both tools reacted within the time constraints imposed by our requirements. TextAE also provides an interface for the comparison of documents annotated by multiple annotators for disagreement resolution. TextAE was stable when annotating our timelines and, although the annotations must be manually saved, there is a warning presented to the annotator before closing. Annotations are saved in a separate file that can be exported as JSON or TSV files. Given the time limit of the hackathon, we did not test the import functionality in JSON format. An active learning API for the tool is in development and was not ready during the event.

Conclusion

The need for annotation tools dedicated to Social Media data, such as Twitter, is becoming more apparent as the interest of the NLP community is growing for this data. Since, to the best of our knowledge, there is no annotation tool dedicated to Twitter available, we evaluated during the 6th edition of the Biomedical Linked Annotation Hackathon two generic annotation tools using 25 Twitter timelines as a way to test their functionalities. After defining a catalog of requirements for an annotation tool dedicated to Twitter, we reviewed 19 tools and selected GATE and TextAE/PubAnnotation for our evaluation. Our results show that, whereas neither of them allows the annotation of all information characterizing Twitter timelines, each may be adapted for this purpose, if annotators are willing to compromise on some functionalities.

ORCID

Davy Weissenbacher: https://orcid.org/0000-0001-8331-3675
Karen O'Connor: https://orcid.org/0000-0001-7709-3813
Aiko T. Hiraki: https://orcid.org/0000-0002-7866-286X
Jin-Dong Kim: https://orcid.org/0000-0002-8877-3248
Graciela Gonzalez-Hernandez: https://orcid.org/0000-0002-6416-9556

Authors' Contribution

Conceptualization: DW, KO, JDK, GGH. Data curation: KO, ATH, DW. Formal analysis: KO, ATH, DW. Funding acquisition: GGH, JDK. Methodology: DW, KO, JDK, GGH. Writing – original draft: DW, KO. Writing – review & editing: DW, KO, GGH, JDK.

Conflicts of Interest

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

Acknowledgments

This work was supported by National Library of Medicine grant

number R01LM011176 to GG-H. The content is solely the responsibility of the authors and does not necessarily represent the official view of National Library of Medicine.

Supplementary Materials

Supplementary data can be found with this article online at http://www.genominfo.org/.

References

1. Shaban H. Twitter reveals its daily active user number for the first time. The Washington Post, 2019. Accessed 2020 Apr 30. Available from: https://www.washingtonpost.com/technology/2019/02/07/twitter-reveals-its-daily-active-user-numbers-first-time/.
2. Rizzo G, Pereira B, Varga A, van Erp M, Basave AE. Lessons learnt from the Named Entity rEcognition and Linking (NEEL) challenge series. Semantic Web J 2017;8:667-700.
3. Workshop on Noisy User-generated Text (WNUT). Stroudsburg: Association for Computational Linguistics, 2020. Accessed 2020 Apr 30. Available from: https://www.aclweb.org/anthology/venues/wnut/.
4. SemEval Portal. Stroudsburg: Association for Computational Linguistics, 2020. Accessed 2020 Apr 30. Available from: https://aclweb.org/aclwiki/SemEval_Portal.
5. Social Media Mining for Health Applications (#SMM4H). Wordpress.com, 2020. Available from: https:// healthlanguageprocessing.org/smm4h/.
6. Lopez C, Partalas I, Balikas G, Derbas N, Martin A, Reutenauer C, et al. CAp 2017 challenge: Twitter Named Entity Recognition. Preprint at https://arxiv.org/abs/1707.07568 (2017).
7. FIRE 2016 Microblog track. Information extraction from Microblogs posted during disasters. Forum for Information Retrieval Evaluation, 2016. Accessed 2020 Apr 30. Available from: https://sites.google.com/site/fire2016microblogtrack/information-extraction-from-microblogs-posted-during-disasters.
8. Cresci S, La Polla MN, Tardelli S, Tesconi M. #tweeTag: a web-based annotation tool for Twitter data. Pisa: Istituto di Informatica e Telematica, 2016.
9. General Architecture for test engineering. Sheffield: University of Sheffield, 2020. Accessed 2020 Apr 30. Available from: https://gate.ac.uk/.
10. Brat rapid annotation tool. Brat, 2020. Accessed 2020 Apr 30. Available from: https://brat.nlplab.org/index.html.
11. TextAE. TextAE, 2020. Accessed 2020 Apr 30. Available from: https://textae.pubannotation.org/.
12. Neves M, Seva J. An extensive review of tools for manual annotation of documents. Brief Bioinform 2019 Dec 15 [Epub]. Available from: https://doi.org/10.1093/bib/bbz130.
13. LightTag. LightTag, 2020. Accessed 2020 Apr 30. Available from: https://www.lighttag.io/.
14. WebAnno. WebAnno, 2020. Accessed 2020 Apr 30. Available from: https://webanno.github.io/webanno/.
15. YEDDA. San Francisco: GibHub, 2020. Accessed 2020 Apr 30. A lightweight collaborative text span annotation tool. Available from: https://github.com/jiesutd/YEDDA.
16. Slate. A super-lightweight annotation tool for experts. San Francisco: GibHub, 2020. Accessed 2020 Apr 30. Available from: https://github.com/jkkummerfeld/slate.
17. EHost. Annotation Tool: The extensible Human Oracle Suite of Tools (eHOST). San Francisco: GibHub, 2020. Accessed 2020 Apr 30. Available from: https://github.com/chrisleng/ehost.
18. Golder S, Chiuve S, Weissenbacher D, Klein A, O'Connor K, Bland M, et al. Pharmacoepidemiologic evaluation of birth defects from health-related postings in social media during pregnancy. Drug Saf 2019;42:389-400.
19. Teamware. GATE Teamware: collaborative annotation factories. Sheffield: University of Sheffield, 2020. Accessed 2020 Apr 30. Available from: https://gate.ac.uk/teamware/.
20. Kholghi M, Sitbon L, Zuccon G, Nguyen A. Active learning reduces annotation time for clinical concept extraction. Int J Med Inform 2017;106:25-31.
21. Cunningham H, Maynard D, Bontcheva K, Tablan V, Dimitrov M, Dowman M, et al. Performance evaluation of language analysers. Sheffield: University of Sheffield, 2020. Accessed 2020 Apr 30. Available from: https://gate.ac.uk/releases/gate-5.1-beta1-build3397-ALL/doc/tao/splitch10.html#sec:eval:annotation-diff.
22. PubAnnotation. Kashiwa: Database Center for Life Science, 2020. Accessed 2020 Apr 30. Available from: https://pubannotation.org/.

https://doi.org/10.5808/GI.2020.18.2.e24

Supplementary Table 1. Requirement review of annotation tools for Twitter

Annotation Tool	Accessibility	Active learning	Export	Import	Multi-annotator support	Multi-level annotation	Normalization	Set-up	Subcategories	Stability
Anafora[a]			NR	NR				NR		NR
Brat										
Djangology[b]			NR	NR				NR		NR
Doccano[a]			NR	NR				NR		NR
eHost										
GATE										
Inception[a]										NR
Lighttag										
MAE										
Slate										
Tagtog								NR		NR
WebAnno[a]										NR
Yedda										

☐ = Requirement not met ▒ = Requirement partially met ■ = Requirement Met NR = Not Reviewed

[a]Our review was done based on the software demonstration.

[b]No demonstration available, our review was done based on the documentation.

We were unable to run 3 of the 19 tools due to installation errors or dependency issues, Argo, Callisto, and Knowtator. Three others, Pubtator, BioQRator, and ezTag, required a specific input file format, such as BioC, and therefore were not suited to annotate tweets. In Supplementary Table 1, we summarize our review of the features of the remaining 13 annotation tools for the most important requirements in our catalog. Note, as we were just reviewing the features of the tools, we did not complete full installations for the tools with external dependencies such as server and/or database installations to run. For those tools, we examined the online demonstrations if available during our assessment.

Supplementary Table 2. Annotation tools reviewed

Annotation tool	URL
Anafora	https://github.com/weitechen/anafora
Argo	http://argo.nactem.ac.uk
BioQRator	http://www.bioqrator.org
Brat	https://brat.nlplab.org
Callisto	https://mitre.github.io/callisto
Djangology	https://sourceforge.net/projects/djangology
Doccano	http://doccano.herokuapp.com
eHost	https://github.com/chrisleng/ehost
ezTag	https://eztag.bioqrator.org
GATE	https://gate.ac.uk/teamware
Inception	https://inception-project.github.io
Knowtator	https://protegewiki.stanford.edu/wiki/Knowtator
Lighttag	https://www.lighttag.io
MAE	http://keighrim.github.io/mae-annotation
Pubtator	https://www.ncbi.nlm.nih.gov/research/pubtator
Slate	https://github.com/jkkummerfeld/slate
Tagtog	https://www.tagtog.net
WebAnno	https://webanno.github.io/webanno
Yedda	https://github.com/jiesutd/YEDDA

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, 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 manuscripts include the following: Title Page, Abstract & Keywords, Introduction, Main Text, References, Tables, and Figure Legends.

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.

Research communications

Research communication (RC) intends to deliver significant scientific discovery with broad interest in a short format. RCs may contain unstructured main text that includes introduction, results and discussion. RCs typically have no more than 2 display items (figures and tables) and the main text (not including abstract, references, tables and figure legends) is limited to 1,500 words. RCs may have online supplementary section.

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 No. 806, 193 Mallijae-ro, Jung-gu, Seoul 04501, 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.

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-

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. Data are available for 5 years at a third-party website (link to be included).	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. Information regarding submitting proposals and accessing data may be found at (link to be provided).	Not applicable

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

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
Room No. 806, 193 Mallijae-ro, Jung-gu, Seoul 04501, 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.