Text mining genome-wide association study literature

Thomas Rowlands and Tim Beck

University of Leicester, University Road, Leicester, LE1 7RH, UK

Abstract

Genome-wide association studies (GWAS) have improved our understanding of disease aetiology by identifying genetic variants associated with complex human traits and disease phenotypes. The GWAS Central resource enables broad and convenient integrative visualisation of, and access to, summary-level GWAS data. GWAS reported in the biomedical literature are manually curated and imported into the database. To accelerate the import process, at a time when the size of GWAS are increasing, we are using natural language processing to standardise publication full text and tables, and text mining to extract GWAS data. We are also developing an annotated full-text GWAS corpus to evaluate the accuracy of the text mining algorithm.

Keywords

Genome-wide association studies, GWAS Central, text mining, bio-ontologies

1. Introduction

Health research is advanced through genome-wide association studies (GWAS) which provide a deeper understanding of disease aetiology by detecting associations between genetic markers and disease phenotypes in population samples. GWAS Central (www.gwascentral.org) is a comprehensive collection of summary-level GWAS data [1]. Genetic marker data are standardised with dbSNP identifiers and phenotypes are annotated using Medical Subject Headings (MeSH) and Human Phenotype Ontology (HPO) terms. Building on the rich semantic phenotype annotation layer, a core subset of GWAS data is made available as RDF nanopublications [2]. Recently, there has been a large increase in the amount of data reported by individual studies that investigate genetic associations with lipidome, metabolome and microbiome phenotypes. This, in turn, requires efforts from biocurators to interpret and extract increased amounts of association data from publications. As the size of GWAS increase, requiring ever larger datasets to be extracted from the literature and imported into databases, there is a need for new bioinformatics capabilities to support scalable data curation.

We have developed a natural language processing (NLP) toolkit to extract GWAS entities and relations from the scientific literature. The toolkit links phenotype entities with bio-ontology terms and uses the text mining BioC standard format for sharing text documents and annotations. The toolkit can be used to process new 'unseen' publications as well as publications that have previously been seen and curated. Entities that have previously been manually extracted by database curators can be mapped back on to publications to generate new annotated corpora in the BioC standard. Here, we present *Auto-CORPus* for standardising publication texts and tables, *GWAS-Tagger* for annotating curated GWAS publications, and *GWAS-Miner* for extracting GWAS entities and relations from unseen publications.

2. Implementation of a natural language processing toolkit

2.1. *Auto-CORPus*: publication text and table standardisation

The Automated pipeline for Consistent Outputs from Research Publications (Auto-CORPus) package [3] converts publication HTML files to the BioC standard. The publication full-text is

SWAT4HCLS 2023: The 14th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences, February 13–16, 2023, Basel, Switzerland

EMAIL: tr142@leicester.ac.uk (A. 1); timbeck@leicester.ac.uk (A. 2)

ORCID: 0000-0002-7912-4203 (A. 1); 0000-0002-0292-7972 (A. 2)

converted to the BioC JSON format with each publication section (e.g., introduction, methods, results) annotated using the Information Artifact Ontology (IAO). A directed graph formed from 21,849 section headers from 2,441 open access publications is used by a section prediction algorithm to assign the IAO classification. Since BioC does not provide native support for tabular data, the publication tables are converted to a BioC compliant table-JSON structure. All abbreviations declared within a publication are converted to a JSON output that relates each abbreviation with the full definition. The abbreviation collection supports text mining tasks such as named entity recognition (NER) by including abbreviations unique to individual publications that are not contained within bio-ontologies.

2.2. *GWAS-Tagger*: annotation of seen publications

The GWAS-Tagger application searches BioC full-text and table-JSON files for GWAS entities and relations that have previously been extracted from a publication because of, for example, manual biocuration or text mining. Entities and relations are written to the BioC and table-JSON files using the BioC annotation format. The following GWAS entities, and relations between them, can be provided to GWAS-Tagger to find within publications:

- Genetic marker given as a dbSNP identifier.
- Phenotype given as a MeSH or HPO term. Synonyms and variations of text such as hyphenation, roman numeral usage and plurals are used if an exact term match is not found.
- P-value association of the genetic marker and phenotype given as a number with optional exponential notation.

2.3. *GWAS-Miner*: text mining of unseen publications

GWAS-Miner is a hybrid-based NLP application that uses the SciSpaCy data model for procedures such as tokenisation, part-of-sentence tagging, dependency parsing, together with rules for NER using a combination of bio-ontologies, vocabularies and pattern matching. BioC full-text files are processed to identify key sentences which include relevant data. Sentence structure and context are analysed using a dependency tree alongside the use of shortest dependency path to isolate relationships. GWAS-Miner is designed to be easily updated with new ontology data as soon as ontologies are updated.

3. Discussion and future work

We have used GWAS-Tagger to annotate 700 publications that have been curated by GWAS Central. We are manually evaluating these annotations, and editing them where necessary, to build a gold standard full-text GWAS corpus. The corpus will be used to assess the precision, recall and F1 score of GWAS-Miner. The supplementary materials of GWAS publications can contain large amounts of GWAS data, so we will extend Auto-CORPus to process supplementary files. Converting heterogenous supplementary formats to the BioC standard will optimise the text and table contents for text mining. Finally, we will incorporate GWAS-Miner into the GWAS Central data import pipeline.

4. References

[1] T. Beck, T. Rowlands, T. Shorter, and A. J. Brookes, GWAS Central: an expanding resource for finding and visualising genotype and phenotype data from genome-wide association studies. Nucleic Acids Res. (2022) doi: 10.1093/nar/gkac1017.

[2] T. Beck, R. C. Free, G. A. Thorisson, and A. J. Brookes, Semantically enabling a genome-wide association study database. J Biomed Semantics. (2012) 3(1):9. doi: 10.1186/2041-1480-3-9.

[3] T. Beck, T. Shorter, Y. Hu, Z. Li, S. Sun, C. M. Popovici, N. A. R. McQuibban, F. Makraduli, C. S. Yeung, T. Rowlands, and J. M. Posma, Auto-CORPus: A Natural Language Processing Tool for Standardizing and Reusing Biomedical Literature. Front Digit Health. (2022) 4:788124. doi: 10.3389/fdgth.2022.788124.