Auto-CORPus: Automated and Consistent Outputs from Research Publications

Yan Hu[1,a], **Shujian Sun**[1,a], **Thomas Rowlands**[2], **Tim Beck**[2,3,b], **and Joram M. Posma**[1,3,b]

[1]Section of Bioinformatics, Division of Systems Medicine, Department of Metabolism, Digestion and Reproduction, Imperial College London, SW7 2AZ, United Kingdom
[2]Department of Genetics and Genome Biology, University of Leicester, LE1 7RH, United Kingdom
[3]Health Data Research (HDR) UK, United Kingdom
[a]These authors contributed equally.
[b]These authors contributed equally. ✉

Abstract

Motivation: The availability of improved natural language processing (NLP) algorithms and models enable researchers to analyse larger corpora using open source tools. Text mining of biomedical literature is one area for which NLP has been used in recent years with large untapped potential. However, in order to generate corpora that can be analyzed using machine learning NLP algorithms, these need to be standardized. Summarizing data from literature to be stored into databases typically requires manual curation, especially for extracting data from result tables.

Results: We present here an automated pipeline that cleans HTML files from biomedical literature. The output is a single JSON file that contains the text for each section, table data in machine-readable format and lists of phenotypes and abbreviations found in the article. We analyzed a total of 2,441 Open Access articles from PubMed Central, from both Genome-Wide and Metabolome-Wide Association Studies, and developed a model to standardize the section headers based on the Information Artifact Ontology. Extraction of table data was developed on PubMed articles and fine-tuned using the equivalent publisher versions.

Availability: The Auto-CORPus package is freely available with detailed instructions from Github at https://github.com/jmp111/AutoCORPus/.

information artefact ontology | natural language processing | text standardization

Correspondence: *timbeck [at] leicester.ac.uk and jmp111 [at] ic.ac.uk*

Introduction

Natural language processing (NLP) is a branch of artificial intelligence that uses computers to process, understand and use human language. NLP is applied in many different fields including language modelling, speech recognition, text mining and translation systems. In the biomedical realm, NLP has been applied to extract for example medication data from electronic health records and patient clinical history from clinical notes, to significantly speed up processes that would otherwise be extracted manually by experts (1). Biomedical publications, unlike structured electronic health records, are semi-structured and this makes it difficult to extract and integrate the relevant information (2). The format of research articles differs between publishers and sections describing the same entity, for example statistical methods, can be found in different locations in the document in different publications. Both unstructured text and semi-structured document elements, such as headings, main texts and tables, can contain important information that can be extracted using text mining (3).

The development of the genome-wide association study (GWAS) has been led to by the on-going revolution in high-throughput genomic screening and a deeper understanding of the relationship between genetic variations and diseases/traits (4). In a typical GWAS, researchers collect data from study participants, use single nucleotide polymorphism (SNP) arrays to detect the common variants among participants, and conduct statistical tests to determine if the association between the variants and traits is significant. The results are mostly represented in publication tables, but can also be found in the main text, and there are multiple community efforts to store these reported associations in queryable, online databases (5, 6). These efforts involve time-intensive and costly manual data curation to transcribe results from the publications, and supplementary information, into databases. Summary-level GWAS results are generally reported in the literature according to community norms (e.g. a SNP associated to a phenotype with a probability value), hence NLP algorithms can be trained to recognize the formats in which data are reported to facilitate faster and scalable information extraction that is less prone to human error.

Development of effective automatic text mining algorithms for GWAS literature can also potentially benefit other fields in biomedical research as the body of biomedical literature grows every day. Yet previous attempts of mining scientific literature focused mainly on information extraction from abstracts and some on the main text, while for the most part ignoring tables. To facilitate the process of preparing a corpus for NLP tasks such as named-entity recognition (NER), text classification or relationship extraction, we have developed an **Auto**mated pipeline for **C**onsistent **O**utputs from **R**esearch **Pu**blication**s** (Auto-CORPus) as a Python package. The main aims of Auto-CORPus are:

- To provide clean text outputs for each publication section with standardized section names

- To represent each publication's tables in a JavaScript Object Notation (JSON) format to facilitate data import into databases

- To use the text outputs to find abbreviations used in the text

We exemplify the package on a corpus of 1,200 Open Access GWAS publications whose data have been manually added to the GWAS Central database to list phenotypes, SNPs and *P*-values found in the cleaned text (Figure 1). In addition, we also include data on 1,200+ Metabolome-Wide Association Studies (MWAS) to ensure the methods are not biased towards one domain. MWAS focus on small molecules, some of which are end-products of cellular regulatory processes, that are the response of the human body to genetic or environmental variations (7).

Materials and Methods

Data. Hypertext Markup Language (HTML) files for 1,200 Open Access GWAS publications whose data exists in the GWAS Central database (5) were downloaded from PubMed Central (PMC) in March 2020. A further 1,241 Open Access publications of MWAS on cancer, gastrointestinal diseases, metabolic syndrome, sepsis and neurodegenerative, psychiatric, and brain illnesses were also downloaded in the same format. Publisher versions of ca. 10% of these publications were downloaded in July 2020 to test the algorithms on publications with different HTML formats. The GWAS dataset was randomly divided into 700 training publications to develop algorithms, and a test set of the remaining 500 publications.

Processing. HTML files were loaded using the Beautifulsoup4 HTML parser package (v4.9.0). Beautifulsoup4 was used to convert HTML files to tree-like structures with each branch representing a HTML section and each leaf a HTML element. After HTML files were loaded, all superscripts, subscripts, and italics were converted to plain text. Auto-CORPus extracts *h1*, *h2* and *h3* tags for titles and headings, and *p* tags for paragraph texts using the default configuration. The headings and paragraphs are saved in a structured JavaScript Object Notation (JSON) file for each HTML file.

Tables are extracted from the document using a different set of configuration files (separate configurations for different table structures can be defined and used) and saved in a new JSON model that ensures tables of all formats and origin, not only restricted to GWAS publications, can be described in the same structured model, so that these can be used as input to rule-based or deep learning algorithms for data extraction. The data cells are stored in the "result" key, and their corresponding section name and header names are stored in "section_name" and "columns" keys respectively. Therefore, extracting relationships between cells only requires simple rules.

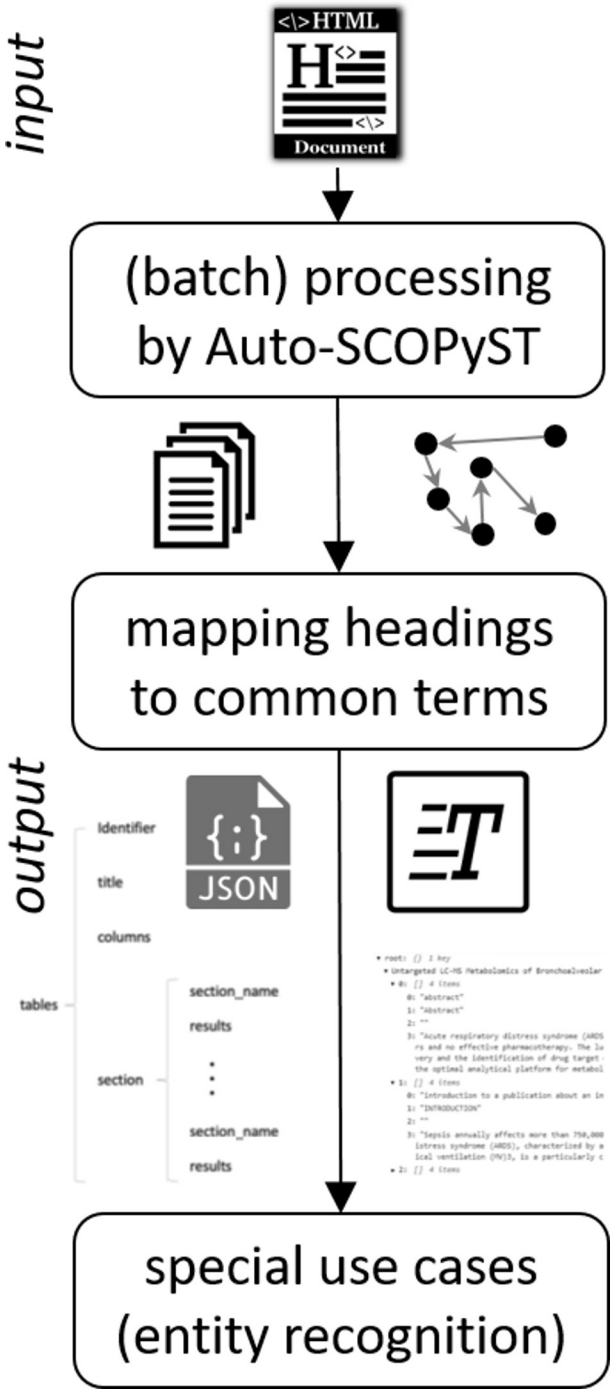

Fig. 1. Workflow of the Auto-CORPus package.

Ontologies for entity recognition. The Information Artifact Ontology (IAO) was created to serve as a domain-neutral resource for the representation of types of information content entities such as documents, databases, and digital images (8). We used the v2020-06-10 model (9) in which 37 different terms exist that describe headers typically found in biomedical literature. The extracted headers in the JSON file were first mapped to the IAO terms using the Lexical OWL Ontology Matcher (10). We use fuzzy matching using the fuzzywuzzy package (v0.17.0) to map headers to the preferred section header terms and synonyms, with a similarity threshold of 0.8. This threshold was evaluated by confirming all matches were accurate by two independent researchers.

After the direct IAO mapping and fuzzy matching, unmapped headers still exist. To map these headings, we developed a new method using a directed graph (digraph) for representation since headers are not repeated within a document, are sequential and have a set order that can be exploited. Digraphs consist of nodes (entities, headers) and edges (links between nodes) and the weight of the nodes and edges is proportional to the number of publications in which these are found. While digraphs from individual publications are acyclic, the combined graph can contain cycles hence digraphs opposed to directed acyclic graphs are used. Unmapped headers are assigned a section based on the digraph and the headers in the publication that could be mapped (anchor points). For example, at this point in this article the main headers are 'abstract' followed by 'introduction' and 'materials and methods' that could make up a digraph. Another article with headers 'abstract', 'background' and 'materials and methods' has two anchor points that match the digraph, and the unmapped header ('background') can be inferred from appearing in between the anchor points in the digraph ('abstract', 'materials and methods'): 'introduction'. We use this process to evaluate new potential synonyms for existing terms and identify new potential terms for sections found in biomedical literature.

We used the Human Phenotype Ontology (HPO) to identify disease traits in the full texts. The HPO was developed with the goal to cover all common phenotypic abnormalities in human monogenic diseases (11).

Use cases: regular expression algorithms. Abbreviations in the full text are found using an adaptation of a previously published methodology (12) based on regular expressions using the abbreviations package (v0.2.5). The brief principle of it is to find all brackets within a corpus. If the number of words in a bracket is <3 it considers if it could be an abbreviation. It searches the characters within the brackets in the text on either side of the brackets one by one. The first character of one of these words must contain the first character within that bracket. And the other characters within that bracket must be contained by other words followed by the previous word whose first character is the same as the first character in that bracket. We combine the output of the package with abbreviations defined in the abbreviations section (if found) from the IAO/digraph model.

For phenotype entity recognition, first any abbreviations in paragraphs extracted from the full text are replaced by their definition. This text is then tokenized using the spacy package (v2.3) (model *en_core_web_sm*) and compared against phenotypes and their synonyms defined by HPO for disease traits matching.

P-values and SNPs were identified in the full text and tables based on regular expressions as they have a standard form. Pairs of P-value-SNP associations are found in the text using dependency parse trees (13).

Use cases: deep learning-based named-entity recognition. The first example of a use case is to recognize the assay with which the data was acquired, however no existing models exist for this purpose. We fine-tuned a pre-existing model trained for biomedical NER, the biomedical Bidirectional Encoder Representations from Transformers (bioBERT) (14), using part of our corpus where only MWAS assays were tagged. We applied our fine-tuned model only on the paragraphs in the materials and methods sections to recognize the assays used. A second bioBERT-based model was fine-tuned on phenotypes, which already exist in the data, and enriched in phenotypes associated with the MWAS publications. This model was applied on only the abstract and paragraphs from the results section. The third example was applied only on paragraphs from the results and discussion sections using an existing model specifically trained to recognize chemical entities, ChemListem (v0.1.0) (15).

Use cases: paragraph classification. It is possible unmapped headers are mapped to multiple sections if the anchor points are far apart. In order to test the applicability of a machine learning model to classify paragraphs we trained a random forest classifier on a dataset consisting of 1,242 abstract paragraphs and 936 non-abstract paragraphs. 80% of the data was used for training and the remainder as the test set.

Results

The order of sections in biomedical literature. A total of 21,849 headers were extracted from the 2,441 publications, mapped to IAO (v2020-06-10) terms and visualized by means of a digraph with 372 unique nodes and 806 directed edges (Figure 2A). The major unmapped node is 'associated data', which is a header specific for PMC articles that appears at the beginning of each article before the abstract. The main structure of biomedical articles that were analyzed is: abstract → introduction → materials → results → discussion → conclusion → acknowledgements → footnotes section → references. IAO has separate definitions for 'materials' (IAO:0000633), 'methods' (IAO:0000317) and 'statistical methods' (IAO:0000644) sections, hence they are separate nodes in the graph and introduction is also often followed by headers to reflect the methods section (and synonyms). There is also a major directed edge from introduction directly to results, with materials and methods placed after the discussion and/or conclusion sections.

All unmapped headers were investigated and evaluated whether some could be used as synonym for existing categories. The digraph was also inspected by means of visualizing individual ego-networks which show the edges around a specific node mapped to an existing IAO term. Figure 2B shows the ego-network for abstract, and four main categories and one potential new synonym (precis, in red) were identified. The majority of unmapped headers (in purple), that follow the abstract, relate to a document that is written as one coherent whole, with specific headers for each section or a general header for the full/main text. An additional four unmapped headers relate to 'materials and methods' in their broader sense and these are data, data description, participants and sample. The remaining two categories of unmapped headers to/from abstract can be classified as new sections 'graphical abstract' and 'highlights'. These headers were found alongside, and appear to be distinct from, the (textual) abstract.

Based on the digraph, we then assigned data and data description to be synonyms of the materials section, and participants and sample as a new category termed 'participants' which is related to, but deemed distinct from, the existing patients section (IAO:0000635). The same process was applied to ego-networks from other nodes linked to existing IAO terms to add additional synonyms to simplify the digraph. Figure 2C shows the resulting digraph with only existing and newly proposed section terms.

New proposed elements for the IAO. Each existing IAO term contains one or more synonyms and extracted headers were first mapped directly to these terms. Any headers that could not be mapped directly are mapped in the second step using fuzzy matching (e.g. the typographical error 'experemintal section' in PMC4286171 is correctly mapped to the methods section). The last step involves mapping remaining unmapped headers to existing terms based on the digraph and using the structure (anchor headers) of the publication. Headers that can be mapped to existing terms in the second and third steps, are included as synonyms in the model. The existing categories for which new potential synonyms were identified are listed in Table 1a and 1b with their existing synonyms and newly identified synonyms.

From the analysis of ego-networks four new potential categories were identified: disclosure, graphical abstract, highlights and participants. Table 2 details the proposed definition and synonyms for these categories. In the digraph in Figure 2C this section is located towards the end of a publication and in some instances is followed by the conflict of interest section.

Table data extraction with different configurations. PMC articles are standardized which makes data extraction more straightforward, however some publications are not deposited into PMC or other repositories and can only be found via publisher websites. While the package has been developed using a large set of PMC articles, we compared the Auto-CORPus output for PMC articles with the output for the equivalent articles made available by the publishers. We found no differences in how headers were extracted and paragraphs were classified based on the digraph. However, the representation of tables does differ substantially between publishers, hence a model developed on PMC articles alone will fail to extract the data. We circumvent this issue by defining configuration files for different table formats and we compare the accuracy of the data represented in the JSON format (Figure 3) between PMC and publisher versions of the same papers.

Using the default (PMC) configuration on non-PMC articles none of the 302 tables are represented accurately in the JSON. Auto-CORPus allows to use a variety of configuration files (a single file, or all as batch) to be used to extract data from tables. One configuration file, different to the default, correctly represented the data in JSON format of 93% (280) of tables. The remaining 22 tables could be represented correctly using 8 different configuration files. When the right configuration file is used for non-PMC articles, all tables (100%) are represented identically to the JSON output from the matching PMC version.

Use cases. The extracted paragraphs were classified as one (or more) categories based on the digraph. This is the purpose of the Auto-CORPus package, to prepare a corpus for analysis so that different sections can be used for specific purposes. We detail how these standardized texts can be used for entity recognition.

Paragraph classification. While many headers can be mapped using fuzzy matching plus the digraph structure, some headers remain unmapped (e.g. the headers in purple in Figure 2B: full text, main text, etc.) while others can be assigned to multiple (possible) sections. The choice of assigning multiple categories to unmapped headers based on the digraph is deliberate as it is to ensure the algorithm does not wrongly assign it to only one (e.g. 'materials' over 'methods'). The next step is to perform the paragraph classification using NLP algorithms to learn from the word usage and context. We show that random forests can be used to this end by training it to distinguish between abstracts and other paragraphs. 435 paragraphs from the test set were predicted using a random forest trained on 1,743 paragraphs. For the test set, we obtained an F1-score of 0.90 for classifying abstracts (precision = 0.91, recall = 0.90) and 0.88 for classifying non-abstracts (precision = 0.87, recall = 0.88).

Abbreviation identification. The abbreviation detection algorithm searches through each paragraph using a rule-based approach to find all abbreviations used. Auto-CORPus then investigates whether a paragraph is mapped to the abbreviations category and, if found, it combines these two lists of abbreviations found in the publication. For example, when applied on an MWAS publication (16) which contains a header titled "ABBREVIATIONS" the algorithm combines the 9 abbreviations listed by the authors and with a further 7 identified from the text (Figure 4), including an abbreviation used with two spellings in the text.

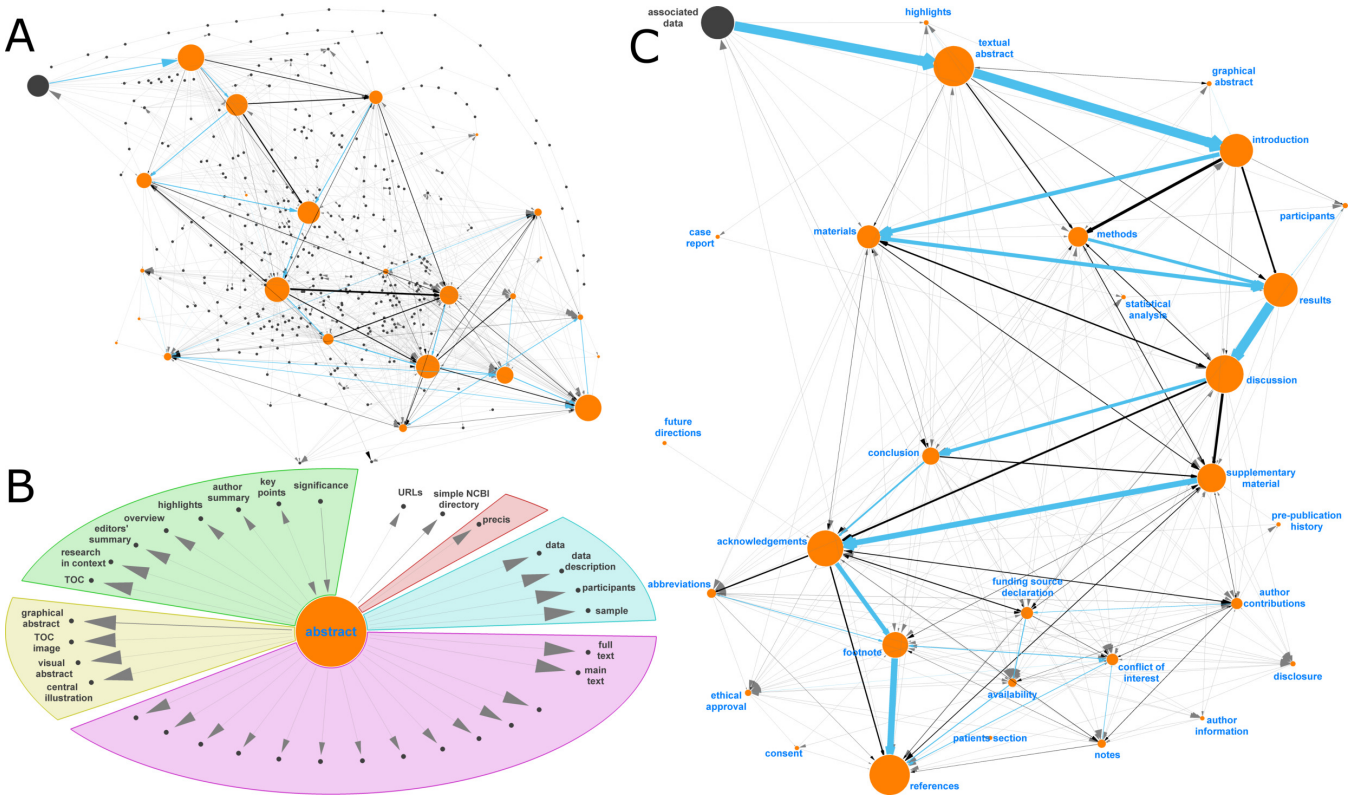

Fig. 2. Digraph generated from analyzing section headers from 2,441 Open Access publications from PubMed Central. **(A)** digraph of the v2020-06-10 IAO model consists of 372 unique nodes, of which 24 could be directly mapped to section terms (in orange) and the remainder are unmapped headers (in grey), and 806 directed edges. Relative node sizes and edge widths are directly proportional to the number of publications with these (subsequent) headers. Blue edges indicate the edge with the highest weight from the source node, edges that exist in fewer than 1% of publications are shown in light grey and the remainder in black. **(B)** Unmapped nodes connected to 'abstract' as ego node, excluding corpus specific nodes, grouped into different categories. Unlabeled nodes are titles of paragraphs in the main text. **(C)** Final digraph model used in Auto-CORPus to classify paragraphs after fuzzy matching. This model includes new (proposed) section terms and each section contains new synonyms identified in this analysis. 'Associated Data' is included as this is a PMC-specific header found before abstracts and can be used to indicate the start of most articles.

Rule-based extraction of GWAS summary-level data. GWAS Central relies on curated data extracted manually from publications or other databases. We investigated whether a rule-based approach to recognize phenotypes, SNPs and *P*-values can correctly identify data from publications contained within the database. A rule-based approach by applying the HPO on the 500 GWAS publications from the test set, identified a total of 9,599 unique disease traits (major and minor) in these publications. 949 traits are recorded for these publications in GWAS Central and the rule-based approach found 449 with a perfect match. For 65% of the publications all traits were correctly identified. SNPs have standardized formats, hence rule-based approaches are well suited for their identification. Likewise, *P*-values in GWAS publications are typically represented using scientific notation and can also be identified using rule-based methods. A total of 26,031 SNP/*P*-value pairs were found across the main text and tables of the 500 publications. For 62.4% of publications all associations recorded in the GWAS Central database are also found using this approach. While 57.6% of these publications present results (SNP/P-value pairs) only in tables, and 94.3% of pairs are found in tables, 276 associations were identified from the main text that are not represented in tables. 2,673 pairs match those recorded in the database (total of 6,969 pairs for these publications), however many associations in the database are not represented in main text/tables but in supplementary materials. Auto-CORPus includes a separate function to convert csv/tsv data to table JSON format (Figure 3), as summary-level results are often saved in these file formats as part of the supplementary information.

Named-entity recognition. Three different deep learning models were used for NER on specific paragraphs of publications. A pre-trained biomedical entity recognition algorithm (14) was fine-tuned using the results from the rule-based approach applied on GWAS data. Example sentences that contain HPO terms were used to fine-tune the transformer model and then applied on 928 MWAS publications from four broad and distinct phenotypes (cancer, gastrointestinal diseases, metabolic syndrome, and neurodegenerative, psychiatric and brain illnesses). The fine-tuned deep learning algorithm obtained accuracies between 0.76 and 0.97, averaging around 82.3% (Table 3).

We then fine-tuned the same base model for recognizing assays in text by training on sentences identified from the text that contain assays routinely used in MWAS. The first pass consisted of a rule-based approach, with fuzzy matching, to find sentences with terms and these were then used to fine-tune the deep learning model. Figure 5 shows the resulting output in JSON format for one MWAS publication (16).

Category (IAO identifier)	**Existing synonyms** (IAO v2020-06-10)	**New synonyms identified** [a]
abstract (IAO:0000315)	abstract	*precis*
acknowledgements (IAO:0000324)	acknowledgements, acknowledgments	*acknowledgement*, *acknowledgment*, acknowledgments and disclaimer
author contributions (IAO:0000323)	author contributions, contributions by the authors	*authors' contribution*, *authors' contributions*, *authors' roles*, *contributorship*, main authors by consortium and author contributions
discussion (IAO:0000319)	discussion, discussion section	*discussions*
footnote (IAO:0000325)	endnote, footnote	*footnotes*
introduction (IAO:0000316)	background, introduction	introductory paragraph
methods (IAO:0000317)	experimental, experimental procedures, experimental section, materials and methods, methods	analytical methods, concise methods, *experimental methods*, *method*, method validation, *methodology*, methods and design, methods and procedures, methods and tools, methods/design, online methods, star methods, study design, study design and methods
references (IAO:0000320)	bibliography, literature cited, references	*literature cited*, *reference*, *references*, *reference list*, selected references, web site references
supplementary material (IAO:0000326)	additional information, appendix, supplemental information, supplementary material, supporting information	*additional file*, *additional files*, additional information and declarations, additional points, *electronic supplementary material*, *electronic supplementary materials*, online content, *supplemental data*, *supplemental material*, *supplementary data*, supplementary figures and tables, *supplementary files*, *supplementary information*, *supplementary materials*, supplementary materials figures, supplementary materials figures and tables, supplementary materials table, supplementary materials tables

Table 1a. Newly identified synonyms for existing IAO terms (00003xx) from the digraph mapping of 2,441 publications. Elements in *italics* have previously been submitted by us for inclusion into IAO and added in the latest release (v2020-12-09).

Lastly, we applied a domain specific algorithm for recognizing chemical entities in the text and tables (15) to identify metabolites in the same publication (Figure 5).

Discussion

The analysis of our corpus of 2,441 Open Access publications has resulted in identifying well over 100 new synonyms for existing terms used in biomedical literature to indicate what a paragraph is about. In addition, we identified four new potential categories not previously included in the IAO. We previously submitted a subset of synonyms reported here and one of the new categories for inclusion in the IAO. These have been accepted by the IAO and are included in the latest release (v2020-12-09), hence we presented our analyses using the previous version of IAO that does not include part of our work. In the latest release, the 'graphical abstract' section has been added (IAO:0000707) based on our contribution. Also, a new 'research participants' (IAO:0000703) section has been added as contribution by others in the same release; therefore synonyms found here for the new category 'participants' section will be proposed in future as synonyms for the 'research participants' section. While the disclosure section appears to be distinct from the conflict of interest section due to a directed edge in the digraph, its synonyms could also be proposed to be part of the existing conflict of interest section in IAO.

Standardization of text for NLP is an important step in preparing a corpus. Auto-CORPus outputs a JSON file of cleaned text, with standardized headers as well as all data presented in tables in JSON format. Standardizing headers is important because some sections are more important than

Category (IAO identifier)	**Existing synonyms** (IAO v2020-06-10)	**New synonyms identified** [a]
abbreviations (IAO:0000606)	abbreviations, abbreviations list, abbreviations used, list of abbreviations, list of abbreviations used	*abbreviation and acronyms*, *abbreviation list*, *abbreviations and acronyms*, abbreviations used in this paper, *definitions for abbreviations*, glossary, key abbreviations, *non-standard abbreviations*, *nonstandard abbreviations*, *nonstandard abbreviations and acronyms*
author information (IAO:0000607)	author information, authors' information	*biographies*, *contributor information*
availability (IAO:0000611)	availability, availability and requirements	*availability of data*, availability of data and materials, *data archiving*, *data availability*, *data availability statement*, *data sharing statement*
conclusion (IAO:0000615)	concluding remarks, conclusion, conclusions, findings, summary	conclusion and perspectives, summary and conclusion
conflict of interest (IAO:0000616)	competing interests, conflict of interest, conflict of interest statement, declaration of competing interests, disclosure of potential conflicts of interest	*authors' disclosures of potential conflicts of interest*, *competing financial interests*, *conflict of interests*, *conflicts of interest*, *declaration of competing interest*, *declaration of interest*, *declaration of interests*, *disclosure of conflict of interest*, *duality of interest*, *statement of interest*
consent (IAO:0000618)	consent	informed consent
ethical approval (IAO:0000620)	ethical approval	ethics approval and consent to participate, *ethical requirements*, *ethics*, *ethics statement*
funding source declaration (IAO:0000623)	funding, funding information, funding sources, funding statement, funding/support, source of funding, sources of funding	*financial support*, *grants*, *role of the funding source*, *study funding*
future directions (IAO:0000625)	future challenges, future considerations, future developments, future directions, future outlook, future perspectives, future plans, future prospects, future research, future research directions, future studies, future work	*outlook*
materials (IAO:0000633)	materials	data, data description
statistical analysis (IAO:0000644)	statistical analysis	statistical methods, statistical methods and analysis, statistics
study limitations (IAO:0000631)	limitations, study limitations	strengths and limitations, study strengths and limitations

Table 1b. Newly identified synonyms for existing IAO terms (00006xx) from the digraph mapping of 2,441 publications. Elements in *italics* have previously been submitted by us for inclusion into IAO and added in the latest release (v2020-12-09).

Proposed category	Proposed definition	Proposed synonyms
disclosure	"A part of a document used to disclose any associations by authors that might be perceived as to potentially interfere with or prevent them from reporting research with complete objectivity."	author disclosure statement, declarations, disclosure, disclosure statement, disclosures
graphical abstract	"*An abstract that is a pictorial summary of the main findings described in a document.*"	central illustration, *graphical abstract*, TOC image, *visual abstract*
highlights	"A short collection of key messages that describe the core findings and essence of the article in concise form. It is distinct and separate from the abstract and only conveys the results and concept of a study. It is devoid of jargon, acronyms and abbreviations and targeted at a broader, non-technical audience."	author summary, editors' summary, highlights, key points, overview, research in context, significance, TOC
participants	"A section describing the recruitment of subjects into a research study. This section is distinct from the 'patients' section and mostly focusses on healthy volunteers."	participants, sample

Table 2. Newly proposed categories of entities found in 2,441 publications in the biomedical literature that could not be mapped to existing terms in IAO. Elements in *italics* have previously been submitted by us for inclusion into IAO and added in the latest release (v2020-12-09).

Known phenotype	Papers	Accuracy
cancer	492	0.84
gastrointestinal diseases	37	0.97
metabolic syndrome	286	0.80
neurodegenerative, psychiatric, brain illnesses	113	0.76

Table 3. Summary of results for named-entity recognition (NER) of phenotypes in MWAS papers.

others for specific tasks. For example, no new findings can be found in an introduction however it is well suited to discover the main phenotypes under study, only in materials/methods can details be found on how these phenotypes are studied and using what technologies, and findings can only be found in results (and discussion) sections. Hence it is important to classify these paragraphs and Auto-CORPus does this by using the structure of the publication and the digraph. We showed that we can further improve the assignment by training machine learning models with good accuracy to distinguish between different types of texts in cases where there may be ambiguity - this can be further improved by using a multi-class classifier and using all paragraphs. These data are then available for use in downstream analyses using dedicated algorithms for entity recognition or other methods. Auto-CORPus is able to process all HTML formatted tables from both GWAS and MWAS corpora, as opposed to previous methods which could only operate on 86% of 3,573 tables (17). It takes Auto-CORPus on average 0.77 seconds to process all tables within a publication compared to several minutes if this is done manually. Moreover, Auto-CORPus also supports parallel computing, thereby further reducing the time needed to process publications as these can be run in batch.

The structured JSON output is machine readable and can be used to support data import into database. Here we used the JSON output of Auto-CORPus in several examples to demonstrate some potential use cases. We demonstrated that existing algorithms trained on biomedical data can be fine-tuned to recognize new entities such as assays and phenotypes, which also opens up the possibility of using these data to train new deep learning algorithms for recognizing new entities such as metabolites (opposed to chemical entities), SNPs and *P*-values, as well as identifying the relationships between them from text. NER algorithms have difficulty with recognizing terms that are abbreviated, therefore the list of abbreviations found by Auto-CORPus can be used to replace all abbreviations in the text to their definitions.

Conclusion

The Auto-CORPus package is freely available and can be deployed on local machines as well as using high-performance computing to process publications in batch. A step-by-step guide to detail how to use Auto-CORPus is supplied with the package. The key features of Auto-CORPus are that it:

1. outputs all text and table data in a standardized JSON format,

2. classifies each paragraph into separate categories of text, and

```
{
  "status": 0,
  "error_message": "",
  "tables": [
    {
      "identifier": "1",
      ...
    },
    {
      "identifier": "2",
      ...
    },
    {
      "identifier": "3",
      "title": "Summary of results for
      named-entity recognition (NER) of
      phenotypes in MWAS papers.",
      "columns": ["Known phenotype","Papers",
      "Accuracy"],
      "section": [ {
        "section_name": "",
        "results": [
          ["cancer", 492, 0.84],
          ["gastrointestinal diseases", 37, 0.97
          ],
          ["metabolic syndrome", 286, 0.80],
          ["neurodegenerative, psychiatric,
          brain illnesses", 113, 0.76]
        ]
      } ],
      "footer": []
    }
  ]
}
```


Fig. 3. Example of JSON format for table data from this work (shown for Table 3). The Auto-CORPus output for tables consists of 'status', 'error message' and 'tables' as top level fields, 'tables' has fields 'identifier', 'title', 'columns', 'section' and 'footer', and 'section' contains 'section name' and 'results'.


```
▼ root: {} 16 keys
  ▼ ARDS: [] 1 item
      0: "acute respiratory distress syndrome"
  ▼ BALF: [] 1 item
      0: "bronchoalveolar lavage fluid"
  ▼ LC: [] 1 item
      0: "liquid chromatography"
  ▶ MS: [] 1 item
  ▶ HPLC: [] 1 item
  ▼ RP: [] 2 items
      0: "reversed phase"
      1: "reverse phase"
  ▶ HILIC: [] 1 item
  ▶ MV: [] 1 item
  ▶ NMR: [] 1 item
```


Fig. 4. Example of JSON output of abbreviation detection using a rule-based approach on an MWAS publication (16).


```
▼ root: {} 1 key
  ▼ acute respiratory distress syndrome: {} 2 keys
    ▼ assay: [] 9 items
        0: "quadrupole time of flight mass spectroscopy"
        1: "hydrophilic interaction chromatography"
        2: "mass spectroscopy"
        3: "electrospray ionization"
        4: "liquid chromatography"
        5: "reversed phase liquid chromatography"
        6: "reversed phase liquid chromatography - mass spectroscopy"
        7: "spectrometry"
        8: "liquid chromatography - mass spectroscopy"
    ▼ chemical entity: [] 39 items
        0: "citrate"
        1: "Guanosine"
        2: "lidocaine"
        3: "silica"
        4: "BALF lactate"
```


Fig. 5. Example of JSON output of named-entity recognition (NER) on an MWAS publication (16) using a fine-tuned transformer-based deep learning model for assays and bidirectional long-short term memory network for chemical entity recognition.

3. is implemented in pure Python code and does not have non-Python dependencies.

ACKNOWLEDGEMENTS

We thank Mohamed Ibrahim (University of Leicester) for identifying different configurations of tables for different HTML formats, and Joy Li and Filip Makraduli (Imperial College London) for testing the package and providing feedback.

AUTHOR CONTRIBUTIONS

TB and JMP designed and supervised the research. SS and YH developed the pipeline and analyzed data. SS developed the initial table extraction algorithm and implemented the phenotype recognition algorithm. YH developed the section header standardization algorithm and implemented the abbreviation recognition algorithm. SS fine-tuned the table extraction algorithm for use on non-PMC texts. TR refined standardization of full texts and contributed algorithms for UTF-8 and UTF-16 conversions of non-ASCII characters to Unicode. SS, YH, TB and JMP wrote the manuscript.

FUNDING

This work has been supported by Health Data Research (HDR) UK and the Medical Research Council via an UKRI Innovation Fellowship to TB (MR/S003703/1) and a Rutherford Fund Fellowship to JMP (MR/S004033/1).

FOOTNOTE

ORCID: 0000-0002-4971-9003 (JMP).

Bibliography

1. Seyedmostafa Sheikhalishahi, Riccardo Miotto, Joel T Dudley, Alberto Lavelli, Fabio Rinaldi, and Venet Osmani. Natural language processing of clinical notes on chronic diseases: Systematic review. *JMIR Med Inform*, 7(2):e12239, 4 2019. ISSN 2291-9694. doi: 10.2196/12239.
2. Ramón A-A. Erhardt, Reinhard Schneider, and Christian Blaschke. Status of text-mining techniques applied to biomedical text. *Drug Discovery Today*, 11(7):315–325, 2006. ISSN 1359-6446. doi: https://doi.org/10.1016/j.drudis.2006.02.011.
3. Nikola Milosevic, Cassie Gregson, Robert Hernandez, and Goran Nenadic. A framework for information extraction from tables in biomedical literature. *International Journal on Document Analysis and Recognition (IJDAR)*, 22(1):55–78, 2 2019. doi: 10.1007/s10032-019-00317-0.
4. Peter M. Visscher, Naomi R. Wray, Qian Zhang, Pamela Sklar, Mark I. McCarthy, Matthew A. Brown, and Jian Yang. 10 years of gwas discovery: Biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5 – 22, 2017. ISSN 0002-9297. doi: https://doi.org/10.1016/j.ajhg.2017.06.005.
5. Tim Beck, Tom Shorter, and Anthony J Brookes. Gwas central: a comprehensive resource for the discovery and comparison of genotype and phenotype data from genome-wide association studies. *Nucleic Acids Research*, 48(D1):D933–D940, 10 2019. ISSN 0305-1048. doi: 10.1093/nar/gkz895.
6. Annalisa Buniello, Jacqueline A L MacArthur, Maria Cerezo, Laura W Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, Joannella Morales, Edward Mountjoy, Elliot Sollis, Daniel Suveges, Olga Vrousgou, Patricia L Whetzel, Ridwan Amode, Jose A Guillen, Harpreet S Riat, Stephen J Trevanion, Peggy Hall, Heather Junkins, Paul Flicek, Tony Burdett, Lucia A Hindorff, Fiona Cunningham, and Helen Parkinson. The NHGRI-EBI GWAS

Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, 47(D1):D1005–D1012, 11 2018. ISSN 0305-1048. doi: 10.1093/nar/gky1120.

7. Jeremy K. Nicholson, Elaine Holmes, and Paul Elliott. The metabolome-wide association study: A new look at human disease risk factors. *Journal of Proteome Research*, 7(9): 3637–3638, 2008. doi: 10.1021/pr8005099. PMID: 18707153.
8. Werner Ceusters. An information artifact ontology perspective on data collections and associated representational artifacts. *Studies in health technology and informatics*, 180:68–72, 2012. ISSN 0926-9630.
9. Alan Ruttenberg, Adam Goldstein, Albert Goldfain, Barry Smith, Bjoern Peters, Carlo Torniai, Chris Mungall, Chris Stoeckert, Christian A. Boelling, Darren Natale, David Osumi-Sutherland, Gwen Frishkoff, Holger Stenzhorn, James A. Overton, James Malone, Jennifer Fostel, Jie Zheng, Jonathan Rees, Larisa Soldatova, Lawrence Hunter, Mathias Brochhausen, Matt Brush, Melanie Courtot, Michel Dumontier, Paolo Ciccarese, Pat Hayes, Philippe Rocca-Serra, Randy Dipert, Ron Rudnicki, Satya Sahoo, Sivaram Arabandi, Werner Ceusters, William Duncan, William Hogan, and Yongqun (Oliver) He. Information artefact ontology (v2020-06-10). https://raw.githubusercontent.com/information-artifact-ontology/IAO/v2020-06-10/iao.owl, 2020. Accessed: 2020-06-21.
10. A. Ghazvinian, N. F. Noy, and M. A. Musen. Creating mappings for ontologies in biomedicine: simple methods work. *AMIA Annu Symp Proc*, 2009:198–202, 11 2009.
11. Peter N. Robinson, Sebastian Köhler, Sebastian Bauer, Dominik Seelow, Denise Horn, and Stefan Mundlos. The human phenotype ontology: A tool for annotating and analyzing human hereditary disease. *The American Journal of Human Genetics*, 83(5):610–615, 2008. ISSN 0002-9297. doi: https://doi.org/10.1016/j.ajhg.2008.09.017.
12. Ariel Schwartz and Marti Hearst. A simple algorithm for identifying abbreviation definitions in biomedical text. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 4:451–62, 02 2003. doi: 10.1142/9789812776303_0042.
13. Katrin Fundel, Robert Küffner, and Ralf Zimmer. RelEx—Relation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371, 12 2006. ISSN 1367-4803. doi: 10.1093/bioinformatics/btl616.
14. Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 09 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz682.
15. Peter Corbett and John Boyle. Chemlistem: chemical named entity recognition using recurrent neural networks. *Journal of Cheminformatics*, 10(1), 12 2018. doi: 10.1186/s13321-018-0313-8.
16. Charles R. Evans, Alla Karnovsky, Melissa A. Kovach, Theodore J. Standiford, Charles F. Burant, and Kathleen A. Stringer. Untargeted LC–MS metabolomics of bronchoalveolar lavage fluid differentiates acute respiratory distress syndrome from health. *Journal of Proteome Research*, 13(2):640–649, 12 2013. doi: 10.1021/pr4007624.
17. Nikola Milosevic, Cassie Gregson, Robert Hernandez, and Goran Nenadic. Disentangling the structure of tables in scientific literature. In Elisabeth Métais, Farid Meziane, Mohamad Saraee, Vijayan Sugumaran, and Sunil Vadera, editors, *Natural Language Processing and Information Systems*, pages 162–174. Springer International Publishing, 2016. ISBN 978-3-319-41754-7. doi: https://doi.org/10.1007/978-3-319-41754-7_14.