BiomedCurator: Data Curation for Biomedical Literature

Mohammad Golam Sohrab[*], **Khoa N. A. Duong**[*], **Masami Ikeda, Goran Topić, Yayoi Natsume-Kitatani, Masakata Kuroda, Mari Nogami Itoh, Hiroya Takamura**
Artificial Intelligence Research Center (AIRC)
National Institute of Advanced Industrial Science and Technology (AIST), Japan
National Institutes of Biomedical Innovation, Health and Nutrition (NIBIOHN), Japan
{sohrab.mohammad, goran.topic}@aist.go.jp
{ikeda-masami, takamura.hiroya}@aist.go.jp
{natsume, m-kuroda, mari}@nibiohn.go.jp

Abstract

We present BiomedCurator[1], a web application that extracts the structured data from scientific articles in PubMed and ClinicalTrials.gov. BiomedCurator uses state-of-the-art natural language processing techniques to fill the fields pre-selected by domain experts in the relevant biomedical area. The BiomedCurator web application includes: text generation based model for relation extraction, entity detection and recognition, text classification model for extracting several fields, information retrieval from external knowledge base to retrieve IDs, and a pattern-based extraction approach that can extract several fields using regular expressions over the PubMed and ClinicalTrials.gov articles. Evaluation results show that different approaches of BiomedCurator web application system are effective for automatic data curation in the biomedical domain.

1 Introduction

Scientific article contains a lot of valuable information. For example, reports on clinical studies provide the pieces of information including the applied drug, the target disease, the dose, the dosing period, the ages of the human subjects, and the results. Such pieces of information are useful in data mining and statistical analysis for drug discovery and drug development, if they are properly structured. We call this structurization process *data curation* in this paper, aiming at two-dimensional spreadsheet style structured data as illustrated in Figure 1. Data curation is usually conducted by human experts, who are supposed to read and understand scientific papers, and fill in the spreadsheet. The purpose of this paper is to develop a web application system for automatic data curation in the biomedical domain, which we name *BiomedCurator*. Specifically, for a PubMed/ClinicalTrials.gov ID given by a user, BiomedCurator returns values for 61 information pieces (henceforth, *fields*).

The task of data curation requires a number of different NLP techniques including named entity recognition (NER), entity linking, relation extraction, and text classification. One notable characteristic of this task is that datasets curated by human experts provide spreadsheet-style supervision signal, but do not tell where in the paper each information piece is described; we cannot annotate BIO tags to the paper unlike the training data for NER.

One approach to perform automatic data curation is to use both structured data obtained from the literature and the original literature as training data. The advantage of this approach is that it can output important fields in a data format that is needed by intended users. On the other hand, disadvantages emerge, as typified by the following. (1) Since only information that is important to intended users is included in the structured data, information that is important in NLP (e.g., where each data field is described in the original literature) tends to be omitted. (2) In the process of creating such structured data, the words are often bundled into a notation different from that used in the original literature for the correction of word distortions. In this study, we have developed a web application that can easily realize automated data curation by solving these technical issues with the methods described in Section 2.2.

2 BiomedCurator: Data Curation System for Biomedical Domain

We first describe the dataset for this task, and then the natural language processing techniques used in the system, followed by the description of our system as a web application.

[*]Equal Contribution

[1]BiomedCurator is publicly available at https://biomed-text.airc.aist.go.jp/biomedcurator/ as well as its GitHub repository at https://github.com/aistairc/BiomedCurator.

Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: System Demonstrations, pages 63–71
November 20–23, 2022.

Reference Information			...	Intervention Characteristics					...	Disease Characreristics		...	Reference details		
reference_type	reference_id	associated_clinical trials	...	drug/therapy	refererence_drug therapy	dose	duration	CAS id	...	disease name	stage	...	source	...	Year
PubMed	23868010	UMIN000001779		Prednisolone	[NA]	80 mg	[NA]	50-24-8		Lung Cancer/Non-Small Cell	IIA, IIB		Full Text		2013
...															
PubMed	27924059	CEEOG0106, ML20033		Erlotinib	[NA]	150 mg	[NA]	183321-74-6		Lung Cancer	IIIB, IV		Full Text		2017
ClinicalTrial	NCT02759835	[NA]		Osimertinib	[NA]	80 mg	[NA]	1421373-65-0		Lung Cancer, Non-Small Cell	[NA]		Clinicaltrial-No result		2016
...															
ClinicalTrial	NCT02773238	[NA]		Radiotherapy	[NA]	[NA]	[NA]	[NA]		Lung Cancer, Non-Small Cell	IIB, IIIB		Clinicaltrial-No result		2016

Figure 1: A quick overview of spreadsheet style structured data. The first and second rows refer to categories and their associated fields. The first column indicates PubMed and ClinicalTrials.gov articles and the other columns are the lists of information pieces of PubMed and ClinicalTrials.gov respectively. "..." indicates more other categories and fields.

2.1 Dataset for BiomedCurator

The information required by the intended user is extracted from the articles in a comprehensive manner and structured. As information required by the intended user, 11 categories of articles in PubMed and ClinicalTrials.gov were selected, and each category was further divided into subcategories for a total of 61 fields (lists of information pieces). In the selection process, freely available PubMed articles from the last five years were screened according to whether they were about Idiopathic Pulmonary Fibrosis (IPF), Idiopathic Pulmonary (IP), or fibrosis. From these, priority was given to those with a text as well as an abstract, and the words were extracted manually to a pre-determined DESCRIPTION. A similar screening was then carried out for papers on lung cancer, with similar prioritization and extraction. To assess the quality of data curation for selecting the 11 categories and its 61 fields is based on two criterion. (1) Determination of items: Necessary information in various processes of drug discovery was extracted by dividing it into categories. This was determined by a pharmacologist with experience in drug discovery in discussion with a curator biologist. (2) For curation, a primary curator and an editor in the field of biology were provided, and further quality assurance and quality control checks were conducted.

We developed NLP models trained on a dataset from which information was manually extracted by biologists with domain knowledge as a supervisory dataset. Figure 1 shows a quick overview of spreadsheet style structured data [2]. We refer to the readers to visit our project page https://github.com/aistairc/BiomedCurator to learn more details about 11 categories and its 61 fields, as well as the models used for each field. See Appendix A for a quick overview of 11 categories and its 61 fields.

[2]Releasing of the structured data set is under consideration through the project page https://github.com/aistairc/BiomedCurator.

2.2 NLP Approaches in BiomedCurator

We address the task of data curation by five main components: (1) Generative relation extraction, (2) Named entity recognition, (3) Text classification, (4) Pattern-based extraction and (5) Information retrieval from external knowledge-base (KB).

2.2.1 Generative Relation Extraction

To extract relations in BiomedCurator, we address two main challenges: (1) the system needs to return the entities and relations where their positions are not given in the training data as mentioned in Introduction, and (2) many entities and relations in our gold data were rephrased/normalized in different ways; pure extraction might not work. These make the preparation of training data and training the discriminative relation extraction model more difficult.

In order to address these challenges, we formalize n-ary relation extraction task as a template generation problem. For a given paragraph, we expect to train a model that can generate a sequence in our predefined structure. For the sequence-to-sequence model, we utilize the BigBirdPegasus model[3] which is designed for summarization tasks to deal with long sequences. For instance, here is a simple training example for extracting relations of drug and dose entities from a given input text: *eligible patients received up to six cycles of pemetrexed, 500 mg/m(2) plus cisplatin, 75 mg/m(2) (day 1) or gemcitabine, 1000 mg/m(2) (days 1 and 8) plus cisplatin, 75 mg/m(2) (day 1). os and toxicity were assessed.*

```
TARGET OUTPUT:
[start]
  [drug] gemcitabine [/drug]
  [dose] 1000 mg/m2 [/dose]
  [and]
  [drug] cisplatin [/drug]
```

[3]https://huggingface.co/google/bigbird-pegasus-large-pubmed

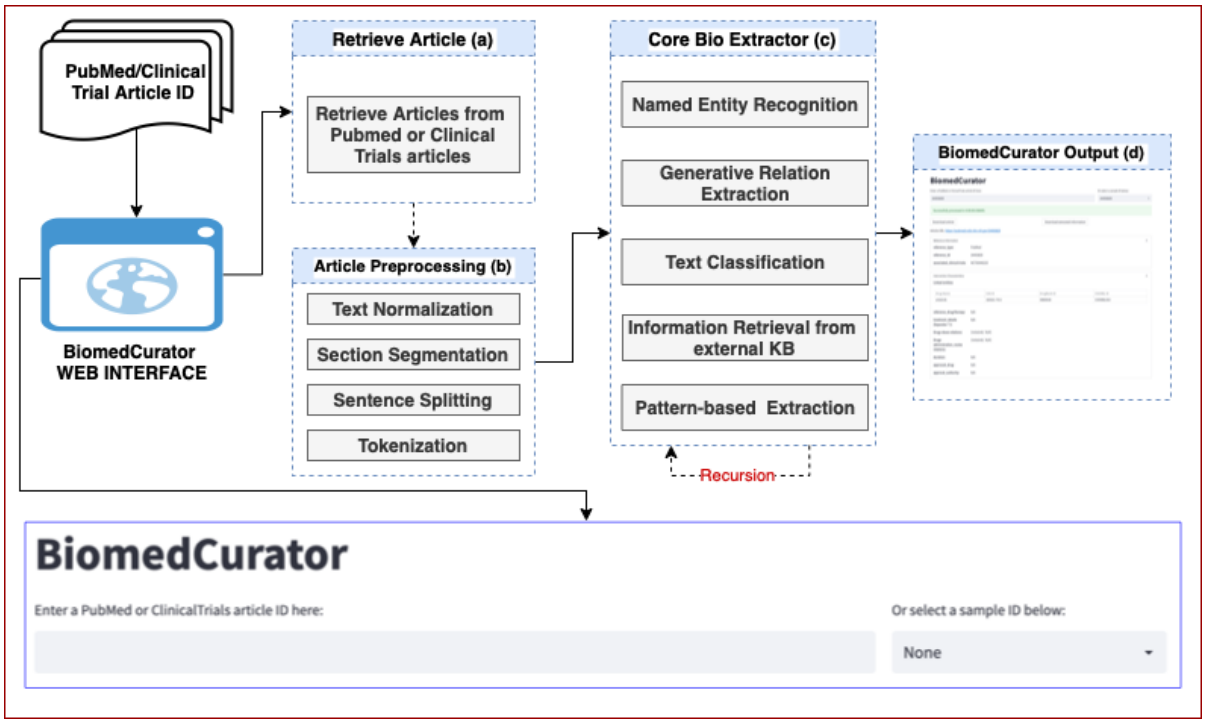

Figure 2: An overview of the workflow of BiomedCurator. The articles from PubMed and ClinicalTrials.gov are parsed through the processes (a) to (d) to extract the value of each field.

```
  [dose] 75 mg/m2 [/dose]
[or]
  [drug] pemetrexed [/drug]
  [dose] 500 mg/m2 [/dose]
  [and]
  [drug] cisplatin [/drug]
  [dose] 75 mg/m2 [/dose]
[end]
```

The target output can be expressed as

```
Combination 1:
Rel_1(gemcitabine, 1000 mg/m2)
+ Rel_2(cisplatin, 75 mg/m2)

Combination 2:
Rel_3(pemetrexed, 500 mg/m2)
+ Rel_4(cisplatin, 75 mg/m2)
```

where `[start]` and `[end]` are special tokens to indicate the beginning and the end of the template. Similarly, `[drug]`, `[/drug]`, `[dose]`, and `[/dose]` are special tokens to denote the beginning and the end of entity drug and dose. `[and]` and `[or]` are special tokens that act as operators for combining different relations together. We propose `[and]` and `[or]` to help the model be able to predict multiple relations at the same time. `[and]` is used to combine two relations together and `[or]` is used to separate two relations. When parsing a generated output to extract relations, `[and]` is greater precedence than `[or]`. `Rel` indicates relation of `[drug]` and `[dose]` entities.

BigBird Encoder-Decoder Model The BigBird architecture can process up to 8x longer sequences than BERT (Devlin et al., 2019). Therefore, for the sequence-to-sequence model, we utilize the BigBirdPegasus (Zaheer et al., 2020) model to extract the relations from a given paragraph which is an input to the BigBirdPegasus model. Unlike discriminative model, we address the relation extraction task based on generative model to fill the fields of `dose`, `drug`, and `route of administration`.

2.2.2 Named Entity Recognition

In the named entity recognition (NER) task, we employ pre-trained BERT-based NER models as they have been proven to be effective in many downstream tasks (Devlin et al., 2019). We also make use of the spaCy[4] library which is very well integrated with BERT-based models to simplify our prediction process. To extract the required information to fill the `ethnicity` field, we use BERT-

[4] https://spacy.io/

based NER model finetuned on OntoNotes 5 (Pradhan et al., 2007) dataset using SciBERT (Beltagy et al., 2019) as initial weights. In contrast to fill the `Biomarker name` field, SciBERT NER model finetuned on BioNLP13CG (Pyysalo et al., 2015) is used. For other fields, we first generate training data using distant supervision as our curated dataset do not provide position information of gold entities. Then, we finetune separate SciBERT NER models on each noisy generated data and use the trained models to extract the required information.

2.2.3 Text Classification

To extract the information of some fields, we implement two multi-class classification models: (1) SciBERT-based and (2) RandomForest-based[5] classification models. The SciBERT-based classification model is used to predict the labels of a given text input. In contrast, the RandomForest-based classification model is used to predict the labels for a combination of feature vectors as an input data. For instance, to predict the labels of the field `association` where we encode the output of three fields `marker_type`, `marker_nature`, `phenotype` as a feature vector.

2.2.4 Pattern-based Extraction

We observe that pattern-based extraction can be applied to extract the information of many fields (e.g. `reference_id`, `grade`, `stage`, `total_sample_number`, etc.). In this approach, it needs to find a substring that matches a pre-specified regular expression pattern in the text and extract the information. We refer to the readers through our project page to know more about the data fields and its corresponding approaches.

2.2.5 Information Retrieval from External Knowledge Base

Given a field information which is extracted from an article, the task is to retrieve its corresponding ID from a knowledge base (KB). This task is an entity linking problem without context. Instead of building our own model from scratch, we use existing KB API services. We look up the fields `CAS ID`, `ChEMBL ID`, `DrugBank ID`, `Entrez ID`, `Uniprot ID`, `HGVS Name`, `Rs ID`, and `KEGG Pathway Name` by using the keywords of `CAS ID`[6], `ChEMBL`[7], `DrugBank Accession`[8], `Entrez ID`[9], `Uniprot ID`[10], `HGVS`[11], `RSID`[12], `Pathway ID`[13], respectively.

2.3 Web Application of BiomedCurator

The overall workflow of BiomedCurator is illustrated in Figure 2.

Given an article ID, the system first retrieves its corresponding article from the online databases; PubMed or ClinicalTrials.gov. The article is then preprocessed before feeding into the five core components, which are designed to extract different types of information from the input article. Finally, the extracted information is returned and displayed to the user. The recursion connection below the core components in the diagram denotes that some predictions are reused and combined as input features to predict other fields. For instance, the system requires the results of `marker_type`, `marker_nature`, and `phenotype` to be able to predict the label for the field `association`.

3 Experimental Settings

In this section, we evaluate our system on our datasets.

3.1 Datasets

We conduct experiments on our curated datasets based on PubMed and ClinicalTrials.gov to address the biomedical data curation tasks. The PubMed and ClinicalTrials.gov datasets consist of 2,570 and 2,371 PubMed and ClinicalTrials.gov related scientific articles respectively. For ClinicalTrials.gov and PubMed datasets, the predefined template is labeled into 11 main categories that labeled further into several subcategories to make 61 fields. The details of 61 fields are stated on the project page[14]. Statistics of both datasets is shown in Table 1.

[5] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[6] https://commonchemistry.cas.org, https://go.drugbank.com

[7] https://go.drugbank.com

[8] https://go.drugbank.com

[9] https://www.ncbi.nlm.nih.gov/gene/, https://www.genecards.org

[10] https://www.uniprot.org/uniprot/, https://www.genecards.org

[11] https://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/LitVar/api.html

[12] https://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/LitVar/api.html

[13] https://www.genome.jp/kegg/pathway.html

[14] https://github.com/aistairc/BiomedCurator

3.2 Data Preprocessing

The data preprocessing component includes 4 main steps: (1) Text normalization, (2) Sentence splitting, (3) Section or paragraph segmentation, and (4) Tokenization.

Text normalization This step is to eliminate XML tags and multiple white spaces in the article, and special characters are converted to spaces. We then apply NFC normalization using ftfy (Speer, 2019) to convert letters followed by combining characters into single combined characters.

Sentence Splitting After the normalization step, we apply the GENIA sentence splitter model[15] to the articles to split into sentences.

Section or Paragraph Segmentation In our curated data, there are several fields that require the system to work on paragraph level instead of sentence level. For example, the relation of fields `drug` and `dose` could span across multiple sentences in the article. Therefore, we propose a simple two-step method to split the entire article into smaller chunks, namely sections and paragraphs. The first step is to leverage the article's metadata provided by PubMed and ClinicalTrials.gov in XML format. Unfortunately, there are cases where we do not have the needed metadata to be able to perform the segmentation, such as not all PubMed articles exist in the PubMed Central[16] database to be downloadable in XML format, or there are sections in ClinicalTrials.gov articles provided in a plain text format. For instance, the section `criteria` in ClinicalTrials.gov articles usually contains sub-sections `Inclusion Criteria` and `Exclusion Criteria` in a plain text format. For that reason, our second step is to utilize a rule-based classifier to predict whether a sentence is a heading/sub-heading or not. Then, we use those headings as splitting points to separate the article into different sections. Our rule-based approach is based on an observation that headings often contain some phrases like Abstract, Introduction, Method, Approach, Results etc. at the beginning of a sentence.

Tokenization Finally, we employ PegasusTokenizer of the BigBirdPegasus model[17] and BertTokenizer of the SciBERT model[18] to tokenize sentences into words.

3.3 NER Model Training

One of the challenges of our curated dataset is that it does not include the position information of the curated entities, which makes the task of training a NER model more difficult. To train the NER model of BiomedCurator, a distantly supervised approach is taken into account to generate the training data. Given a set of entities, we retrieve all the sentences that are associated with the entities (with case-insensitive and a string matching threshold of 90%) and only use those as input data for training.

3.4 Implementation

We optimize all of our models using AdamW (Loshchilov and Hutter, 2019) with a learning rate of 3e-5. For curriculum learning, we trained our generative relation extraction models with 50 epochs and a total batch size of 32 on 8 GPUs (4 examples per GPU). We trained our NER models with 5 epochs and a batch size of 32 on a single GPU with half precision enabled. We conducted each experiment on a server with 8x NVIDIA A100 for NVLink 40GiB. For NER models, we set the max input length up to 512 tokens. For relation extraction models, we use the max length of 768 tokens for encoder input and 512 tokens for decoder output.

4 Results and Discussion

Table 2 shows the performance of 17 fields in terms of precision (P), recall (R), and F-score (F) over the PubMed dataset. In this table, most of the fields performance based on F-score performing well where some fields including `duration`, `grade`, `disease_name`, and `phenotype` are performing comparatively lower than other fields. For `disease_name`, the model is trained on a distantly supervised dataset, which is filtered on gold entity mentions. Since many disease names have multiple variant forms, many were left out by the strict match filtering of the noisy dataset, which led to a poor recall score.

In contrast, Table 3 shows the accuracy performance on six other fields on PubMed dataset. We compute the accuracy for evaluating the fields that have only one answer in an article. For example,

[15] http://www.nactem.ac.uk/y-matsu/geniass/

[16] https://www.ncbi.nlm.nih.gov/pmc/

[17] https://huggingface.co/google/bigbird-pegasus-large-pubmed

[18] https://huggingface.co/allenai/scibert_scivocab_cased

Dataset	Statistics				
	Split	#Docs	Avg. Tokens/Doc	Avg. Sec./Doc	Avg. Para./Doc
PubMed	Train	1542	3296.52	10.90	37.89
	Dev	514	3037.13	10.38	35.32
	Test	514	3277.14	10.87	36.96
ClinicalTrials.gov	Train	1421	1395.08	8.41	15.34
	Dev	475	1296.97	8.39	15.00
	Test	475	1343.81	8.43	15.07

Table 1: Statistics of curated dataset based on PubMed and ClinicalTrials.gov

Field Name	P	R	F (%)
associated_clinical trials	54.24	60.38	57.10
Relation of (drug/therapy-dose)	53.77	50.59	52.13
duration	4.91	38.57	8.71
Cell line/Model Name	44.07	41.67	42.83
study_type	85.80	82.43	84.08
ethnicity	27.72	73.29	40.23
grade	10.53	21.43	14.12
phase	18.07	82.86	29.67
disease_name	91.67	5.66	10.66
stage	44.34	70.19	54.35
association	97.00	97.00	97.00
phenotype	9.09	45.19	15.14
p_value	33.37	32.68	33.02
application	94.00	95.00	94.50
allocation	39.02	72.73	50.79
masking	37.50	46.15	41.38
authors	86.35	87.35	86.85

Table 2: Performance of extraction on PubMed dataset. The performances are based on F-score for evaluating fields that have multiple answers.

Field Name	Accuracy (%)
type of alteration	87.44
phenotype_alteration	93.00
significance	99.95
author_conclusion	100.00
title	95.33
year	99.61

Table 3: Performance of extraction on PubMed Dataset. The performances are based on accuracy for evaluating fields that have single answer "correct" or "incorrect".

an article has only one published year information, and our system just needs to predict only one answer. So there are 2 possibilities: correct and incorrect. We compute the F-score for evaluating the fields that have multiple answers. For example, a certain document contains two gold answers and our system predicts one or more predictions.

In the ClinicalTrials.gov dataset, Tables 4 and 5 show the performance of different fields in terms of F-score and accuracy. In these tables, the results show the extraction performances over most of the fields are good except `duration`, `disease_sub_category`, and `BNAMIR` are relatively very poor. The low performance of the field `duration` can be explained by the fact that gold entities of the field `duration` are usually made up of 1-3 digits followed by a single word representing the unit of time (e.g. 24 hours, 120 days, 2 weeks, 3 months, etc.). As we use distant supervision to generate training examples, this results in the generation of many noisy sentences from which the entities were not actually curated. Training on this noisy data causes our model to ignore the context around entities. This explains why the model has low precision and high recall because the model tends to predict entities whenever it sees number-like tokens. Another major challenge that leads to poor scores in some fields: during manual data curation by domain experts, some information normalized or rephrased in different ways or collected in different ways which is hard to find in the article that leads to difficulty to evaluate.

5 Related Work

Several web-based tools exist that support the retrieval of biomedical information using text mining. Huang et al. (2021) addresses document-level entity-based extraction (EE), aiming at extracting entity-centric information such as entity types and entity relations, which is a key to automatic knowledge acquisition from text corpora for various domains. The authors propose a generative frame-

Field Name	P	R	F (%)
Relation of (drug/therapy-dose)	38.36	33.15	35.5
duration	2.53	81.82	4.90
disease_name	61.66	73.44	67.04
disease_sub_category	11.75	51.43	19.13
stage	23.44	85.80	36.82
BNAMIR	1.29	7.38	2.20
phenotype	22.43	51.03	31.16
total_sample_number	74.95	75.11	75.03
patient_number (case)	74.95	75.11	75.03
age(case)	87.77	87.96	87.86
gender(case)	99.37	100.00	99.68
ethnicity (case)	55.56	71.43	62.50
sponsor & collaborator	66.67	88.93	76.20
phase	97.31	97.97	97.64
inclusion_criteria	73.05	83.41	77.89
authors	95.61	94.37	94.99
intervention_model	96.25	96.48	96.37
masking	97.92	98.38	98.15
primary_purpose	98.84	99.07	98.96
association	98.00	98.00	98.00
application	99.00	99.00	99.00

Table 4: Performances on ClinicalTrials.gov Dataset over the 21 fields. BNAMIR indicates biomarker_name_as_mentioned_in_reference.

Field Name	Accuracy (%)
trial_status	56.21
title	93.89
year	86.11

Table 5: Performance of data extraction on ClinicalTrials.gov Dataset based on accuracy.

work for two document-level EE tasks: role-filler entity extraction (REE) and relation extraction (RE) to address the issue of long-term dependencies among entities at the document-level. In this work, the authors first formulate the task as a template generation problem, allowing models to efficiently capture cross-entity dependencies, exploit label semantics, and avoid the exponential computation complexity of identifying n-ary relations. Other works such as Christopoulou et al. (2019) and Jia et al. (2019) addressed the document-level relation extraction. Christopoulou et al. (2019) introduced constructing a document-level graph from sentence encoding, then extracting entity relations from edge representations in the graph. Where, Jia et al. (2019) proposed a layer classifiers-based pipeline architecture to obtain hierarchical representation of n-ary relations.

Li et al. (2022) proposed pubmedKB, a web server designed to extract and visualize semantic relationships between four biomedical entity types: variants, genes, diseases, and chemicals. pubmedKB uses state-of-the-art natural language processing techniques to extract semantic relations from the large number of PubMed abstracts. Wang et al. (2018) proposed a novel framework CPIE (Clause+Pattern-guided Information Extraction) that incorporates clause extraction and meta-pattern discovery to extract structured relation tuples. Deng et al. (2021) addressed an extraction of gene-disease association using a BERT-based language model. Xing et al. (2018) proposed a pipeline based approach to extract the relation between gene-phenotype from biomedical literature.

In contrast, our work is broader in the sense that it addresses entity and relation-based extraction along with entity linking based on external KB from PubMed and ClinicalTrials.gov datasets. We also introduce multi-class classification and pattern-based approaches for data curation.

6 Conclusion

We propose BiomedCurator based on several data curation approaches from biomedical literature. Our approach is distantly supervised based approach to create training data. Besides, it follows the state-of-the-art NLP techniques that extracts the information from PubMed and ClinicalTrials.gov articles and fill the 61 data fields. We also present an interactive web application of BiomedCurator to facilitate the biomedical research. Experimental results on two datasets show that BiomedCurator performs very well to extract the template fields information in terms of both F-score and accuracy. The BiomedCurator system is continually evolving; we will continue to improve the system as well as to implement new functions such as n-ary relation extraction to further facilitate BiomedCurator research.

Acknowledgements

This work is based on results obtained from a project commissioned by the Public/Private R&D Investment Strategic Expansion PrograM (PRISM). We appreciate insightful feedback from the anonymous reviewers.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Connecting the dots: Document-level neural relation extraction with edge-oriented graphs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4925–4936, Hong Kong, China. Association for Computational Linguistics.

Chuan Deng, Jiahui Zou, Jingwen Deng, and Mingze Bai. 2021. Extraction of gene-disease association from literature using biobert. *The 2nd International Conference on Computing and Data Science*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kung-Hsiang Huang, Sam Tang, and Nanyun Peng. 2021. Document-level entity-based extraction as template generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5257–5269, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Robin Jia, Cliff Wong, and Hoifung Poon. 2019. Document-level n-ary relation extraction with multiscale representation learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3693–3704, Minneapolis, Minnesota. Association for Computational Linguistics.

Peng-Hsuan Li, Ting-Fu Chen, Jheng-Ying Yu, Shang-Hung Shih, Chan-Hung Su, Yin-Hung Lin, Huai-Kuang Tsai, Hsueh-Fen Juan, Chien-Yu Chen, and Jia-Hsin Huang. 2022. pubmedKB: an interactive web server for exploring biomedical entity relations in the biomedical literature. *Nucleic Acids Research*, 50(W1):W616–W622.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations, ICLR 2019, New Orleans, USA*.

Sameer S. Pradhan, Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2007. Ontonotes: A unified relational semantic representation. In *International Conference on Semantic Computing (ICSC 2007)*, pages 517–526.

Sampo Pyysalo, Tomoko Ohta, Rafal Rak, Andrew Rowley, Hong-Woo Chun, Sung-Jae Jung, Sung-Pil Choi, Jun'ichi Tsujii, and Sophia Ananiadou. 2015. Overview of the cancer genetics and pathway curation tasks of BioNLP shared task 2013. *BMC bioinformatics*, 16(10):1–19.

Robyn Speer. 2019. ftfy. Zenodo. Version 5.5.

Xuan Wang, Yu Zhang, Qi Li, Yinyin Chen, and Jiawei Han. 2018. Open information extraction with meta-pattern discovery in biomedical literature. *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*.

Wenhui Xing, Junsheng Qi, Xiaohui Yuan, Lin Li, Xiaoyu Zhang, Yuhua Fu, Shengwu Xiong, Lun Hu, and Jing Peng. 2018. A gene–phenotype relationship extraction pipeline from the biomedical literature using a representation learning approach. *Bioinformatics*, 34(13):i386–i394.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. *CoRR*, abs/2007.14062.

A Curated Data Fields

The overview of categories, fields, related natural language processing techniques (NLPT), and descriptions are illustrated in Table 6. The first and second columns indicate category and its fields name. The third column stands for different NLP approaches applied in each field. In this column, PE, RE, EE, EL, and TC refer to pattern-based extraction, relation extraction, entity extraction, entity linking, and text classification-based approaches are applied for data curation. Besides, Fixed Value (FV) means a specific value in the curated data and NA means Not Available at the moment. Fields Descriptions are also added in column four.

Category	Field Name	NLPT	Description
Reference Information	reference_type	FV	Source of the article. Ex: PubMed or Clinical trial
	reference_id	PE	Unique Pubmed ID or Clinical trial id of the curated document
	associated_clinical trials	PE	Provides the associated clinicaltrial ids for which the results were published
	s_no	NA	Each assertion has given a unique number
Intervention Characteristics	drug/therapy	RE	Captured the list of authors focus drug/s of case group.
	reference_drug/therapy	RE	Captured the list of authors focus drug/s of reference group.
	treatment_details	NA	Detail description of the treatment, including but not limited to patient details, drug/therapy, dose/cycles, duration, route, schedule, analysis.
	dose	RE	It represents the concentration value of the drug used in the given reference
	route of administration	RE	The route through which the drug is administered.
	duration	EE+PE	Time period of the treatment.
	CAS id	EL	Chemical abstracts service registry number of the drug.
	ChEMBL	EL	Unique id as provided by ChEMBL.
	drug bank id	EL	Drug bank id for the given drug.
	approved_drug	NA	Name of the drug which is approved by any approval authority.
	approval_authority	NA	Name of the organization/institution has the authority to approve the respective drug. Ex: FDA.
Disease Characteristics	disease_name	EE+PE	Name of the focused indication for which the biomarker was studied.
	disease_sub_category	EE+PE	Represents the subtype or any state of the disease mentioned in the given reference.
	Stage	PE	Stages of the disease Eg: Stage I, II, III, IV, etc.
	Grade	PE	Grading of the disease Eg: Grade I, II, III, IV, etc.
	Histopathology	EE	Additional details of the disease mentioned in the article Ex: Stage, histopathology etc.
Biomarker Details	BNAMIR	EE	Complete name of the biomarker. Abbreviations are extended for ease of understanding.
	marker_type	EE	Represents the type of the biomarker based on the techniques used to measure the biomarker Eg: Biochemical, Genomic etc.
	marker_nature	EE	Represents the chemical nature of the biomarker based on the techniques used to measure the biomarker Eg: Protein, Gene, Lipid etc.
	Entrez id	EL	Unique ID as provided by the NCBI Entrez gene database for each gene.
	Uniprot id	EL	Protein accession number of UniprotKB database.
	type_of_variation	EL	Represents standard HGVS constructs unique for each variation.
	rs_id	EL	Represents the unique reference number for each SNP at a specific position. Taken from NCBI site – (dbSNP) Eg: rs763110.
	HGVS Name	EL	Field describes nucleotide/DNA (c.) change as per the HGVS format (the nucleotide/genomic numbering should be as in article only)
Biomarker association with outcomes	association	TC	Describes about the high level type/category of biomarker association with outcomes. Associations are of 5 types: Gene - drug relationships; Gene - gene interactions; Gene - pathway relationships; Gene - phenotype relationships; Gene - transcript information
	marker_alteration	EE	Represents the type of alteration or measurement done for biomarker Eg: Gene expression, Polymorphism, Biomarker level etc.
	type of alteration	PE	Represents the modification of the marker mentioned in the article i.e change of biomarker expression or levels. Eg: High; Low; Decreases; Association; Upregulation etc.
	phenotype	TC	Biomarker associates with any phenotype character, end point, outcome, any physiological process and other biomarkers of the study sample.
	phenotype_alteration	PE	Represents the state of change for the outcome variables which are associated with the studied biomarker.
	significance	PE	Represents the level of significance of P value between different groups Eg: Non-significant or Significant.
	p_value	PE	P value (Significance) between the different groups for comparison of biomarker result values or any other values related to biomarker. Ex: P=0.016
Utility	application	TC	Denotes the utility of the biomarker for a given condition in a specific reference (either clinical trial or pubmed article).
	author_conclusion	TC	Represents the utility of the biomarker from the author's perspective in the given reference. Yes indicates that author, in the reference, supports the application of the biomarker for the given indication. No indicates that author in the reference does not support the application of the biomarker for the given indication.
	evidence_statement	NA	Gives the structured description of the application text of the biomarker in a given condition specific to each reference and clinical status.
Study characteristics	study_type (Clinical/PreClinical)	PE	Represents the status of the clinical study Ex: Clinical, Preclinical etc.
	Cell line/ Model Name	EE	Represents the cell lines used in the preclinical model/It represents the preclinical model Eg: Mouse, rat etc.
	total_sample_number	PE	Denotes total number of participants from both study and reference sample group in a particular study.
	patient_number (case)	PE	To capture the study group sample size for the curated assertion from the article.
	patient_number (reference)	PE	To capture the reference group sample size for the curated assertion.
	age (case)	PE	Used to capture the study sample age from the article.
	gender (case)	PE	Used to capture the gender for studied samples from the article.
	ethnicity (case)	EE	This represents the nationality/ethnicity of the study group as stated in the article.
Trial level information	trial_status	PE	Current stage of a clinical study. Ex: Completed, Terminated etc.
	sponsor & collaborator	PE	Sponsors/collaborators of the clinical study.
	phase	PE	Represents the clinical phase of the trial. Ex: 0, I, II, III, IV.
	inclusion_criteria	PE	Description on the Inclusion criteria for the patients in the clinical study.
	exclusion_criteria	PE	Description on the Exclusion criteria for the patients in the clinical study.
Study design	allocation	PE	Assigning trial subjects to treatment or control groups. Ex: Non-randomized, Randomised.
	intervention_model	PE	Type of intervention model from the study. Ex: Single Group Design, Parallel Design, Crossover Design and Factorial Design.
	masking	PE	Types of Masking include None, Open Label, Single and Double Blind Masking.
	primary_purpose	PE	Represents purpose of the study primarily under taken for the research.
Additional details	pathway_name	EL+PE	Names of the pathways in which a biomarker has a role. Taken from KEGG database.
Reference details	Source	FV	whether the curated data is from full-text or abstract of the article or ClinicalTrials.
	Title	PE	Title of the article.
	Authors	PE	Authors of the article.
	Article/URL	PE	Name of the journal or specific links from which the information is captured
	Year	PE	Year in which the given article published Ex: Article published year. for Pubmed articles and First received year is considered for Clinicaltrials.

Table 6: An overview of category, field with corresponding description, and methodology.