Beyond Metadata: What Paper Authors Say About Corpora They Use

Nikolay Kolyada[1] **Martin Potthast**[2] **Benno Stein**[1]
[1] Bauhaus-Universität Weimar, Germany, <first>.<last>@uni-weimar.de
[2] Leipzig University, Germany, martin.potthast@uni-leipzig.de

Abstract

The growing ecosystem of data sharing in science has put dataset search into the focus. To make data sharing and reuse more feasible, new retrieval tools and services are being developed. Currently, dataset retrieval relies almost exclusively on metadata provided by the publishers. To extend this knowledge source our work studies the task of "dataset review mining" in scientific publications. For the field of Natural Language Processing we collect metadata about datasets from established resources such as the ELRA and LDC catalogs, and then extract review statements about the datasets from ACL Anthology Corpus publications, compiling the Webis-Dataset-Reviews-21 corpus. By analyzing the reviews we identify different categories of what paper authors write about data. To the best of our knowledge, this is the first analysis of this kind in the field of Natural Language Processing, albeit similar analyses have been carried out in the social and medical sciences. Our corpus and the underlying code are shared alongside this paper.[1,2,3]

1 Introduction

Recently, Google introduced its Dataset Search service (Brickley et al., 2019).[4] Around 30 million datasets have been indexed to date,[5] hosted at web pages and repositories all across the web. Although data sharing and dataset search has long been a best practice in natural language processing, Google's service has accelerated and standardized data sharing across scientific communities, since only those datasets are indexed, for which a certain metadata description prescribed by Google is provided.

[1] https://github.com/webis-de/ACL-21
[2] https://webis.de/data.html#Webis-Dataset-Reviews-21
[3] https://doi.org/10.5281/zenodo.4889032
[4] https://datasetsearch.research.google.com/
[5] https://ai.googleblog.com/2020/08/an-analysis-of-online-datasets-using.html

Chapman et al. (2020) and Koesten et al. (2017) survey the state of the art in dataset search and analyze how it differs from classical information retrieval. Approaches to dataset search from related fields, such as databases, semantic web, entity-based search, and tabular search are reviewed, but neither solves the task comprehensively. From a user's perspective, searching for datasets entails two steps: (1) its retrieval using standard keyword search, and (2) examining the search results to identify the datasets suitable to one's needs. The second step strongly differs from web search, since a (large) dataset cannot be easily reviewed in the browser; merely it's metadata is available—as well as the descriptions found on the download page. We argue that a crucial piece of information is missing to inform the examination of dataset search results: dataset user experience.

Unlike for commercial products in online shops, for datasets, there are hardly any platforms where users share their experiences using them. On the few that do exist, the amount of published reviews cannot be compared with that of commercial products. In any case, due to the specificity of most datasets, few people actually use them. Their feedback is typically found only as part of their reports and scientific publications. However, here, authors are at liberty to provide testimonials and criticisms that can be more thorough and extended in comparison to commercial product reviews.

To assess this potential source of information about datasets, and to eventually harness it for dataset search, this paper takes the first step in this direction by compiling and analyzing Webis-Dataset-Reviews-21, a corpus of dataset reviews from a large corpus of scientific publications, namely the ACL Anthology. Compiling a list of datasets in the field of Natural Language Processing and Computational Linguistics using authoritative sources including catalogs of the European

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 5085–5090

Language Resources Association (ELRA) and its LRE Map, the Linguistic Data Consortium (LDC), and others, we extract from the ACL Anthology's 57,608 papers 466,567 sentences mentioning a dataset. Exploring these mention statements, we organize them in a taxonomy, assessing their usefulness for dataset retrieval.

2 Related Work

Dataset review mining can be seen as part of the task of dataset retrieval, as well as that of extracting information from scientific documents. The demand for dataset search has increased with more data being published across scientific communities, by industry, and by government bodies. Multiple initiatives facilitate data sharing and encourage scientists to do so, such as Zenodo,[6] and DataCite[7] (Rueda et al., 2016), which also provide digital object identifiers (DOIs) for datasets, and a platform for publishing them. Several open software frameworks implement data collection management. Currently, most common are CKAN (Solr), Socrata, and OpenDataSoft. Chapman et al. (2020) show that most of the search features implemented rely exclusively on metadata. Brickley et al. (2019) observe that this restricts the effectiveness of dataset retrieval engines as metadata quality varies significantly. Notwithstanding this shortcoming, specialized dataset search engines still dramatically improve dataset discovery and sharing.

Several commonly used schemas and formats for dataset metadata exist, such as the JavaScript Object Notation for Linked Data (JSON-LD), and the Dataset schema[8] based on W3C DCAT (Erickson et al., 2013). The latter has become a de-facto standard, as it is used by Google Dataset Search to index data collections found in the open web. Many data publishers meanwhile adopted these standards of making metadata about their published datasets machine-readable, and data portal software platforms, such as CKAN, integrate this functionality by default. In addition, improvements on tracking citations of data are being actively developed. Projects, like Semantic Scholar, Microsoft Academic Graph,[9] and Google Dataset Search, along with Google Scholar, couple data and publications to improve data discovery.

The task of extracting structured information from scientific publications has been tackled many times. Gupta and Manning (2011) extract key aspects of scientific papers, including focus, technique, and domain from the ACL Anthology. Mesbah et al. (2018), Luan et al. (2018), and Jain et al. (2020) propose approaches to identify entities and their relations in scientific documents. Gábor et al. (2016) creates an annotated corpus for concepts and semantic relations based on the ACL Anthology. Duck et al. (2016) employed text mining to process dataset and software mentions in biological and medical publications from PubMed Central. Boland et al. (2012) identify references to datasets in social science publications. Closely related to our work, *software* mentions in scientific documents can be mined using a Grobid library module,[10] e.g., to give research software more credit.

In 2019, a shared task focusing on the tasks of dataset review mining and extraction of scientific methods and fields was organized by the Coleridge Initiative.[11] The training data provided was based on the Inter-university Consortium for Political and Social Research (ICPSR) data catalog,[12] comprising around 10,000 datasets used in the social sciences and a labeled corpus of 5,000 publications matched with mentioned datasets. The best-performing solution by The Allen Institute for Artificial Intelligence (AI2) implements a set of rule-based candidate citations by exact string matching mentions and datasets as a first step. An ad-hoc named entity recognition model has been trained based on matches identified. Mention labels were generated by string matching mentions in the provided annotations against the full text of papers. Then, candidate datasets were scored based on TF-IDF-weighted token overlap between the mention text and the dataset title, linking them using a binary classifier.

Färber et al.'s (2021) work is perhaps the most closely related one to our paper. It recognizes datasets along with scientific methods in a corpus of 510,000 publications. The TSE-NER approach by Mesbah et al. (2018) was used to identify methods and datasets in order to integrate this data into the Microsoft Academic Knowledge Graph (MAKG). Moreover, identified dataset references has been classified into used vs. not-used based on the textual context.

[6] https://zenodo.org/
[7] https://datacite.org/
[8] http://schema.org/Dataset
[9] This project has been discontinued: https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/
[10] https://github.com/ourresearch/software-mentions
[11] https://coleridgeinitiative.org/richcontextcompetition/
[12] https://www.icpsr.umich.edu/icpsrweb/

3 Collecting the Metdata of NLP Datasets

The two major steps to construct the Webis-Dataset-Reviews-21 corpus are the compilation of a collection of metadata of datasets used in the field of Natural Language Processing, and the extraction of mentions of these datasets from the ACL Anthology. To tackle the first step, we crawl authoritative NLP dataset catalogs, collect their metadata, clean and normalize it, merge duplicates, and format this compilation according to the specification of https://schema.org/Dataset.

The catalogs crawled are shown in Table 1. Although these catalogs may not list all existing datasets in Natural Language Processing, they certainly contain many of the most prominent and the most commonly used ones in the field. Nevertheless, there are bound to be many datasets in the "long tail": Their identification will require tailored named entity recognition approaches along the lines of the ones employed in aforementioned shared tasks. Our current collection of dataset metadata contains 13,372 entries in total.

The metadata enclosed in the catalogs are based on different schemas. In order to render them amenable for subsequent processing, we went about cleaning, normalizing, and unifying them across catalogs. Duplicates were identified and merged into a normalized schema: For each dataset, we collect properties relating to its source, original title, acronym, DOI, description, year, creator, and URL, as well as additional properties, such as language, format, size, and metadata about the associated paper, if available. Our final, unified metadata catalog of NLP datasets forms part of the Webis-Dataset-Reviews-21 corpus.

Catalog	Datasets
Language Resources monitoring (LRE Map)[13]	6,143
European Language Association (ELRA)[14]	5,398
Linguistic Data Consortium (LDC)[15]	950
Big Bad NLP Database[16]	791
NLP Progress[17]	90
Σ	13,372

Table 1: NLP resources crawled for dataset metadata.

[13] http://lremap.elra.info/
[14] http://www.elra.info/
[15] https://www.ldc.upenn.edu/
[16] https://datasets.quantumstat.com/
[17] http://nlpprogress.com/

4 Mining Dataset Reviews

The second step of constructing our corpus was mining for dataset reviews from the ACL Anthology in the form of sentences that mention a dataset found in our unified catalog. We extract dataset mentions from the anthology's papers, annotate them, and develop a taxonomy with regard to the information a dataset mention provides about the data. We further analyze common patterns used by authors to describe the data.

4.1 Preprocessing the ACL Anthology

The ACL Anthology[18] compiles all papers published in the fields of computer linguistics and natural language processing published between the late 1960s up to today and makes them available open access. We used a collection of papers from that includes the papers up until 2020, 57,606 PDF files in total, omitting ones not written in English.

To enable text processing the papers, we use the Grobid (GRO, 2008–2021) library for parsing scientific publications given as PDF and extracting structured XML/TEI-encoded documents.[19] In the subsequent review mining step, we decided to analze only the publications' body sections (including the introduction), as a pilot study showed datasets mentioned in the abstract or references sections hardly contain any information about them. Harnessing the annotations provided by Grobid, we resolve references and footnotes within the text of a paper and add the titles of cited items to the document full text. If a paper contains exact references in a form of DOI or URL to the corpora used, we consider this information as additional features in the mention extraction step. Information about paper sections and paragraphs within the body are preserved as well.

4.2 Dataset Mentions Extraction

For the dataset mention extraction, we exploit the fact that most datasets in natural language processing research have proper names and/or distinct acronyms. This is unlike in the social sciences, where many datasets rather have descriptions for names (e.g., "national census data on population development 2020", which then gets shortened ad-hoc by the authors while writing their paper). By contrast, for NLP dataset, their creators often conceive of catchy acronyms (e.g., "WordNet") that

[18] https://www.aclweb.org/anthology/
[19] https://tei-c.org/

Characteristic	Number of elements
Mentions	466,567
Coreference mentions	93,176
Publications with at least one mention	(92%) 53,129
Unique datasets mentioned	(22%) 2,986

Table 2: Mentions of datasets in the ACL Anthology.

Dataset	Mentions	Coreferences	Publications
WordNet	**43,746**	**8,925**	6,591
Wikipedia	36,954	8,682	6,602
Penn Treebank	20,453	5,661	**8,039**
FrameNet	9,236	2,047	1,386
Brown	8,666	2,317	2,752
PropBank	5,366	1,202	1,032
Europarl	5,075	1,477	1,624
BNC	4,607	1,403	895
PDTB	4,232	967	399
Freebase	3,799	925	759
Gigaword	3,792	1,046	1,720
UMLS	3,705	783	687
VerbNet	3,372	636	592
Wiktionary	3,156	680	553
DUC	3.022	656	474
OntoNotes	2,973	821	782
SQuAD	2,872	764	432
ATIS	2,737	750	483
DBpedia	2,625	642	576
TimeML	2,584	614	384
PubMed	2,437	585	648

Table 3: Number of mentions and papers for top 20 datasets mentioned in the ACL Anthology publications.

are then picked up by paper authors. Therefore, we get by with basic string matching to extract 466,567 mentions of 2,986 unique datasets out of the 13,371 ones in our catalog.[20] At the same time, with this basic approach, we introduce a baseline for the task of dataset mention extraction, which, in practice, has a similar recall compared to that of a full-text search performed over an indexed collection of papers.

We define a dataset review as "an assessment from the person who used the data" (be it praise or criticism). Reviews are often spread across sentences around a dataset mention, but are usually confined to within the same paragraph or section. We employ a co-reference resolution model (Clark and Manning, 2016) to extract further discussion of a dataset from adjacent sentences around a dataset mention. Tables 2 and 3 overview relevant corpus statistics.

[20] In future work, we plan on investigating if the recall of mentioned datasets can be increased with heuristic matching approaches as well as citation-based ones; cursory attempts, however, showed especially the former to introduce a lot of noisy mentions, which is why we omitted them in this analysis.

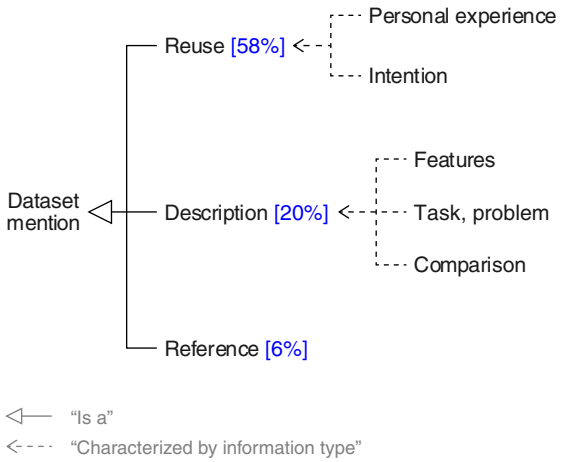

Figure 1: Taxonomy of dataset mentions. The type "personal experience" corrsponds to a dataset review, and implies reuse of the data.

4.3 A Taxonomy of Dataset Mentions

We manually reviewed a random sample of 1,000 dataset mentions and derived a taxonomy of three classes as shown in Figure 1: reuse, description, and reference. We primarily distinguish between "active" cases of reuse of a dataset from "passive" descriptions and references. Statements of dataset reuse include cases where data is used as a basis for experiments, training models, and synthesis of new or sub-datasets. For instance:

(1) *"Second, we reuse the RCV1-V2, using a version that contained a selected 5,000 term vocabulary."*

Descriptions and references rather give details about a dataset or compare it:

(2) *"For instance, two words are said to be synonyms if they belong in the same synset in the WordNet."*;

(3) *"MovieQA is a challenging dataset for movie understanding. The dataset consists of 14,944 multiple choice questions about 408 movies."*

Altogether, 58% of the sample are reuse mentions, 20% descriptions, and 6% references. The remaining mentions are ambiguous and/or incorrect.

From the reuse mentions, 63% share a "personal experience" which we operationalize as a form of dataset review. Projected to the entire set of dataset mentions, this potentially results in about 150,000 such cases. For instance:

(4) *"When only WordNet, not BabelNet, is used for identifying lexico-semantic relations, per-*

formance increases slightly, which we attribute to noise that comes with using BabelNet.";

(5) *"The TimeBank and the data used in the two TempEval challenges are important, as they have annotations describing not just dates and times, but also events and temporal relations between these entities.*";

(6) *"While AQuA may prove useful for training, it is inappropriate as an evaluation set."*

Unsurprisingly, commonly used datasets, such as Wikipedia, WordNet, the Penn Treebank, the Brown Corpus, and GigaWord receive the most mentions with over 5,000 each. The majority are mentioned only a few times; only 25% of the datasets are mentioned more than 10 times. Altogether, description mentions provide information about the data, its features, and common tasks it is used for. Reuse mentions conveying a personal experience provide information about suitability of the dataset for the paper's envisioned task, as well as details about the data.

5 Conclusion

We compiled a comprehensive list of NLP-related datasets and extracted their mentions from the ACL Anthology. Analyzing the mentions, three basic categories can be distinguished, namely reuse, descriptions, and references. The former two categories of mentions are considered to be useful as a source of information on a dataset in the task of dataset retrieval, enabling dataset search engines to display more in-depth information about a dataset on a search results page, as well as informing their retrieval models. Other future directions include improving the extraction of mentions, e.g., via "few-shot" domain-specific named entity recognition, as well as, expanding to other fields of research.

References

2008–2021. Grobid. https://github.com/kermitt2/grobid.

Katarina Boland, Dominique Ritze, Kai Eckert, and Brigitte Mathiak. 2012. Identifying references to datasets in publications. In *International Conference on Theory and Practice of Digital Libraries*, pages 150–161. Springer.

Dan Brickley, Matthew Burgess, and Natasha Noy. 2019. Google dataset search: Building a search engine for datasets in an open web ecosystem. In *The World Wide Web Conference*, pages 1365–1375.

Adriane Chapman, Elena Simperl, Laura Koesten, George Konstantinidis, Luis-Daniel Ibáñez, Emilia Kacprzak, and Paul Groth. 2020. Dataset search: a survey. *The VLDB Journal*, 29(1):251–272.

Kevin Clark and Christopher D Manning. 2016. Deep reinforcement learning for mention-ranking coreference models. *arXiv preprint arXiv:1609.08667*.

Geraint Duck, Goran Nenadic, Michele Filannino, Andy Brass, David L Robertson, and Robert Stevens. 2016. A survey of bioinformatics database and software usage through mining the literature. *PloS one*, 11(6):e0157989.

John S. Erickson, Amar Viswanathan, Joshua Shinavier, Yongmei Shi, and James A. Hendler. 2013. Open government data: A data analytics approach. *IEEE Intell. Syst.*, 28(5):19–23.

Michael Färber, Alexander Albers, and Felix Schüber. 2021. Identifying used methods and datasets in scientific publications.

Kata Gábor, Haïfa Zargayouna, Davide Buscaldi, Isabelle Tellier, and Thierry Charnois. 2016. Semantic annotation of the ACL anthology corpus for the automatic analysis of scientific literature. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).

Sonal Gupta and Christopher D. Manning. 2011. Analyzing the dynamics of research by extracting key aspects of scientific papers. In *Fifth International Joint Conference on Natural Language Processing, IJCNLP 2011, Chiang Mai, Thailand, November 8-13, 2011*, pages 1–9. The Association for Computer Linguistics.

Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. 2020. Scirex: A challenge dataset for document-level information extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7506–7516. Association for Computational Linguistics.

Laura M Koesten, Emilia Kacprzak, Jenifer FA Tennison, and Elena Simperl. 2017. The trials and tribulations of working with structured data: -a study on information seeking behaviour. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 1277–1289.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018*

Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, pages 3219–3232. Association for Computational Linguistics.

Sepideh Mesbah, Christoph Lofi, Manuel Valle Torre, Alessandro Bozzon, and Geert-Jan Houben. 2018. TSE-NER: an iterative approach for long-tail entity extraction in scientific publications. In *The Semantic Web - ISWC 2018 - 17th International Semantic Web Conference, Monterey, CA, USA, October 8-12, 2018, Proceedings, Part I*, volume 11136 of *Lecture Notes in Computer Science*, pages 127–143. Springer.

Laura Rueda, Martin Fenner, and Patricia Cruse. 2016. Datacite: Lessons learned on persistent identifiers for research data. *Int. J. Digit. Curation*, 11(2):39–47.