# DataQuest: An Approach to Automatically Extract Dataset Mentions from Scientific Papers

Sandeep Kumar[1(✉)], Tirthankar Ghosal[2], and Asif Ekbal[1]

[1] Department of Computer Science and Engineering, Indian Institute of Technology Patna, Bihta, India
{19_11mc12,asif}@iitp.ac.in

[2] Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic
ghosal@ufal.mff.cuni.cz

**Abstract.** The rapid growth of scientific literature is presenting several challenges for the search and discovery of research artifacts. Datasets are the backbone of scientific experiments. It is crucial to locate the datasets used or generated by previous research as building suitable datasets is costly in terms of time, money, and human labor. Hence automated mechanisms to aid the search and discovery of datasets from scientific publications can aid reproducibility and reusability of these valuable scientific artifacts. Here in this work, utilizing the *next sentence prediction* capability of language models, we show that a BERT-based entity recognition model with POS aware embedding can be effectively used to address this problem. Our investigation shows that identifying sentences containing dataset mentions in the first place proves critical to the task. Our method outperforms earlier ones and achieves an F1 score of 56.2 in extracting dataset mentions from research papers on a popular corpus of social science publications. We make our codes available at https://github.com/sandeep82945/data_discovery.

**Keywords:** Dataset discovery · Dataset mention extraction · Publication mining · Deep learning

## 1 Introduction

*Data is the new oil* for as they say, and datasets are crucial for scientific research. There has been an enormous growth of data and rapid advancement in data science technologies a generation or two ago, which has opened considerable opportunities to conduct empirical research. Now the researchers can rapidly acquire and develop massive, rich datasets, routinely fit complex statistical models, and conduct their science in increasingly fine-grained ways. Finding a good dataset to support/carry out the investigation or creating a new one is crucial to research.

S. Kumar and T. Ghosal—Equal contribution.


H.-R. Ke et al. (Eds.): ICADL 2021, LNCS 13133, pp. 43–53, 2021.
https://doi.org/10.1007/978-3-030-91669-5_4

Faced with a never-ending stream of new findings and datasets generated using different code and analytical techniques, researchers cannot readily determine who has worked in an area before, what methods were used, what was produced, and where those products can be found. However, many datasets go unnoticed due to lack of proper dataset discovery tools, and hence many efforts are duplicated. A survey [16] even suggests that data users' and analysts' productivity grow less because more than a third of their time is spent finding out about data rather than in model development and production. The links from scientific publications to the underlying datasets and vice versa are helpful in many scenarios, including building a dataset recommendation system, determining the impact of a given dataset, or identifying the most used datasets in a given community, sharing available datasets through the research community.

*Empirical researchers and analysts who want to use data for evidence and policy mostly face challenges in finding out who else worked with the data. Hence, good research is underused, great data go undiscovered and are undervalued, and time and resources are wasted redoing empirical work* [1]. It will also help governments modernize their data management practices and building policies based on evidence and science [3]. Too often, scientific data and outputs cannot be easily discovered, even if publicly available, which leads to the reproducibility crisis of empirical science, thereby threatening its legitimacy and utility [12,22]. Automatically detecting dataset references is challenging even within one research community because of a wide variety of dataset citations and the variety of places in which datasets can be referenced in articles [14].

A significant effort towards this problem were made in the Rich Context Competition [4] (RCC). This paper improves the previously used state-of-the-art approaches for dataset extraction from scientific publications by proposing an end-to-end pipeline. Our approach consists of two stages: *(1) Dataset Sentence Classification, (2) Identification of Actual Dataset Mentions within that sentence*. To the best of our knowledge, our approach is novel in this domain.

## 2 Related Work

Researchers have long investigated extracting entities, artifacts from research paper full text to make knowledge computable [23,25,28]. However, here in this work, we concentrate on the investigations that specifically address dataset extraction and discovery. Recently Google released their Dataset Discovery engine [26] which relies on an open ecosystem, where dataset owners and providers publish semantically enhanced metadata on their sites. Singhal et al. [32] leverage on a user profile-based search and a keyword-based search from open-source web resources such as scholarly articles repositories and academic search engines to discover the datasets. Lu et al. [21] extracted dataset from publications using handcrafted features. Ghavimi et al. [15] proposed a semi-automatic three-step approach for finding explicit references to datasets in social sciences articles. To identify references to datasets in publications, Katarina Boland et al. [8] proposed a pattern induction approach to induce patterns

iteratively using a bootstrapping strategy. The task of identifying biomedical dataset is addressed by [9] open source biomedical data discovery system called DataMed. Within the RCC challenge [2], the winner was the Semantic Scholar team from Allen AI [18]. They built a rule-based extraction system with Named Entity Recognition (NER) model using Bidirectional Long Short-Term Memory (Bi-LSTM) model with a conditional random field (CRF) decoding layer to predict dataset mentions. The honorable mention KAIST team [17] used a machine-learning-based question answering system for retrieving data sets by generating questions regarding datasets. Another finalist, team GESIS [27] also explored a named entity recognition (NER) approach using SPACY for full text. The DICE team [24] from Paderborn University trained an entity extraction model based on CRFs and combined it with a simple dataset mention search to detect datasets in an article. The team from Singapore Management University (SMU) [30] used SVM for dataset detection followed by rules to extract dataset names. The work reported in [29,33] by SU and NUS describes a method for extracting dataset-mentions using various BiLSTM variants with CRF attention models for the dataset extraction task.

The previous works have some limitations in generalizing unseen datasets, discriminating ambiguous names to datasets, and reducing noise. Our current work aims to tackle the limitation and improve the results by combining the transfer capabilities of Bi-Directional Encoder Representations from Transformers (BERT).

## 3 Methodology

RCC organizers provided a labeled corpus of 5000 publications with an additional development fold of 100 publications. Overall, there are around 8 lakhs and 32k sentences, not containing dataset mention and dataset mention, respectively. Each publication was labeled to indicate which of the datasets from the list were referenced within and what specific text was used to refer to each dataset. However, many of the listed datasets do not appear in the corpus. We consider only those publications that contain a mention of the dataset and filtered out the rest for training the dataset-mention extraction model.

We employ a pipeline of two tasks in sequence: Dataset Sentence Classification, followed by Dataset Mention Extraction, as shown in Fig. 1. The sentences that contain dataset mentions are considered further for the dataset mention extraction task. The first task helps us quickly filter out the sentences that do not refer to any dataset.

### 3.1 Dataset-Sentence Classification

We propose a SciBERT+MLP model (a sentence-level binary classifier), which encodes hidden semantics and long-distance dependency. In this module, the goal is to classify each sentence in a sequence of $n$ sentences in a document to find out whether it contains a dataset reference or not. For this purpose, we

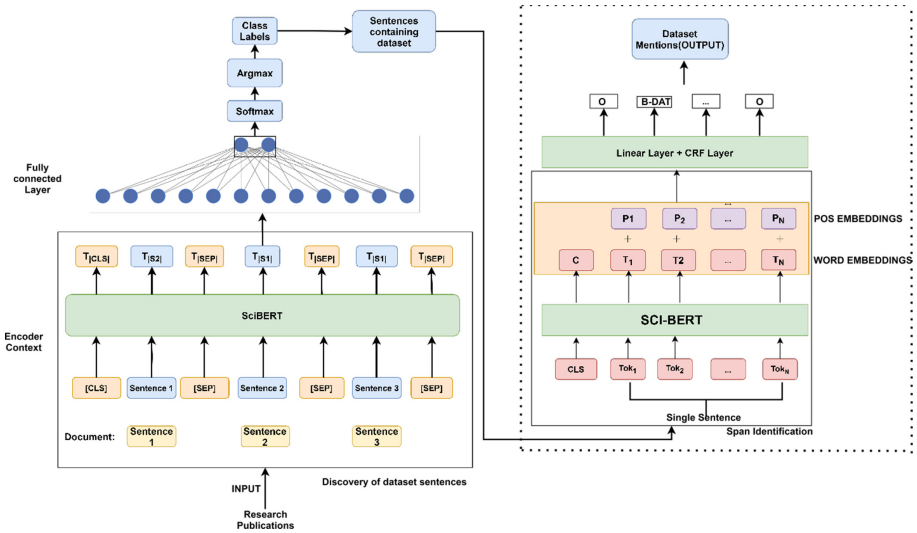


**Fig. 1.** Overall architecture diagram showing: (a) Dataset Sentence Classification (on the left), (b) Dataset Mention Extraction (on the right)

develop a technique based on the Sequential Sentence Classification [6] (SSC) model. The SSC model is based on SciBERT [7], a variant of BERT [10] pre-trained on a large multi-domain corpus of scientific publications. Figure 1(a) gives an overview of our dataset sentence identification module. Consider the training dataset as $\mathrm{T} = D_1, D_2, .., D_i, .., D_Z$ comprising of Z documents. Each $D_i$ can be represented as $D_i = s_{i1}, s_{i2}, .., s_{ij}, .., s_{iN}$ where N is the number of sentences in the document and $s_{ij}$ is the $j^{th}$ sentence of document $D_i$. Each sentence is assigned a ground-truth label where label "1" represents a sentence containing dataset mention reference and label "0" a sentence doesn't contain dataset mention reference. The standard [CLS] is inserted as the first token of the sequence, and another delimiter token [SEP] is used for separating the segments. The initial input embedding ($\mathrm{E}_{Tok}$) is calculated by summing up the token, sentence, and positional embedding. The transformer layers [11] allow the model to fine-tune the weights of these special tokens according to the task-specific training data (RCC corpus). We use a multi-layer feedforward network on top of each sentence's [SEP] representations to classify them to their corresponding categories (Has Dataset Mention or Not?). During fine-tuning, the model learns appropriate weights for the [SEP] token to capture contextual information and learns sentence structure and relations between continuous sentences (through the next sentence objective). Further, we use a softmax classifier on top of the MLP to predict the label's probability. The last linear layer consists of two units corresponding to label "0" and label "1". The final output label is the label whose corresponding unit has a higher score in the last linear layer. Our loss function is weighted binary cross entropy loss, whose weights are decided by

the number of samples in each class. We use the AllenNLP [13] toolkit for the model implementation. As in prior work [10], for training we use dropout of 0.1, the Adam optimizer for 2–5 epochs, and learning rates of 5e−6, 1e−5, 2e−5, or 5e−5.

### 3.2 Dataset Mention Extraction

Dataset Mention Extraction is a binary sequence tagging task where we classified each token to indicate whether it is part of a dataset mention phrase fragment. Here, the goal is to extract the dataset mentions from the sentences which contain at least one mention of the dataset. To detect the boundary of a dataset mention, we use the BIO tagging scheme[1]. We finetune the pre-trained SciBERT model using the annotated corpus with the BIO-schema for dataset mention recognition. While BERT has its tokenization with Byte-Pair encoding and will assign tags to its extracted tokens, we should take care of it. BERT extracted tokens are always equal to or smaller than our main tokens because BERT takes tokens of our dataset one by one, as described by [31]. As a result, we will have intra-tokens that take X tag (meaning don't mention). We employ masking to ignore the padded elements in the sequences.

To add syntactic features to the BERT model, we create a syntax-infused vector for each word by adding a POS embedding vector of dimension d = D to the BERT embedding vector of the word. To determine the POS label of each word of a sentence, we use the pretrained spacy model [5]. We make a POS embedding vector from the BERT embedding of the POS label of the word. Here D is the input dimension of the encoder (D = 768). We add a token-level classifier on top of the BERT layer followed by a Linear-Chain CRF to classify the *dataset mention tokens*. For an input sequence of n tokens, BERT outputs an encoded token sequence with hidden dimension *H*. The classification model projects each token's encoded representation to the tag space, i.e. $\mathbb{R}^H$-> $\mathbb{R}^K$ where $K$ is the number of tags and depends on the number of classes and the tagging scheme. The output scores $\mathbf{P} \in \mathbb{R}^{n \times K}$ of the classification model are then fed to the CRF layer. The matrix $\mathbf{A}$ is such that $A_{i,j}$ represents the score of transitioning from tag $i$ to tag $j$ including two more additional states representing start and end of sequence.

As described by [20] for an input sequence $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)$ and a sequence of tag predictions $\mathbf{y} = (y_1, ..., y_n), y_i \in \{1, ..., K\}$ the score of the sequence is defined as:-

$$s(\mathbf{X}, \mathbf{y}) = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} P_{i, y_i} \tag{1}$$

where $y_0$ and $y_{n+1}$ are the start and end tags. A softmax over all possible sequences yields the probability for sequence $\mathbf{y}$

$$p(\mathbf{y}|\mathbf{X}) = \frac{e^{s(\boldsymbol{X}, \boldsymbol{y})}}{\sum_{\tilde{\boldsymbol{y}} \in \boldsymbol{Y}_{\boldsymbol{X}}} e^{s(\boldsymbol{X}, \tilde{\boldsymbol{y}})}} \tag{2}$$

[1] B, I, and O denote the beginning, intermediate, and outside of dataset mention.

The model is trained to maximize the log probability of the correct tag sequence:-

$$log(p(\mathbf{y}|\mathbf{X})) = s(\mathbf{X}, \mathbf{y}) - log(\sum_{\hat{\boldsymbol{y}} \in \boldsymbol{Y_X}} e^{s(\mathbf{X}, \hat{\boldsymbol{y}})}) \quad (3)$$

where $\boldsymbol{Y_X}$ are all possible tag sequences. Equation 3 is computed using dynamic programming. During evaluation, the most likely sequence is obtained by Viterbi decoding. As per [10] we compute predictions and losses only for the first sub-token of each token. While we tried different batch sizes and learning rates for fine-tuning while we report the best. We use a learning rate of 5e-6 for the Adam optimizer [19], with a batch size of 16 for 10 epochs. Gradient clipping was used, with a max gradient of 1. This module's output will be the BIO-tagged sentence from where we can extract the B followed by I-tagged tokens signifying the dataset mention.

**Table 1.** Result of Dataset-Sentence Classification (P → Precision, R → Recall, F1 → F1-Score)

| Result of Task-1 | P | R | F1 |
|---|---|---|---|
| Sentence not containing data | 0.99 | 0.98 | 0.99 |
| Sentence containing data | 0.83 | 0.82 | 0.83 |
| Macro average | 0.92 | 0.91 | 0.91 |

**Table 2.** Result of Dataset Mention Extraction, Details of each of these comparison systems is described in Sect. 2

| Model | Partial match | | | Exact match | | |
|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 |
| SMU [30] | – | – | – | 34.0 | 30.0 | 32.0 |
| BiLSTM(NUS) [29] | 71.4 | 64.4 | 67.7 | 31.3 | 34 | 32.6 |
| SL-E-C(NUS) [29] | 72.2 | 72.6 | 74.8 | 39.9 | 41.6 | 40.7 |
| CNN-BiLSTM(NUS) [29] | 77.5 | 75.5 | 76.5 | 41.4 | 44.6 | 43.0 |
| CNN-BiLSTM-CRF(NUS) [29] | 79.1 | 71.1 | 74.9 | 42.7 | 44.6 | 43.6 |
| SU [33] | 88.2 | 88.4 | 88.3 | – | – | – |
| CNN-BiLSTM-Att-CRF(NUS) [29] | 76.1 | 73.8 | 74.9 | 39.4 | 47.7 | 43.2 |
| Allen AI [18] | – | – | – | 52.4 | 50.3 | 51.8 |
| **Our model** | **89.2** | 88.1 | **88.6** | **60.24** | **52.8** | **56.2** |
| GESIS [27] | 93.0 | 95.0 | 93.8 | 80.0 | 81.0 | 80.4 |
| **Our model** | **94.2** | **95.2** | **94.6** | **85.2** | **86.7** | **85.9** |

## 4 Results and Analysis

Table 1 shows the result of Task-1 (Dataset Sentence Classification). Our model has reported a 0.91 macro average for Task 1. While Table 2 shows the result of

Task-2 (Dataset Mention Extraction) and the comparison with other baselines. We evaluate our model for strict and partial (relaxed) F1-score. While strict criterion contributes a true positive count if and only if the ground truth tokens are exactly predicted, whereas matched correctly predicted assigns the credit if and only if the exact boundaries are matched, for partial (or relaxed) criterion, a partial match to the ground truth is also treated as the true positive count.

As expected, our proposed model results are better for the partial match than the exact match, which means we can find the proper context with very high precision even if we could not match the full dataset mention in the text exactly. Results also show that our proposed system performs the best for both strict and relaxed evaluation metrics than the other existing methods. The closest system, AllenAI [18], reported having achieved the F1-scores of 51.8 for the strict. We observe a relative improvement of 6.4% F1-score compared to AllenAI wrt strict. The closest system, GESIS [27], reported having achieved the F1-scores of 80.4 for the strict and 93.8 for the relaxed criterion, respectively. We observe a relative improvement of 5.4% F1-score compared to GESIS wrt strict and almost equal F1-score to relaxed criterion (All results are on the development set while GESIS divided the training set into the split of 80:20, 80% for training and 20% for testing; we also report for the same set). The other participants, including AllenAI [18] and GESIS [27], have not tried transformer-based NER and have also performed NER on the paper's full context. In contrast, we filtered out the irrelevant sentences (not containing the dataset mention) and then used the relevant sentences for mention extraction. Also, the BERT-based NER understood the context better, resulting in better results. We also perform test-of-significance (T-test) and observe that the obtained results are statistically significant w.r.t. the state-of-the-art with p-values $< 0.05$.

**Table 3.** Ablation study

| Model | Precision | Recall | F1 |
|---|---|---|---|
| BERT+CRF (On full test set) | 54.9 | 51.2 | 52.9 |
| BERT+CRF (After dataset sentence classification) | 58.2 | 51.3 | 54.4 |
| BERT+CRF+POS (On full test set) | 55.2 | 52.3 | 53.7 |
| BERT+CRF+POS (After dataset sentence classification) | 60.2 | 52.8 | 56.2 |

### 4.1 Analysis

Table 3 shows the ablation study examining our system's various components' importance. We observe dataset sentence classification before dataset mention extraction, and POS-aware BERT embedding for dataset mention extraction boosts the overall model's performance for this task.

**Table 4.** Examples of the dataset sentence identification task, where the red coloured text indicate sentences being filtered out whereas blue colored text indicate sentences passed for the next dataset mention extraction task.

| |
|---|
| ***Dataset Name: "SWAN' Title: "Study of Women's Health across the Nation"*** |
| 1. Swans are bird of family Anatidae within the genus Cygnus (-) |
| 2. Several enduring themes have emerged from SWAN that have associated certain patterns of hormones and symptoms with metabolic status. (+) |
| 3. SWAN Energy LTd. is an emerging "green energy" company (-) |
| 4. Weight gain has been observed to occur in conjunction with the menopausal transition, but studies prior to SWAN have concluded that weight gain is driven primarly by age (+) |
| ***Dataset Name: "SUPPORT' Title: "Study to understand Prognoses and Preferences for Outcomes and Risks of Treatments"*** |
| 5. Peng's findings do not appear to support his conclusions (-) |
| 6. Phase I of SUPPORT collected data from patients accessioned during 1989-1991 to characterize the care, treatment preferences, and pattern of decision-making among critically ill patients. (+) |

**Role of Sentence Identification.** As the string may occur multiple times in the document, and all occurrences may or may not be correct dataset mentions; this is especially problematic when the string is a common word which may have multiple meanings in different contexts. As shown in Table 4, we provide some examples to show how the sentence identification task can overcome other participants' limitations, including that of GESIS. 'SWAN' is a dataset mention of a dataset with the title "Study of Women's Health Across Nation," which is also the name of a bird, company, etc. Similarly, 'SUPPORT' is a dataset mention of a dataset with the title "Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments." However, it is also a commonly used word in the English language with a different meaning. Using NER directly does not discriminate these confusion cases and mislabels all of them as dataset names. While the sentence identification task understands the context of the sentences and filters out these irrelevant sentences (red), and preserves only relevant sentences (blue) before feeding them to NER.

**Table 5.** Examples showing the use of adding POS embedding to word embedding (red: wrongly identified dataset mention, blue: correctly identified dataset mention)

| **Without POS embedding** | **With POS embedding** |
|---|---|
| The data has been used from progress in international reading literacy study.... There is a high level of risk in the rise in the number of cases. | The data has been used from progress in international reading literacy study.... There is a high level of risk in the rise in the number of cases. |
| We combine two micro datasets provided by the deutsche bundesbank ,the its and the midi and complement them[..] | We combine two micro datasets provided by the deutsche bundesbank ,the its and the midi and complement them [..] |

**Role of POS Embedding.** Dataset mentions are usually noun phrases, such as in Table 5 *"National health and educational survey"*, *"coastal erosion study"*, etc. The examples *"progress in"* and *"rise in"* are misclassified by the NER, as the dataset mentions. However, adding the POS embedding gives more weightage to the noun chunks. Hence, some misclassified verbs or other POS dataset phrases are reduced.

### 4.2 Error Analysis

- **Roman numbers:** Our model finds difficulty in determining full dataset names having roman names. For example *"[..]add health (waves i, ii, and iii) with obesity[..]"*, contains roman letters in the dataset name (*"add health and add health waves i ii and iii"*). However, the model predicts only *add health*, i.e., does not predict the full dataset name.
- **Too many numbers or punctuations:** Our model confuses when there are too many numbers or punctuations in the sentence. For example *"002 hospital beds per 100,000 population −0:002***[..] national profile of local health departments[..]"* shows the example having the dataset mentions *"national profile of local health departments,"* but the model fails to understand the context due to many punctuation or numbers, hence fails to predict the dataset name.

## 5 Conclusion and Future Work

In this work, we report a novel BERT-based model for extracting dataset mentions from scientific publications. Our model is simple and outperforms earlier approaches. Our overall goal is to understand the impact of any given dataset (*Data Impact Factor*) in the community. The critical observation we make here is that *identifying sentences containing the dataset-mentions are highly useful before proceeding with the task of dataset-mention extraction* and using BERT with POS embedding can enhance the task of dataset-mention extraction. In the future, we intend to explore extracting other helpful information (tasks, methods, metrics) from research publications to automate automated literature comparison.

**Acknowledgement.** Sandeep Kumar acknowledges the Prime Minister Research Fellowship (PMRF) program of the Government of India for its support. Asif Ekbal is a recipient of the Visvesvaraya Young Faculty Award and acknowledges Digital India Corporation, Ministry of Electronics and Information Technology, Government of India for supporting this research.

## References

1. The coleridge initiative announces rich context competition—NYU cusp. https://cusp.nyu.edu/blog/the-coleridge-initiative-announces-rich-context-competition/. Accessed 14 July 2021
2. Github - rich-context-competition/rich-context-book-2019. https://github.com/rich-context-competition/rich-context-book-2019. Accessed 14 July 2021
3. Rich context project - coleridge initiative. https://coleridgeinitiative.org/rich-context-project/. Accessed 14 July 2021
4. Richcontextcompetition - coleridge initiative. https://coleridgeinitiative.org/richcontext/richcontextcompetition/. Accessed 14 July 2021
5. Spacy industrial-strength natural language processing in python. https://spacy.io/. Accessed 15 July 2021
6. Cohan, A., Beltagy, I., King, D., Dalvi, B., Weld, D.S.: Pretrained language models for sequential sentence classification. In: EMNLP (2019)

7. Beltagy, I., Lo, K., Cohan, A.: SciBERT: a pretrained language model for scientific text. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, 3–7 November 2019, pp. 3613–3618. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/D19-1371
8. Boland, K., Ritze, D., Eckert, K., Mathiak, B.: Identifying references to datasets in publications. In: Zaphiris, P., Buchanan, G., Rasmussen, E., Loizides, F. (eds.) TPDL 2012. LNCS, vol. 7489, pp. 150–161. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33290-6_17
9. Chen, X., et al.: DataMed - an open source discovery index for finding biomedical datasets. J. Am. Medical Informatics Assoc. **25**(3), 300–308 (2018)
10. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding (2019)
11. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, 2–7 June 2019 (Long and Short Papers), vol. 1, pp. 4171–4186. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/n19-1423
12. Feger, S.S.: Interactive tools for reproducible science - understanding, supporting, and motivating reproducible science practices. CoRR abs/2012.02570 (2020). https://arxiv.org/abs/2012.02570
13. Gardner, M., et al.: AllenNLP: a deep semantic natural language processing platform (2018). http://arxiv.org/abs/1803.07640
14. Ghavimi, B., Mayr, P., Lange, C., Vahdati, S., Auer, S.: A semi-automatic approach for detecting dataset references in social science texts. Inf. Serv. Use **36**(3–4), 171–187 (2016)
15. Ghavimi, B., Mayr, P., Vahdati, S., Lange, C.: Identifying and improving dataset references in social sciences full texts. In: Loizides, F., Schmidt, B. (eds.) Positioning and Power in Academic Publishing: Players, Agents and Agendas, 20th International Conference on Electronic Publishing, Göttingen, Germany, 7–9 June 2016, pp. 105–114. IOS Press (2016). https://doi.org/10.3233/978-1-61499-649-1-105
16. Grover, M.: Amundsen - Lyft's data discovery & metadata engine—by mark grover—Lyft engineering, April 2019. https://eng.lyft.com/amundsen-lyfts-data-discovery-metadata-engine-62d27254fbb9. Accessed 31 Oct 2020
17. Hong, G., Cao, M.S., Puerto-San-Roman, H.: Rich text competition. In: Rich Search and Discovery for Research Datasets: Building the Next Generation of Scholarly Infrastructure. Sage, London (2020)
18. King, D., Ammar, W., Beltagy, I., Betts, C., Gururangan, S., van Zuylen, M.: The AI2 submission at the rich context competition. In: Rich Search and Discovery for Research Datasets: Building the Next Generation of Scholarly Infrastructure. Sage, London (2020)
19. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization (2015). http://arxiv.org/abs/1412.6980
20. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. CoRR abs/1603.01360 (2016). http://arxiv.org/abs/1603.01360

21. Lu, M., Bangalore, S., Cormode, G., Hadjieleftheriou, M., Srivastava, D.: A dataset search engine for the research document corpus. In: 2012 IEEE 28th International Conference on Data Engineering, pp. 1237–1240. IEEE (2012)
22. Munafò, M., et al.: A manifesto for reproducible science. Nat. Hum. Behav. **1**, 0021 (2017). https://doi.org/10.1038/s41562-016-0021
23. Nasar, Z., Jaffry, S.W., Malik, M.K.: Information extraction from scientific articles: a survey. Scientometrics **117**(3), 1931–1990 (2018)
24. Ngonga, P.D.A., Srivastava, N., Jalota, R.: Dice @ rich context competition. In: Rich Search and Discovery for Research Datasets: Building the Next Generation of Scholarly Infrastructure. Sage, London (2020)
25. Nguyen, T.D., Kan, M.-Y.: Keyphrase extraction in scientific publications. In: Goh, D.H.-L., Cao, T.H., Sølvberg, I.T., Rasmussen, E. (eds.) ICADL 2007. LNCS, vol. 4822, pp. 317–326. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-77094-7_41
26. Noy, N., Burgess, M., Brickley, D.: Google dataset search: building a search engine for datasets in an open web ecosystem. In: 28th Web Conference (WebConf 2019) (2019)
27. Otto, W., Zielinski, A., Ghavimi, B., Dimitrov, D., Tavakolpoursaleh, N.: Rich context competition phase 2. In: Rich Search and Discovery for Research Datasets: Building the Next Generation of Scholarly Infrastructure. Sage, London (2020)
28. Peng, F., McCallum, A.: Information extraction from research papers using conditional random fields. Inf. Process. Manag. **42**(4), 963–979 (2006)
29. Prasad, A., Si, C., Kan, M.Y.: Dataset mention extraction and classification. In: Proceedings of the Workshop on Extracting Structured Knowledge from Scientific Publications, Minneapolis, Minnesota, pp. 31–36. Association for Computational Linguistics, June 2019. https://doi.org/10.18653/v1/W19-2604. https://www.aclweb.org/anthology/W19-2604
30. Prasetyo, P.K., Silva, A., Lim, E.P., Achananuparp, P.: Simple extraction for social science publications. In: Rich Search and Discovery for Research Datasets: Building the Next Generation of Scholarly Infrastructure. Sage, London (2020)
31. Shamsfard, M., Jafari, H.S., Ilbeygi, M.: Step-1: a set of fundamental tools for Persian text processing. In: Calzolari, N., et al. (eds.) Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, Valletta, Malta, 17–23 May 2010. European Language Resources Association (2010). http://www.lrec-conf.org/proceedings/lrec2010/summaries/809.html
32. Singhal, A., Srivastava, J.: Research dataset discovery from research publications using web context. In: Web Intelligence, vol. 15, pp. 81–99. IOS Press (2017)
33. Zeng, T., Acuna, D.: Dataset mention extraction in scientific articles using a BiLSTM-CRF model. In: Rich Search and Discovery for Research Datasets: Building the Next Generation of Scholarly Infrastructure. Sage, London (2020)