Simple Extraction for Social Science Publications

Philips Kokoh Prasetyo, Amila Silva, Ee-Peng Lim, Palakorn Achananuparp
Living Analytics Research Centre
Singapore Management University
{pprasetyo,amilasilva,eplim,palakorna}@smu.edu.sg

Abstract

With the vast number of datasets and literature collections available for research today, it is very difficult to keep track on the use of datasets and literature articles for scientific research and discovery. Many datasets and research work using them are left undiscovered and under-utilized due to the lack of available search tools to automatically find out who worked with the data, on what research topics, using what research methods and generating what results. The Coleridge Rich Context Competition (RCC) therefore aims to build automated dataset discovery tools for analysing and searching social science research publications. In this paper, we describe our approach to solving the first phase of Coleridge Rich Context Competition.

1 Introduction

Automated discovery from scientific research publications is an important task for analysts, researchers, and learners as they develop the scientific knowledge and use them to gain new insights. More specifically, on the tasks of discovering datasets and methods mentioned in a research publication, we have seen a lack of available tools to easily find who else worked on a particular dataset, what research methods people apply on the dataset, and what results they have found using the dataset. Furthermore, new datasets are not easy to discover, and as a result, good datasets and methods are often neglected.

The Coleridge Rich Context Competition (RCC) aims to build automated datasets discovery from social science research publications, filling the gap of this problem. In this competition, given a corpus of social science research publications, we have to automatically identify datasets used, and then infer the research methods and research fields in the publications. Note that no labeled data are given for research methods and fields identification.

This manuscript describes summary of our submission for the first phase of RCC. We begin with related work in section 2. We present our analysis on RCC dataset in section 3, describe our approach in section 4, and discuss our experiment results in section 5. Finally, we wrap up with conclusion and future work in section 6.

2 Related Work

Extracting information from scientific text has been explored in the past (Peng and McCallum, 2004; Nguyen et al., 2015; Singh et al., 2016). One type of information extraction from scientific articles is extracting keyphrases and relation between them (Augenstein et al., 2017). Luan et al. (2017) propose semi-supervised sequence tagging approach to extract keyphrases. Augenstein and Søgaard (2017) explore multi-task deep recurrent neural network approach with several auxiliary tasks to extract keyphrases.

Another type of extraction is citation extraction. Two citation extraction settings have been explored before: reference mining inside the full text (Alves et al., 2018), and citation metadata extraction (Hetzner, 2008; Anzaroot et al., 2014; An et al., 2017). Nasar et al. (2018) write a survey on information extraction from scientific articles.

Recently, there are some work to explore dataset extraction from scientific text (Boland et al., 2012; Ghavimi et al., 2016a,b). Boland et al. (2012) propose weakly supervised pattern induction to identify references in social science publications. Ghavimi et al. (2016a,b) propose a semi automatic approach for detecting dataset references for social science texts. Dataset extraction is a challenging task because of the inconsistency and wide range of dataset mention styles in research publi-

cations (Ghavimi et al., 2016b).

3 Data Analysis

The first phase of RCC dataset consists of a labeled corpus of 5,000 publications for training set, and additional 100 publications for development set. The RCC organizer keeps a separate corpus of 5,000 publications for evaluation. Each article in the dataset contains full text article and dataset citation labels. The metadata of cited datasets in the corpus are also provided. For research methods and fields, no label information is provided, only SAGE social science research method graph and research fields vocabulary are provided.

Preprocessing. In order to reliably access important structures of paper publications, we parse all papers using AllenAI Science Parse[1] (Ammar et al., 2018). AllenAI Science Parse reads PDF file, and returns title, authors, abstract, sections, and bibliography (references). Since this parser utilizes machine learning models to parse PDF file, the parsing results may not be 100% accurate. Furthermore, this parser is unable to parse scan copy of old publication. In the situation where we are unable to access parsed fields, we fall back to the given text files.

Mention Analysis. There are 5,499 and 123 dataset citations in training and development set respectively. Among these citations, 320 citations in training set and 6 citations in development set do not have mentions information. We analyze the paper sections where the dataset mentions commonly occur. Table 1 and 2 show top 12 most common sections mentioning dataset in training and development set. The tables suggest that abstract, reference titles, discussion, results, and methods are the most common sections where the dataset mentions occur. We exploit reference titles for dataset extraction.

Citation Analysis. We build citation network from training set. Each node in the network is a paper publication, and an edge between two node A and B is generated if a paper A cites paper B. Table 3 shows the statistics of the citation network.

Initially, we propose an approach utilizing citation network based on an intuition that datasets, research methods, and research fields are shared by: 1) same or similar issues, 2) same or similar context, 3) same or similar authors and communities, 4) same or similar metrics used in the publication. However, based on table 3, we learn that exploring rich context using paper-paper citation network is not viable at this stage because most papers listed in publications' bibliography are not available in the training set, and therefore, paper-paper citation network becomes very sparse with many unknown information. Due to this reason, we drop our idea on utilizing paper-paper citation graph at this stage. Nevertheless, we believe that bibliography contains important signals and information about datasets, and research fields.

[1] https://github.com/allenai/science-parse

Section Header	Mention Frequency
Abstract	2,548
Reference Titles	1,997
Discussion	1,390
Results	836
Methods	804
Introduction	530
Statistical Analysis	285
Comment	279
Acknowledgements	261
Materials and Methods	254
Study Population	227
Data	214

Table 1: Top 12 Sections Mentioning Datasets in Training Set

Section Header	Mention Frequency
Abstract	78
Reference Titles	37
Discussion	19
Introduction	14
Results	12
Statistical Analyses	9
Methods	8
Ethics	7
Population	7
Population Impact	7
Price	7
2.1 Data	5

Table 2: Top 12 Sections Mentioning Datasets in Development Set

Number of nodes	5,000
Number of edges	998
Network density	0.008%

Table 3: Statistics of Citation Network

4 Methods

In this section, we describe our approach for RCC tasks: dataset extraction, research methods identification, and research fields identification.

4.1 Dataset Extraction

We employ a pipeline of two subtasks for dataset extraction: dataset detection, followed by dataset recognition. The goal of dataset detection is to detect whether a publication cites a dataset or not. This first subtask helps us to quickly filter out non-dataset publications. After the first subtask, we mine dataset mentions for the remaining publications in dataset recognition subtask.

For dataset detection, we utilize paper title in bibliography (reference list) combined with explicit research methods mentions to detect whether a publication citing a dataset or not. Explicit research methods mentions are determined based on exact match between paper title and SAGE research methods vocabulary. We train an SVM classifier using explicit research method mentions and n-gram features from paper titles in bibliography. We use the SVM classifier to classify each publication, if the classifier gives positive label, then we proceed to dataset recognition subtask, otherwise we ignore the publication.

For dataset recognition, we use an implicit entity linking approach. We start with the Naive Bayes model, which can be regarded as a standard information retrieval baseline, and entity indicative weighting strategy is used to improve the model. In order to calculate the word distribution of each dataset, we represent each dataset using its title, dataset mentions (provided in the training set), and dataset relevant sentences, filtered from the relevant publications using the rule based approach proposed in Ghavimi et al. (2016b). All these text sections related to a particular dataset are considered as a single text chunk, and we calculate the word distribution as follows. Let $\mathbf{w}$ be the set of words in a dataset. In our problem setting, we assume the dataset prior probability $p(d)$ to be uniform. The probability of dataset d given $w \in \mathbf{w}$ is:

$$\begin{aligned} p(d|\mathbf{w}) &\propto \prod_{w \in \mathbf{w}} p(w|d) \\ &= \prod_{w \in \mathbf{w}} \frac{f(d,w) + \gamma}{\sum_{w'} f(d,w') + |W|\gamma} \end{aligned} \quad (1)$$

where $f(d,w)$ is the number of co-occurrences of word w with entity d, γ is the smoothing parameter, and $|W|$ is the vocabulary size. For each dataset d, we derive $f(d,w)$ by the count of w occurrences in the text extracted for each dataset. In order to stress more priority for dataset indicative words, we improved the final objective function of our model as follows:

$$ln(p(d|\mathbf{w})) \propto \sum_{w \in \mathbf{w}} \beta(w) * ln(p(w|d)) \quad (2)$$

where $\beta(w)$ is the entity-indicative weight for word w. This weight $\beta(w)$ is added as an exponent to the term $p(w|d)$. $\beta(w)$ is calculated as:

$$\beta(w) = log(1 + E/df(w)) \quad (3)$$

where E is the number of distinct datasets considered and $df(w)$ counts the number of datasets with at least one occurrence of w.

Then for a given unseen publication, we use same rule based approach (Ghavimi et al., 2016b) to filter a few relevant sentences, and datasets are ranked by $ln(p(d|w))$ to select the most suitable datasets. In order to select exact datasets related to particular publication, we select top 10 datasets ranked using above approach. And then the confidence probability related to the top 10 datasets are normalized and select the datasets with the normalized probability higher than a predefined threshold value. We return the entity indicative words as relevant dataset mentions.

4.2 Research Methods Identification

Since we do not have labeled training data for this task, we use explicit research method mentions (based on exact match with SAGE research methods vocabulary) in a publication as weak signals on research methods used in the publication. When these mentions frequently appear in a publication, there is a high chance that this publication is using these particular research methods.

Based on this intuition, we generate training set for research method classification utilizing sentences that explicitly mention research method in a publication. Publication title and the sentences mentioning research method serve as context information of a specific research method. In order to reduce noisy weak signals, we apply minimum support of three sentences in a publication. We

exclude research methods which only being mentioned one or two times in a publication. We also exclude research methods that only being mentioned in less than 10 different publications from the training set. Finally, we have 133 research methods having sufficient context information for training data. This number is 20.18% of 659 research methods in SAGE research method graph.

We use the training data to train logistic regression classifier to classify research methods from publication title and sentences. We utilize n-gram features from publication title and sentences for the classifier. We apply the logistic regression classifier to recommend top 3 research methods based on logistic regression probability score.

This approach can be extended by utilizing research method graph to expand the context. Context information does not only comes from sentences in publication, but also comes from related research methods as well as broader concept information. By using this information, we can potentially expand to more than 133 research methods and perform more accurate prediction.

4.3 Research Fields Identification

Similar to research methods identification, this task does not have labeled training data. We only have access to list of SAGE research fields. SAGE research fields are organized hierarchically into three levels, namely L1, L2, and L3, for example: Soc-2-4 (*kinship*) is under Soc (*sociology*) in L1, and under Soc-2 (*anthropology*) in L2.

To gain more understanding about the characteristic of each field, we crawl top search results from SAGE Knowledge[2]. From the search result snippets, we collect information such as title and abstract on various publications including case, major work, books, handbooks, and dictionary. We exclude video and encyclopedia. Due to sparseness of the SAGE Knowledge, we exclude all research fields with less than 10 search results. In the end, we have samples of 414 L3 research fields under 101 L2 research fields and 10 L1 research fields. This numbers cover 20.87% of 1,984 L3 research fields, and 67.79% of 149 L2 research fields in the list of SAGE research fields. We use this data to train research fields classifiers.

We build three SVM classifiers for L1, L2, and L3 to classify a publication using paper title and abstract. Instead of taking the highest score, we take top-k research fields and perform re-ranking considering agreement among L1, L2, L3. We return a research field if its upper level are also in top ranks. Since level L1 is too general, we only output research fields from L2, and L3. We outline our heuristic to reorder the ranking below:

1. Get top-5 L3 research fields, top-4 L2 research fields, and top-3 L1 research fields.

2. Assign initial score v for each research field based on its ranking.

$$v(f_i) = (K - i)/K \tag{4}$$

   where K is the length of top-k, and i is the ranking of a research field f. For example, research fields in top-5 L3 have initial score of $[1, 0.8, 0.6, 0.4, 0.2]$, top-4 L2 have initial score of $[1, 0.75, 0.5, 0.25]$, and top-3 L1 have $[1, 0.666, 0.333]$

3. Update the score by multiplying each score with the score of matching research fields at upper level, and 0 otherwise.

$$score(f_i^l) = \begin{cases} \prod_{l \in L} v(f^l) & \text{if field matched} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

   where L is the level of research field f and its upper levels. Here are examples of score update:

   - Soc-2-4 at rank-2 in L3, Soc-2 at rank-3 in L2, and Soc at rank-1 in L1. In this case, the score of Soc-2-4 is $0.8 * 0.5 * 1 = 0.4$.
   - Soc-2-4 at rank-1 in L3, Soc-2 at rank-2 in L2, but Soc is not found in top rank in L1. In this case, the score of Soc-2-4 is 0.

4. Collect score from L2 and L3, and exclude L2 if we see more specific of L2 in top-5 L3.

5. Re-rank L2 and L3 research fields based on the score.

6. Return research fields having score $>= 0.4$.

To expand to more context from paper list in bibliography section, we also build other three Naive Bayes classifiers for L1, L2, and L3 using

[2]http://sk.sagepub.com/browse/

Classifier	Prec.	Rec.	F1
Naive Bayes	0.85	0.92	0.88
SVM	0.96	0.88	0.92

Table 4: Dataset Detection Results on Development Set

Method	Prec.	Rec.	F1
No Dataset Detection	0.18	0.33	0.24
With Dataset Detection	0.34	0.30	0.32

Table 5: Dataset Extraction Results on Development Set

Dataset	Prec.	Rec.	F1
Test Set (phase1)	0.17	0.10	0.13

Table 6: Dataset Extraction Result on Test Set

paper title feature only. We believe that a publication from a certain field also cites other publications from same or similar fields. For each publication in the bibliography, we apply the same procedure as mentioned above, then we average the score to get top research fields from bibliography. Finally, we combine top research fields from paper titles and abstract with results from bibliography.

5 Experiment Results

We discuss our experiment results for each task in this section. We use standard precision, recall, and F1 as evaluation metrics.

Dataset Extraction. First, we analyze our experiment for dataset detection subtask comparing Naive Bayes and SVM classifier. Using only paper titles in bibliography and explicit research method mentions, Naive Bayes and SVM classifiers are able to reach 0.88 & 0.92 F1 score respectively. Since SVM outperforms Naive Bayes, we use SVM for our dataset detection module. Table 4 shows detail dataset detection results on development set.

To see the impact of performing dataset detection, we test the performance of dataset extraction with and without dataset detection on development set. Table 5 summarizes the results. As shown in the table, performing dataset detection before extraction significantly improves the dataset extraction on development set.

Table 6 shows dataset extraction performance on test set (phase 1). The significant drop from development set result suggests that the test set might have different distribution compare to the training and development set. It might also contain dataset citations that are never been seen in training set.

Classifier	F1
Naive Bayes	0.55
Logistic Regression	0.86

Table 7: F1 Score for Research Method Classification

Classifier	L1	L2	L3
Naive Bayes	0.78	0.37	0.13
SVM	0.82	0.38	0.12

Table 8: F1 Score for Research Field Classification on Publication Level using Paper Title and Abstract

Classifier	L1	L2	L3
Naive Bayes	0.80	0.35	0.12
SVM	0.81	0.35	0.11

Table 9: F1 Score for Research Field Classification on Bibliography Level using Paper Title Only

Research Methods Identification. We only consider Naive Bayes and Logistic Regression classifiers for research method identification because they naturally outputs probability score. We perform 5-fold cross validation to evaluate classification performance, and the result can be seen in table 7. Logistic regression classifier outperforms Naive Bayes with 0.86 F1 score in classifying 133 research methods.

Research Fields Identification. We perform 5-fold cross validation to evaluate our classifiers to classify L1, L2, and L3 research fields. Table 8 shows the results using n-gram features from paper title and abstract, whereas table 9 shows the results using n-gram features from title only. Naive Bayes tends to perform slightly better on L3 research fields where we have large number of research field labels. We decide to use SVM for research field identification on publication level because SVM is generally better than Naive Bayes. On the other hand, we decide to use Naive Bayes for research field identification on bibliography level because Naive Bayes prefer to have more accurate L2 and L3 research fields.

Method	Features (n-gram)
Dataset extraction	
SVM for dataset detection	paper titles in bibliography and explicit research method mentions
Implicit entity linking	paper title and full text
Research method identification	
Logistic regression	paper title, abstract, and full text
Research field identification	
SVM (on paper)	paper title and abstract
Naive Bayes (on bibliography)	paper titles in bibliography

Table 10: Summary of Our Approach

6 Conclusion

Extraction of research datasets, associated research methods and fields from social science publication is challenging, yet an important problem to organize social science publications. We have described our approach for the RCC challenge, and table 10 summarizes our approach. Beside publication content such as paper titles, abstract, full text, our approach also leverages on the information from bibliography. Furthermore, we also collect external information from SAGE Knowledge to get more information about research fields.

Apart from F1 score on 5-fold cross validation, we have no good way to evaluate research method and research field identification without ground truth label. Our methods are unable to automatically extract and recognize new datasets, research methods, and fields. An extension to automatically handle such cases using advance Natural Language Processing (NLP) approach is a promising direction.

From this competition, we have learned that lacks of labelled training data is a huge challenge, and it directs us to other external resources (i.e., SAGE Knowledge) as proxy for our label. Another challenge is data sparsity. Although we see many paper listed in bibliography, lacks of access to these publication make us difficult to exploit citation network.

Unfortunately, our model did not advance to the second phase. We are interested in exploring more advanced information extraction methods on the RCC datasets, and we hope that the organizer will release the RCC datasets for future research. We thank the organizers for organizing a competition and workshop on this important, interesting, and challenging problem.

References

Danny Rodrigues Alves, Giovanni Colavizza, and Frédéric Kaplan. 2018. Deep reference mining from scholarly literature in the arts and humanities. In *Front. Res. Metr. Anal.*

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew E. Peters, Joanna Power, Sam Skjonsberg, Lucy Lu Wang, Chris Wilhelm, Zheng Yuan, Madeleine van Zuylen, and Oren Etzioni. 2018. Construction of the literature graph in semantic scholar. In *NAACL-HTL*.

Dong An, Liangcai Gao, Zhuoren Jiang, Runtao Liu, and Zhi Tang. 2017. Citation metadata extraction via deep neural network-based segment sequence labeling. In *CIKM*.

Sam Anzaroot, Alexandre Passos, David Belanger, and Andrew McCallum. 2014. Learning soft linear constraints with application to citation field extraction. In *ACL*.

Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. 2017. Semeval 2017 task 10: Scienceie - extracting keyphrases and relations from scientific publications. In *SemEval@ACL*.

Isabelle Augenstein and Anders Søgaard. 2017. Multi-task learning of keyphrase boundary classification. In *ACL*.

Katarina Boland, Dominique Ritze, Kai Eckert, and Brigitte Mathiak. 2012. Identifying references to datasets in publications. In *TPDL*.

Behnam Ghavimi, Philipp Mayr, Christoph Lange, Sahar Vahdati, and Sören Auer. 2016a. A semi-automatic approach for detecting dataset references in social science texts. *Inf. Services and Use*, 36:171–187.

Behnam Ghavimi, Philipp Mayr, Sahar Vahdati, and Christoph Lange. 2016b. Identifying and improving dataset references in social sciences full texts. In *ELPUB*.

Erik Hetzner. 2008. A simple method for citation metadata extraction using hidden markov models. In *JCDL 2008*.

Yi Luan, Mari Ostendorf, and Hannaneh Hajishirzi. 2017. Scientific information extraction with semi-supervised neural tagging. In *EMNLP*.

Zara Nasar, S. W. Jaffry, and Muhammad Kamran Malik. 2018. Information extraction from scientific articles: a survey. *Scientometrics*, 117:1931–1990.

Viet Cuong Nguyen, Muthu Kumar Chandrasekaran, Min-Yen Kan, and Wee Sun Lee. 2015. Scholarly document information extraction using extensible features for efficient higher order semi-crfs. In *JCDL 2015*.

Fuchun Peng and Andrew McCallum. 2004. Accurate information extraction from research papers using conditional random fields. In *HLT-NAACL 2004*.

Mayank Singh, Barnopriyo Barua, Priyank Palod, Manvi Garg, Sidhartha Satapathy, Samuel Bushi, Kumar Ayush, Krishna Sai Rohith, Tulasi Gamidi, Pawan Goyal, and Animesh Mukherjee. 2016. Ocr++: A robust framework for information extraction from scholarly articles. In *COLING*.