DICE @ Rich Context Competition 2018 – Combining Embeddings and Conditional Random Fields for Research Dataset, Field and Method Recognition and Linking

Rricha Jalota, Nikit Srivastava, Daniel Vollmers, René Speck, Michael Röder, Ricardo Usbeck[0000−0002−0191−7211], and Axel-Cyrille Ngonga Ngomo

Data Science Group, Paderborn University, Germany
firstname.lastname@uni-paderborn.de

Abstract. The steadily increasing number of publications available to researchers makes it difficult to keep track of the state of the art. In particular, tracking the datasets used, topics addressed, experiments performed and results achieved by peers becomes increasingly tedious. Current academic search engines render a limited number of entries pertaining to this information. However, having this knowledge would be beneficial for researchers to become acquainted with all results and baselines relevant to the problems they aim to address. With our participation in NYU Coleridge Initiative's Rich Context Competition, we aimed to provide approaches to automate the discovery of datasets, research fields and methods used in publications in the domain of Social Sciences. We trained an Entity Extraction model based on Conditional Random Fields and combined it with the results from a Simple Dataset Mention Search to detect datasets in an article. For the identification of Fields and Methods, we used word embeddings. In this paper, we present how our approaches performed, their limitations, some of the challenges we encountered and our future agenda.

1 Introduction

1.1 Rich Context Competition

The goal of the Rich Context Competition[1], organized by the New York University under their Coleridge Initiative, was to automate the discovery of research datasets, associated research methods and fields in research publications belonging to the domain of Social Sciences. It was carried out in two phases. In the first phase (Phase-1), we were provided with a list of datasets along with their metadata (dataset vocabulary), a training corpus of 5000 publications containing publication metadata (2500 of them were labeled) and an additional dev fold of 100 publications. Apart from this, we were also given Social Science Methods and

[1] https://coleridgeinitiative.org/richcontextcompetition

Fields vocabularies by SAGE Publications[2]. To carry out Phase-1 evaluation, a separate corpus of 5000 labeled publications was held back.

In the second phase (Phase-2), in addition to the Phase-1 data for training, we were provided with the Phase-1 holdout set consisting of 5000 labeled publications and an additional corpus of 5000 unlabeled publications. The evaluation of the second phase was carried out by the organizers on another corpus that contained 5000 unlabeled publications. Note that the labeled data in both the phases was for Dataset Detection only. There was no ground truth for Research Methods and Fields.

1.2 Project Architecture

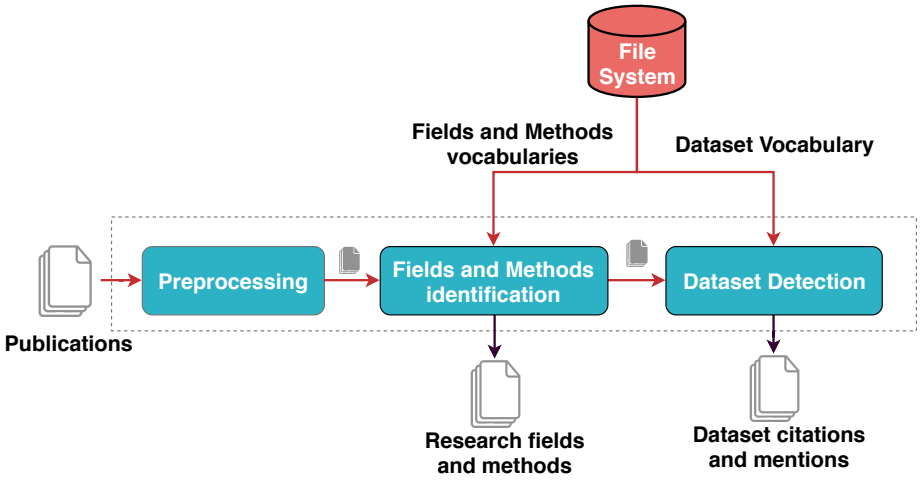

Fig. 1: Data Flow Pipeline (Red lines depict the flow of given and generated files between components whereas black lines represent the generation of final output files)

Our pipeline (shown in Fig. 1) consisted of three main components: 1) Preprocessing, 2) Fields and Methods Identification and 3) Dataset Extraction. The Preprocessing module read the text from publications and generated some additional files (see Section 2 for details). These files along with the given Fields and Methods vocabularies were used to infer Research Fields and Methods from the publications. Then, the information regarding fields was passed onto the Dataset Detection module and using the Dataset Vocabulary, it identified Dataset Citations and Mentions. The following sections provide a detailed overview of each of these components.

[2] https://uk.sagepub.com/en-gb/eur/home

2 Preprocessing

The publications were provided to us in two formats: PDF and text. For Phase-1, we used the given text files, however during Phase-2, we came across many articles in the training files that had not been properly converted to text and contained mostly non-ASCII characters. In order to work with such articles, we relied on the open source tool `pdf2text` from `poppler suite`[3] to extract text from PDFs.

Once we had the text files, we followed the rule-based approach as proposed in [3] for pre-processing. The following series of operations based mostly on regular expressions were performed:

- Words split by hyphens were de-hyphenated
- Irrelevant data was removed (i.e., equations, tables, acknowledgment, references);
- Main sections (i.e., abstract, keywords, JEL-Classification, methodology/data, summary, conclusion) were identified and extracted;
- Noun phrases from these sections were extracted (using the python library, spaCy[4]).

If a section was not found in the article (because of no explicit mention), then only the sections that could be detected were extracted. The remaining content was saved as 'reduced_content' after cleaning to prevent loss of any meaningful data.

Another open source tool that we used was `pdfinfo` from `poppler suite` to extract PDF metadata that very often contained the keywords and subject of an article. This tool was helpful in those cases where the keywords were not found by the regular expression.
In the end, the preprocessing module generated four text files for a publication: PDF-converted text, PDF-metadata, processed articles containing relevant data, and noun phrases from the relevant sections, respectively. These files were then passed on to the other two components of the pipeline, which have been discussed below.

3 Approach

3.1 Research Fields and Methods Identification

Vocabulary Generation and Model Preperation

1. **Research Methods Vocabulary**: In Phase-1 of the challenge, we used the given methods vocabulary. However, in Phase-2, based on the evaluation feedback, we created our own Research Methods Vocabulary using

[3] https://manpages.debian.org/testing/poppler-utils
[4] https://github.com/explosion/spaCy

Wikipedia and DBpedia.[5] We manually curated a list of all the relevant statistical methods from Wikipedia[6] and fetched their descriptions from the corresponding DBpedia resources. For each label in the vocabulary, we extracted noun phrases from its description and added them to the vocabulary.

2. **Research Fields Vocabulary**: For both the phases, we used the given research fields vocabulary and, just like the methods vocabulary, added noun phrases from the description of the labels to it. In addition, we created a blacklist of terms that didn't contain any domain-specific information, such as; Mixed Methods, Meta Analysis, Narrative Analysis and the like for Phase-2.
3. **Word2Vec Model generation**: In this pre-processing step, we used the above-mentioned vocabulary files to generate a vector model for both research fields and methods. The vector model was generated by using the labels and description of the available research fields and methods and then using the noun phrases present in them to form a sum vector. The sum vector was basically the sum of all the vectors of the words present in a particular noun phrase. A pre-trained Word2Vec model [2] was used to extract the vectors of the individual words.
4. **Research Method training results creation**: For research methods, we generated an intermediate result file for the publications present in the training data. It was generated using a `naïve finder algorithm` that, for each publication, selected the research method with the highest cosine similarity to any of its noun phrase's vectors. This file was later used to assign weights to research methods using Inverse Document Frequency.

Processing with Trained Models

- **Finding Research Fields and Methods:** To find the research fields and methods for a given list of publications, we performed the following steps: (At first, Step 1 was executed for all the publications, thereafter Step 2 and 3 were executed iteratively for each publication).
  1. **Naïve Research Method Finder run** - In this step, we executed the `naïve research method finding algorithm` against all the current publications and then merged the results with the existing result from the `research methods' preprocessing step`. The combined result was then used to generate IDF weight values for each `research method`, in order to compute the significance of recurring terms.
  2. **IDF-based Research Method Selection** - We re-ran the algorithm to find the closest research method to each noun phrase and then sorted the pairs based on their weighted cosine similarity. The weights were taken from the IDF values generated in the first step and the manual

[5] https://wiki.dbpedia.org/services-resources/ontology
[6] https://en.wikipedia.org/wiki/Category:Statistical_methods

weights assigned (section-wise weightage). Here, the noun phrases that came from the methodology section and from the methods listed in JEL-classification (if present) were given a higher preference. The pair with the highest weighted cosine similarity was then chosen as the Research Method of the article.

3. **Research Field Finder run** - In this step, we first found the closest research field from each noun phrase in the publication. Then we selected the Top N (= 10) pairs that had the highest weighted cosine similarity. Afterwards, the noun phrases that had a similarity score less than a given threshold (= 0.9) were filtered out. The end-result was then passed on to the post-processing algorithm.
   For weighted cosine similarity, the weights were assigned manually based on the section of publication from which the noun phrases came. In general, noun phrases from title and keywords were given a higher preference than other sections.
4. **Research Field Selection** - The top-ranked term from the result of step 3, which was not present in the blacklist of irrelevant terms, was marked as the research field of the article.

3.2 Dataset Extraction

For identifying the datasets in a publication, we followed two approaches and later combined results from both. Both the approaches have been described below.

1. **Simple Dataset Mention Search:** We chose the dataset citations from the given Dataset Vocabulary that occurred for one dataset only and used these unique mentions to search for the corresponding datasets in the text documents. Then, we computed a frequency distribution of the datasets. As can be seen from Fig. 2, certain dataset mentions occurred more often than others, which increased the number of false positives. Therefore, we filtered out those mentions that occured more than a certain threshold value (=1.20) multiplied by the median of the frequency distribution and passed the remaining mentions to an interim result file.
2. **Rasa-based Dataset Detection:** In our second approach, we trained an entity extraction model based on conditional random fields using Rasa NLU [1]. We particularly tested two configurations for training the CRF-based NER model. In Phase-1, the 2500 labeled publications from the training dataset were used for training the Rasa NLU[7] model. Later in Phase-2, when the Phase-1 holdout corpus was released, we combined its 5000 labeled publications with the previously given 2500 labeled publications and then retrained the model again with these 7500 labeled publications.
   Running the CRF-Model: The trained model was run against the pre-processed data to detect dataset citations and mentions. Only the entities

[7] https://rasa.com/docs/nlu

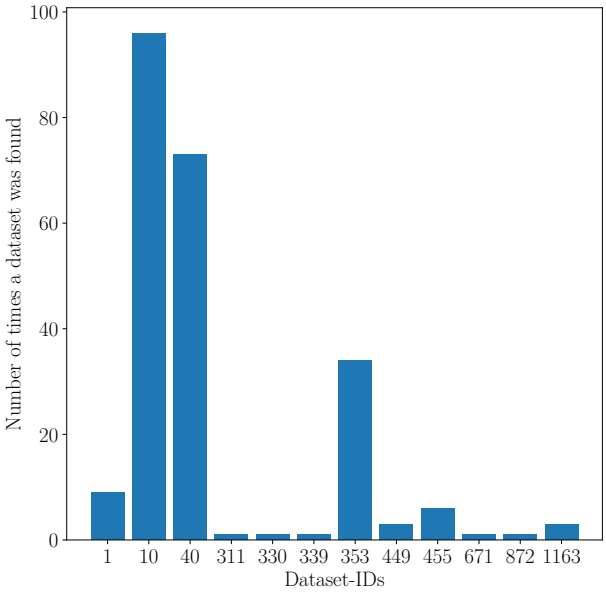

Fig. 2: Frequency Distribution of Dataset Citations

that had a confidence score greater than a certain threshold value (=0.72) were considered as dataset mentions. A dataset mention was considered as a citation only if it was found in the given Dataset Vocabulary and if it belonged to the research field of the article. To check if a dataset belonged to the field of research, we found the cosine similarity of the terms in the 'subjects' field of the Dataset Vocabulary with the keywords, and identified the Research Field of the article.

3. **Combining the two approaches:** The output generated by the rasa-based approach was first checked for irrelevant citations before a union was performed to combine the results. This was done by checking if a given dataset_id occured more than a threshold value (=1.20) multiplied by the median of the frequency distribution (same as the filtering process of the Simple Dataset Mention Search).

Note that, the threshold values mentioned above were set after some experiments of trial and testing. For dataset extraction, the goal was to keep the number of false positives low while not compromising the true positives. For research methods and fields, a manual evaluation (see the next section for details) was done to test if the results made sense with the articles.

4 Evaluation

We performed a quantitative evaluation for Dataset Extraction using the evaluation script provided by the competition organizers. This evaluation (see Table 1) was carried out against the validation data, wherein we compared four different configurations. As can be inferred from the table, there was only a slight increase in performance for the Rasa-based model, when the training samples were increased. However, combining it with the Simple Dataset Mention Search, increased the performance by *19.42%*. Interestingly, there was no improvement in performance in the combined approach even when the training samples for Rasa Model were increased. This might be because of the removal of frequently-occuring terms from the Rasa-generated output, based on the frequency distribution of dataset mentions as computed in the Simple Dataset Mention Search.

Table 1: Quantitative Evaluation of Datasets against Validation Data. (The numbers inside brackets indicate training samples)

	Phase-1	**Phase-2**		
Metrics	**Rasa-based Approach** (2500)	**Rasa-based Approach** (7500)	**Combined Approach** (2500)	**Combined Approach** (7500)
Precision	0.382	0.388	**0.456**	**0.456**
Recall	0.26	0.26	**0.31**	**0.31**
F1	0.309	0.311	**0.369**	**0.369**

For Research Fields and Methods, we carried out a qualitative evaluation against 10 randomly selected articles from Phase-1 holdout corpus. Tables 2 and 3 depict a comparison between the predicted fields and methods in Phase-1 and Phase-2. In general, our models returned a more granular output in the second phase, solely because of the modifications we made in the vocabularies.

Table 2: Evaluation of Research Fields against Phase-1 holdout

pub_id	**Keywords**	**Phase-1**	**Phase-2**
10328	Cycling for transport, leisure and sport cyclists	Health evaluation	**Public health and health promotion**
7270	Older adult drug users, harm reduction	Health Education	**Correctional health care**
6053	Economic conditions - crime relationship, homicide	Homicide	**Gangs and crime**

Table 3: Evaluation of Research Methods against Phase-1 holdout

pub_id	Keywords	Phase-1	Phase-2
10328	Thematic content analysis	Thematic analysis	**Sidak correction**
7270	Interviews conducted face to face, finding systematic patterns or relationships among categories identified by reading the interview transcript	Qualitative interviewing	**Sampling design**
6053	Autoregressive integrated moving average (ARIMA) time-series model	Methodological pluralism	**Multivariate statistics**

5 Discussion

Throughout the course of this competition, we encountered several challenges and limitations in all the three stages of the pipeline. In the preprocessing step, the appropriate extraction of text from PDFs turned out to be rather challenging. This was especially due to the varied formats of the publications, which made the extraction of specific sections—that contained all data relevant to our work—demanding. As mentioned before, if there was no explicit mention of the key-terms like `Abstract, Keywords, Introduction, Methodology/Data, Summary, Conclusion` in the text, then the content was saved as 'reduced_content' after applying all other preprocessing steps and filtering out any irrelevant data.

Our experiments suggest that the labeled publications we received for dataset detection were not uniform in the dataset mentions provided, which made it difficult to train an entity extraction model even with an increased number of training samples. Hence, there was only a slight improvement in performance when the Rasa-model was trained with 7500 publications instead of 2500. This was also why we combined the Rasa-based approach with the Simple Dataset Mention Search, so that at least the datasets that were present in the vocabulary didn't get missed.

Regarding the fields and methods, vocabularies played an immense role in their identification. The vocabularies that were provided by the SAGE publications contained some terms that were either polysemous or very high-level and therefore, were picked up by our model very often. Hence, for research methods, we created our own vocabulary containing all the relevant statistical methods, and for fields, we introduced a blacklist of irrelevant terms and looked it up each time, before writing the result to the output file. Since the focus was on more granulated results, we tried to look for open ontologies for Social Science Fields and Methods and unfortunately, could not find any. It is worth mentioning that since our approach for Fields and Methods identification relied heavily upon vocabularies, it could not find any new methods or fields from the publications.

6 Future Agenda

The data provided to us in the competition displayed a cornucopia of inconsistencies even after human processing. We hence propose that machine-aided methods for computing correct and complete structured representation of publications are of central importance for scientific research. Previous works on never-ending learning have shown how humans and extraction algorithms can work together to achieve high-precision and high-recall knowledge extraction from unstructured sources. In our future work, we hence aim to populate **scientific knowledge graphs** based on never-ending learning. The methodology we plan to develop will be domain-independent and rely on active learning to classify, extract, link and publish scientific research artifacts extracted from open-access papers. The resulting graphs will

- rely on advanced distributed storage for RDF to scale to the large number of publications available;
- be self-feeding, i.e., crawl the web for potentially relevant content and make this content available for processing;
- be self-repairing, i.e., be able to update previous extraction results based on insights gathered from new content;
- be weakly supervised by humans, who would assist in correcting wrong hypotheses;
- provide standardized access via W3C Standards such as SPARQL.

Having such knowledge graphs would make it easier for the researchers (both young and veteran) to easily follow along with their domain of fast-paced research and eliminate the need to manually update the domain-specific ontologies for fields, methods and other metadata as new research innovations come up.

References

1. Tom Bocklisch, Joey Faulkner, Nick Pawlowski, and Alan Nichol. Rasa: Open source language understanding and dialogue management. *CoRR*, abs/1712.05181, 2017.
2. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.
3. David Westergaard, Hans Henrik Stærfeldt, Christian Tønsberg, Lars Juhl Jensen, and Søren Brunak. A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. *PLoS Computational Biology*, 14(2), 2018.