Rich Context Competition Phase 2 (RCC-05)

Wolfgang Otto[1]*, Andrea Zielinski[1], Bahnam Ghavimi[1], Dimitar Dimitrov[1], Narges Tavakolpoursaleh[1]

Abstract
This document describes the approach, results, and software submitted by team RCC-05 to the Rich Context Competition (RCC). The team consists of members of the department *Knowledge Technologies in the Social Sciences* of *GESIS - Leibniz Institute for the Social Sciences* The goal of the RCC is to automatically discover and link research datasets, methods, and fields in social science publications.

Keywords
Dataset Mention Extraction — Dataset Linking — Research Method Extraction — Research Field Classification

[1] *Knowledge Technologies in the Social Sciences, GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany*
***Corresponding author**: wolfgang.otto@gesis.org

Contents

1. Introduction

Scientists and analysts often face the problem of finding interesting research datasets and identifying who else used the data, in which research fields, and how the data has been analyzed from a methodological perspective. To address these problems, the Coleridge Initiative organized the Rich Context Competition[1](RCC). The competition invited international research teams to develop text analysis and machine learning tools that can discover relationships between research datasets, methods, and fields in scientific literature. The competition took place between October 2018 and February 2019 and included two phases[2]. The first phase was open for all teams which have submitted a letter of intent. Teams are then provided with a corpus of social science publications to develop and train machine learning algorithms for automatic research dataset, methods and field detection and linking. More concretely, one major subtask consisted of linking dataset mentions to a given set of around 10,000 dataset descriptions from the ICPSR's research data index.[3] Only the best four teams from the first phase are invited to the second phase of the competition and asked to discover research datasets, methods, and fields in a larger corpus of social science publications. All submitted algorithms have to be made publicly available as open source tools. With this document, we (team RCC-5) aim to fulfill another requirement, i.e., the documentation and summary of the developed approach including data pre-processing, algorithms, and software.

1.1 General Approach and Software Components

One of the central tasks in the RCC is the extraction of dataset mentions from text. Nevertheless, we considered the methods and fields discovery equally important. To this end, we decided to follow a module-based approach and developed tools that can be used separately but also as parts of a data processing pipeline. Figure 1 shows an overview of the software modules developed for the RCC competition, including their dependencies. Here, the upper three modules (gray) describe the pre-processing steps (cf. Section 2.3). The lower four modules (blue) are used to generate the output in a pre-specified format. The pre-processing step consists of extracting metadata and pure text from PDF documents. The extraction itself is done using the Cermine Tool[4] which returns a Journal Article Tag Suite[5](Jats) XML document. Then, in a second step, text, metadata and references are extracted. The output of the pre-processing is then used by the software modules responsible for tackling the individual sub-tasks, i.e., discovering research datasets (cf. Section 3), methods (cf. Section 4) and fields (cf. Section 5). Section 6 provides the technical details of the modules, i.e., input, output, and how to run the modules.

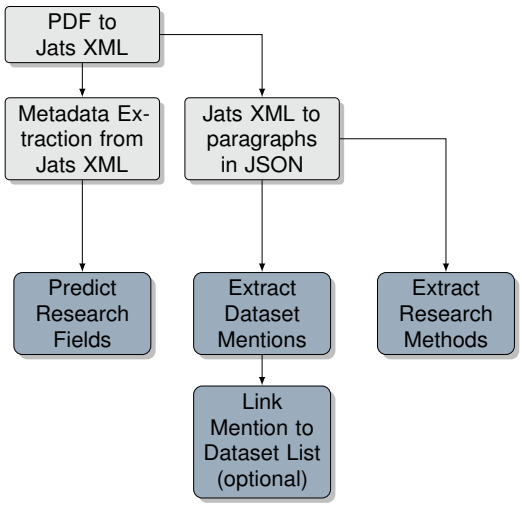

Figure 1. Software modules. The figure shows an overview of the individual software modules described in this document and their dependencies. Modules colored in gray represent our pre-processing pipeline, whereas blue-colored modules represent the three main tasks of the RCC.

1.2 First Phase Feedback

After the first phase, each team received feedback from the organizers of the RCC. The feedback is twofold and consists of a quantitative and qualitative evaluation. Unfortunately, our team did not perform very well regarding precision and recall. In contrast to this, our approach has been found convincing regarding the quality of results. The qualitative feedback result from a random sample of ten documents that are given to four judges. Judges are then asked to manually extract dataset mentions and calculate the overlap between their dataset extractions and the output of our algorithm. Other factors that judges took into consideration are specificity, uniqueness and multiple occurrences of dataset mentions. As for the extraction of research methods and fields no ground truth has been provided, these tasks were evaluated against the judges' expert knowledge. Similarly to the extraction of dataset mentions, specificity and uniqueness have been considered for these two tasks. The feedback our team received acknowledged the fact that no ground truth has been provided and our efforts regarding the extraction of research methods and fields.

2. Data and Pre-processing

This section describes the data provided by the organizers of the RCC, the external data sources we used as well as our pre-processing steps.

[1] https://coleridgeinitiative.org/richcontextcompetition
[2] https://coleridgeinitiative.org/richcontextcompetition#competitionschedule
[3] https://www.icpsr.umich.edu/index.html
[4] https://github.com/CeON/CERMINE
[5] https://jats.nlm.nih.gov

2.1 The RCC Corpus

For the first phase, the data provided by the organizers consisted of 5,000 publications. Additionally, a development fold of 100 plain text publications, their metadata, a list of datasets of interest (including all datasets that were explicitly referenced in the curated corpus) were given. The list of datasets should not be considered complete as there could be additional datasets mentioned in these publications. The organizers also provided examples of social science research methods and fields vocabularies in term of SAGE Publications research field and method vocabularies. In the second phase of the competition, an additional set of 5,000 publications from the social sciences has been provided.

2.2 External Data Sources

For developing our algorithms, we also utilized two external data sources. For the discovery of research methods and fields, we resort to data from Social Science Open Access Repository[6] (SSOAR). SSOAR is maintained at GESIS – Leibniz Institute for the Social Sciences collects and archives literature of relevance to the social sciences. In SSOAR, full texts are indexed using controlled social science vocabulary (Thesaurus[7], Classification[8]) and are assigned rich metadata. SSOAR offers documents in various languages. The corpus of English language publications that can be used for purposes of the competition consists of a total of 13,175 documents. All SSOAR documents can be accessed through the OAI-PMH[9] interface. Another external source that we used for discovery of research methods is the ACL Anthology Reference Corpus [1]. ACL ARC is a corpus of scholarly publications about computational linguistics. The corpus consists of a total of 22,878 articles.

2.3 Pre-processing

Although the organizers of the RCC, offered plain texts for the publication, we decided to build our own pre-process pipeline. The pipeline uses the Cermine Tool to extract information from PDF documents. The main benefit of using this tool is the structured metadata output including better disambiguation of sections and paragraphs in the publications. The output XML file uses the Journal Article Tag Suite[10]. For the competition, there are only two interesting elements of the Jats XML format, i.e., <front> and <body>. The <front> element contains the metadata of the publication, whereas the <body> contains the publication text. Another advantage of Cermine is that the hyphenation and segmentation of paragraphs are carried out automatically. As a last step of the pre-processing, we remove all linebreaks from the publication text and output a list of metadata fields and values as shown in Table 1 for each publication paragraph.

Table 1. An Example output of our pre-processing for a paragraph in a given publication.

	Example Text Field Data
publication_id	12744
label	paragraph_text
text	A careful reading of text, word for word, was ...
section_title	Data Analysis
annotations	[{'start': 270, 'end': 295, 'type': 'bibref', ...
section_nr	[3, 2]
text_field_nr	31
para_in_section	1

3. Dataset Extraction

3.1 Task Description

In scientific literature, datasets are specified to indicate, e.g., the data on which a analysis is performed, a certain finding or a claim is based on. In this competition, we focus on (i) extracting and (ii) linking datasets mention from social science publications to a list of given dataset references. Identifying dataset mention in literature is a challenging problem due to the lack of an established style of citing datasets. Furthermore, in many research publication, a correct citation of datasets is entirely missing [2]. The following two sentences exemplify the problem.

Example 1: *P-values are reported for the one-tail paired t-test on* Allbus *(dataset mention) and* ISSP *(dataset mention).*
Example 2: *We used* WHO data from 2001 *(dataset mention) to estimate the spreading degree of AIDS in Uganda.*
We treat the problem of detecting dataset mentions in full-text as a Named Entity Recognition (NER) task.

Formal problem definition Let D denote a set of existing datasets d and the knowledgebase K as a set of known dataset references k. Furthermore, each element of K is referencing an existing dataset d. The Named Entity Recognition and linking task is defined as (i) the identification of dataset mentions m in a sentence, where m references a dataset d and (ii) linking them, when possible, to one element in K (i.e., the reference dataset list given by the RCC).

3.2 Challenges

With our method, we focus on the extraction of dataset mentions in the body of the full-text of scientific publications. We recognize three types a dataset can be mentioned: (i) The full name of a dataset like "National Health and Nutrition Examination Survey", (ii) an abbreviation ("NHaNES") or (iii) a vague reference, e.g., "the monthly statistic". By each

[6] https://www.gesis.org/ssoar/home
[7] https://www.gesis.org/en/services/research/tools/thesaurus-for-the-social-sciences
[8] https://www.gesis.org/angebot/recherchieren/tools-zur-recherche/klassifikation-sozialwissenschaften (in German)
[9] http://www.openarchives.org
[10] https://jats.nlm.nih.gov

of these varieties, the NER task faces particular challenges. For the first type, the used dataset name can vary in different publications. Where one publication cites the dataset with "National Health and Nutrition Examination Survey" the other could use the words "Health and Nutrition Survey". In a case where abbreviations are used a disambiguation problem occurs, e.g., in "WHO data". WHO may describe the World Health Organization or the White House Office. The biggest challenge is again the lack of a precise gold standard that can be used to train a classifier. In the following we describe how we have dealt with this lack of ground truth data.

3.3 Phase one approach

The challenge of missing ground truth data is the main problem to handle during this competition. To this end, supervised learning methods for dataset mentions extraction from text are not directly applicable. To overcome this limitation, we resort to the provided list of dataset mentions and publication pairs and re-annotate the particular sentences in the publication text. This re-annotation is then used to train Spacy's neural network based NER model[11]. We created a holdout set of 1000 publications and a training set of size 4000. We train our model using publication paragraphs as training samples. In the training set, 0.45 percent of the paragraphs contained mentions. For each positive training example, we added a negative example that does not contain dataset mentions and is sampled at random. We used a batch size of 25 and a dropout rate of 0.4. The model was trained for 300 iterations.

Evaluation We evaluated our model with respect to four metrics: strict precision and recall, and partial precision and recall. While the former are standard evaluation metrics, the latter are their relaxed variants in which the degree to which dataset mentions have to match can vary. Consider the following example of a partial match: "National Health and Nutrition Examination Survey" is the extracted dataset mention whereas, "National Health and Nutrition Examination Survey (NHANES)" represents the true dataset mention.

Table 2 show the results of the dataset mention extraction on the holdout set. The model is able to achieve high strict precision and recall values. As expected, the results are even better for the partial version of the metrics. But, this version indicates that even if we are not able to exactly match the dataset mention in text, we can find the right context with very high precision at least.

[11] spacy.io

Table 2. Results(phase one).

Metric	Value
Partial Precision	0.93
Partial Recall	0.95
Strict Precision	0.80
Strict Recall	0.81

3.4 Phase two approach

In the second phase of the competition additional 5,000 publications have been provided. We extended our approach to consider the list with dataset names supplied by the organizers and re-annotated the complete corpus of 15.000 publication in the same manner as in phase one to obtain training data. This time we split the data in 80% for training and 20% for test.

Evaluation We resort to the same evaluation metrics as in phase one. However, we calculate precision and recall on the full-text of the publication and not on the paragraphs as in the first phase. Table 3 show the results achieved by our model. We observe a lower precision and recall values. Compared to phase one, there is also a smaller difference between the precision an recall values for the strict and partial version of the metrics.

Table 3. Results(phase two).

Metric	Value
Partial Precision	0.51
Partial Recall	0.90
Strict Precision	0.49
Strict Recall	0.87

4. Research Method Extraction

4.1 Task Description

Inspired by a recent work of Nasar et al. [3], we define a list of basic entity types that give key-insights into scholarly publications. We adapted the list of semantic entity types to the domain of the social sciences with a focus on *research methods*, but also including related entity types such as *Theory, Model, Measurement, Tool, Performance*. We suspect that the division into semantic types might be helpful to find *research methods*, because related semantic entities types might provide clues or might be directly related to the research method itself. For instance, in order to realize a certain research objective, an experiment is instrumented where a specific combination of *methods* is applied to a *data set* that might be intellectual or *software*, thus achieving a specific *performance* and result in that context.
Example: *P-values* (measurement) are reported for the *one-tail paired t-test* (method) on *Allbus* (dataset) and *ISSP* (dataset).

Formal problem definition Let E denote a set of entities. The Named Entity Recognition and Linking task consists of (i) identifying entity mentions m in a sentence and, (ii) linking them, when possible, to a reference knowledge base

K (i.e, the SAGE Thesaurus[12]) and (iii) assigning a type to the entity, e.g., *research method*, selected from a set of given types. Given a textual named entity mention m along with the unstructured text in which it appears, the goal is to produce a mapping from the mention m to its referent real world entity e in K.

4.2 Challenges

There are some major challenges that any named entity recognition, classification and linking system needs to handle. First, regarding NER, identifying the entities boundary is important, thus detecting the exact sequence span. Second, ambiguity errors might arise in classification. For instance,'range' might be a domain-specific term from the knowledge base or belong to the general domain vocabulary. This is a challenging task for which context information is required. In the literature, this relates to the problem of **domain adaptation** which includes fine-tuning to specific named entity classes[13]. With respect to entity linking, another challenge is detecting name variations, since entities can be referred to in many different ways. Semantically similar words, synonyms or related words, which might be lexically or syntactically different, are often not listed in the knowledge base (e.g., the lack of certain terms like 'questioning' but not 'questionnaire'). This problem of automatically detecting these relationships is generally known as **linking problem**. Note that part of this problem also results from PDF-to-text conversion which is error-prone. Dealing with incomplete knowledge bases, i.e. **handling of out of vocabulary (OOV) items**, is also a major issue, since knowledge bases are often not exhaustive enough and do not cover specific terms or novel concepts from recent research. Last but not least, the combination of different semantic types gives a more coherent picture of a research article. We hypothesize that such information would be helpful and results in an insightful co-occurrence statistics, and provides additional detail directly related to entity resolution, and finally helps to assess the **relevance of terms** by means of a score.

4.3 Our Approach - Overview

Our context-aware framework builds on Stanford's CoreNLP and Named Entity Recognition System[14]. The information extraction process follows the workflow depicted in Figure 1, using separate modules for pre-processing, classification, linking and term filtering.

We envision the task of finding entities in scientific publications as a sequence labeling problem, where each input word is classified as being of a dedicated semantic type or not. In order to handle entities related to our domain, we train a novel machine learning classifier with major semantic classes, using training material from the ACL RD-TEC 2.0 dataset

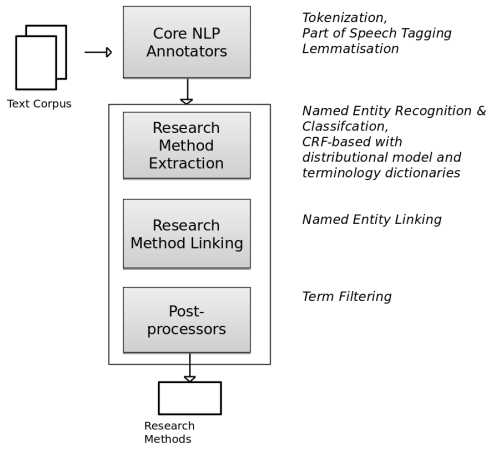

Figure 2. Overview of the entity extraction pipeline

[5]. Apart from this, we follow a domain adaptation approach inspired by [6] and ingest semantic background knowledge extracted from external scientific corpora, in particular the ACL Anthology [1, 7]. We perform entity linking by means of a new gazetteer-based SAGE dictionary of Social Research Methods [8], thus putting a special emphasis on the social sciences. The linking component addresses the synonymy problem and matches an entity despite name variations such as spelling variations. Finally, term filtering is carried out based on a termhood and unithood, while scoring is achieved by calculating a relevance score based on TF-IDF (cf. Section 4.3).

Our research experiments are based on the repository for the Social Sciences SSOAR as well as the train and test data of the Rich Context Competition corpus[15]. Our work extends previous work on this topic (cf. [9]) in various ways: First, we do not limit our study to abstracts, but use the entire fulltext. Second, we focus on a broader range of semantic classes, i.e. *Research Method*, *Research Theory*, *Research Tool* and *Research Measurement*, tackling also the problem of identifying novel entities.

Distributed Semantic Models For domain adaptation, we integrate further background knowledge. We use vector embeddings of words trained on additional corpora and which serve as input features to the CRF model. Semantic representations of words are a successful extension of common features, resulting in higher NER performance [10] and can be trained offline.

In this work, the word vectors were learned from the scientific ACL ARC[16] using Gensim with the skip gram model (cf. [11]) and a pre-clustering algorithm[17]. A summary of

[12] http://methods.sagepub.com

[13] apart from those used in traditional NER systems like *Person*, *Location*, or *Organization* with abundant training data, as covered in the Stanford NER system[4]

[14] https://nlp.stanford.edu/projects/project-ner.shtml

[15] https://coleridgeinitiative.org/richcontextcompetition with a total of 5,000 English documents

[16] https://acl-arc.comp.nus.edu.sg/

[17] Word embeddings are trained with a skip gram model using embedding

Table 4. English data used for Training Word Embeddings

Corpus	Articles	Documents/Tokens
ACL Corpus	22,878	806,791/2.5 GB

Table 5. Features used for NER

Type	Features
Token unigrams	w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2}, ...
POS unigrams	p_i, p_{i-1}, p_{i-2}
Shapes	shape and capitalization
NE-Tag	t_{i-1}, t_{i-2}
WordPair	(p_i, w_i, c_i)
WordTag	(w_i, c_i)
Gazetteer	SAGE gazetteer
Distributional Model	ACL Anthology model

the size of the unlabeled English data used for training word embeddings can be found in Table 4.

Features The features incorporated into the linear chain CRF are shown in the Table 5. The features depend mainly on the observations and on pairs of adjacent labels, using a log-linear combination. However, since simple token level training of CRFs leads to poor performance, more effective text features such as word shape, orthographic, gazetteer, Part-Of-Speech (POS) tags, along with word clustering (see Section 4.3) have been used.

Knowledge Resources We use the SAGE thesaurus which includes well-defined concepts, an explicit taxonomic hierarchy between concepts as well as labels that specify synonyms of the same concept. A portion of terms is unique to the social science domain (e. g., 'dependent interviewing'), while others are drawn from related disciplines such as statistics (e. g., 'conditional likelihood ratio test')[18]. However, since the thesaurus is not exhaustive and covers only the top-level concepts related to social science methods, our aim was to extend it by automatically extracting further terms from domain-specific texts, in particular from the Social Science Open Access Repository. More concretely, we carried out the following steps to extend SAGE as an off-line step. For step 2 and 3, candidate terms have been extracted by our pipeline for the entire SSOAR corpus.

1. Assignment of semantic types to concepts (manual)
2. Extracting terms variants such as abbreviations, synonyms, related terms from SSOAR (semi-automatic)
3. Computation of Term and Document Frequency Scores for SSOAR (automatic)

size equal to 100, word window equal to 5, minimal occurrences of a word to be considered 10. Word embeddings are clustered using agglomerative clustering with a number of clusters set to 500,600,700 Ward linkage with euclidean distance is used to minimize the variance within the clusters.

[18] A glossary of statistical terms as provided in `https://www.statistics.com/resources/glossary/` has been added as well.

Table 6. Most relevant terms from SAGE by Semantic Type

SAGE Term	TF-IDF Score	Semantic Class
Fuzzy logic	591,29	Research Method
arts-based research	547,21	Research Method
cognitive interviewing	521,13	Research Method
QCA	463,13	Research Method
oral history	399,68	Research Method
market research	345,37	Research Field
life events	186,61	Research Field
Realism	314,34	Research Theory
Marxism	206,77	Research Theory
ATLAS.ti	544,51	Research Tool
GIS	486,01	Research Tool
SPSS	136,52	Research Tool

Extracting term variants such as abbreviations, synonyms, and related terms 26.082 candidate terms have been recognized and classified by our pipeline and manually inspected to a) find synonyms and related words that could be linked to SAGE, and b) build a post-filter for incorrectly classified terms. Moreover, abbreviations have been extracted using the algorithm of Schwartz and Hearst [12]. This way, a Named Entity gazetteer could be built and will be used at run-time. It comprises 1,111 terms from SAGE and 447 terms from the Statistics glossary as well as 54 previously unseen terms detected by the model-based classifier.

Computation of Term and Document Frequency Scores Term frequency statistics have been calculated off-line for the entire SSOAR corpus. The term frequency at corpus level will be used at run time to determine the term relevance at the document level by calculating the TF-IDF scores. The most relevant terms from SAGE are listed in Table 6.

Definition of a Relevance Score Relevance of terminology is often assessed using the notion of *unithood*, i.e. 'the degree of strength or stability of syntagmatic combinations of collections', and *termhood*, i.e. 'the degree that a linguistic unit is related to domain-specific concepts' [13]. Regarding *unithood*, the NER model implicitly contains heuristics about legal POS tag sequences for candidate terms, consisting of at least one noun (NN), preceeded or followed by modifiers such as adjectives (JJ), participles (VB*) or cardinal numbers (CD), complemented by wordshape features.

In order to find out if the candidate term also fulfills the *termhood* requirement, domain-specific term frequency statistics have been computed on the SSOAR repository, and set in contrast to general domain vocabulary terms. It has to be noted that only a small portion of the social science terms is actually unique to the domain (e.g., 'dependent interviewing'),

while others might be drawn from related disciplines such as statistics (e.g., 'conditional likelihood ratio test').

Preliminary Results Our method has been tested on 100 fulltext papers from SSOAR and 10 documents from the Rich Context Competition (RCC), all randomly selected from hold out corpora. In our experiments on SSOAR Social Science publications, we compared results to the given metadata information. The main finding was that while most entities from the SAGE thesaurus could be extracted and linked reliably (e.g., 'Paired t-test'), they could not be easily mapped to the SSOAR metadata terms, which consist of only a few abstract classes (e.g., 'quantitative analysis'). Furthermore, our tool was tested by the RCC organizer, were the judges reviewed 10 random publications and generated qualitative scores for each document.

4.4 Conclusion and Future Work

We plan to carry out a more detailed evaluation on fulltext scholarly publications and assess the impact of different features used in the ML model, including background resources such as embeddings and dictionaries.

5. Research Field Classification

5.1 Task Description

The goal of this task is to identify the research fields covered in social science publications. The RCC data does not provide a gold standard —annotated training data— for that task. To this end, we decided to train a classifier using annotated data from SSOAR. In this way, our interpretation of the task is to select one or more labels from a given set of labels for each publication. This approach is known as a mulit-label classification. In our case, a label represents a research field.

5.2 Our approach - Overview

Due to the unequal distribution of labels in the dataset, we need to guaranty enough training data for each label. We selected only labels with frequency over 300 for training the model which results in a total of 44 labels representing research fields. We decided to train a classification model based on the fasttext framework [14]. To train our model we resort to the abstracts of the publication, as this approach worked better than using the full-texts.

5.3 Evaluation

Figure 3 shows the performance of the model regarding various evaluation metrics for different thresholds. A label is assigned to a publication if the model outputs a probability for the label above the defined threshold. In multi-label classification, this allows us to evaluate our model from different perspectives.

The intersection point of micro-recall and micro-precision is at threshold point around 0.1. In this point, the model has some predictions for almost all publications, but about 0.8 publication have only wrongly predicted labels (cf. red line).

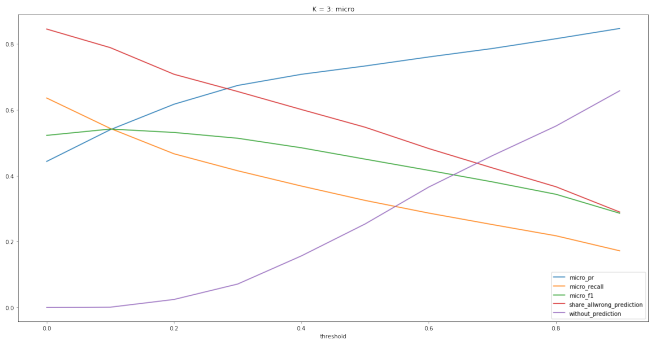

Figure 3. Precision-Recall vs. Threshold

For the default threshold (0.5), precision is more than 0.7 while the number of publications without prediction is about 0.3 (cf. purple line). Since inclination of micro precision is lower than the line of the number of publications without prediction, picking a threshold like 0.4 looks better than default one.

6. Technical Documentation

The project contains the following modules listed in the order in which they are excecuted.

6.1 Pre-processing

6.1.1 PDF Text Extraction

Module name :

Cermine_NlmJat_extractor

Function:

Converts each PDF files of a given folder to JATS XML Format. Each input PDF File is transformed to one XML File.

Bash function call:

```
java -jar target/cermineXMLextraction
    -1.0.0-jar-with-dependencies.jar
```

Parameter (2):

```
-s <source folder>\
-t <target folder>
```

Returns:

XML Files in JATS XML Format.

Build:

This is a Java program using Maven build tool.

build call:

mvn install

6.1.2 Extraction of text from JATS XML

Module name :

preprocess-rcc-data

Function:

Transform text from JATX XML Format into a JSON File containing a list of textfields with essential metadata for each JATS XML file of a given folder.

Bash function call:

```
‘python3 ./jats_text_extractor.py ‘
```

Parameter (4):

```
<source folder>
<target folder>
<limiting number of files to transform
    (-1: all)>
<number of cores to use for
    multiprocessing (-1: all)>
```

Returns:
A JSON File for each given XML File in the source folder

6.1.3 Metadata Extraction

Module name :
preprocess-rcc-data
Function:
Extracts structured metadata and references from all JATS XML files in a given folder into two Files. One containing the metadata from all Publications in JATS XML files and one containing all references from the JATS XML Files. The target file format is JSON.
Bash function call:

```
python3 ./jats_metadata_extractor.py
```

Parameter (3):

```
<source folder>
<target filename for metadata>
<target filename for references>
```

Returns:
Two JSON files containing metadata and references from all XML Files

6.2 Dataset Mentions

6.2.1 Dataset Mention Extraction

Module:
dataset-mention-extraction
Function:
Extract dataset mentions from all JSON Files from a given folder with a given spacy model.
Bash function call:

```
python3 ./predict_mentions.py
```

Parameter (4):

```
<source folder>
<name of spacy model folder>
<target filename rcc-output>
<target filename internal format>
```

Returns:
Two JSON files containing the found dataset mentions in all given JSON Files. One in RCC defined output. One in Internal format including the senctence the dataset mention occures.
Train:
For training we submit a jupyter notebook with all needed code. *Train_spacy_ner_prod.ipynb*
Build (For training only):
Install english spacy language model. This can be done with ‘python -m spacy download en‘.

6.2.2 Dataset Linking (only Phase 1)

Module:
dataset-prediction
Function:
Links dataset mentions given a JSON file in internal format to datasets listed in a given JSON File.
Bash function call:

```
python3 ./retrieve.py
```

Parameter (3):

```
<JSON filename of extracted mentions>
<JSON filename of dataset list to match>
<output filename for dataset citations>
```

Returns:
JSON file in the format defined by the competition containing information about links between publications and datasets.

6.3 Research Method Extraction

Module name :
research-method-extractor
Function:
Extracts research method terms from JSON files with text information from publications.
Bash function call:

```
java -jar target/gesisents-0.1-jar-with-
    dependencies.jar
```

Parameter (3):

```
<source folder>
<target file name>
<Limit to reduce the number of processed
    files (-1: all)>
```

Returns:
A JSON file in the format defined by the competition containing information about publications and research methods.
Build:
This is a Java program using Maven build tool.
build call:
mvn install

6.4 Research Field Classifier

Module name :
research-field-detector
Function:
Classifies given abstracts with classoz Labels
Bash function call:

```
python3 ./fasttext_predictor.py
```

Parameter (4):

```
<filename of JSON file with abstracts>
<filename of fasttext model>
<filename of label dictionary in JSON>
<target filename labels in >
```

Returns:

A JSON file in the format defined by the competition containing information about publications and research fields.

Acknowledgments

We would like to thank GESIS for giving us the time and resources to participate in the competition.

References

[1] Steven Bird, Robert Dale, Bonnie J Dorr, Bryan Gibson, Mark Thomas Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir R Radev, and Yee Fan Tan. The acl anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*. European Language Resources Association (ELRA), 2008.

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

[3] Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik. Information extraction from scientific articles: a survey. *Scientometrics*, 117(3):1931–1990, 2018.

[4] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 363–370. Association for Computational Linguistics, 2005.

[5] Behrang QasemiZadeh and Anne-Kathrin Schumann. The acl rd-tec 2.0: A language resource for evaluating term extraction and entity recognition methods. In *LREC*, 2016.

[6] Rodrigo Agerri and German Rigau. Robust multilingual named entity recognition with shallow semi-supervised features. *Artificial Intelligence*, 238:63–82, 2016.

[7] Daniel Gildea, Min-Yen Kan, Nitin Madnani, Christoph Teichmann, and Martin Villalba. The acl anthology: Current state and future directions. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 23–28, 2018.

[8] Michael Lewis-Beck, Alan E Bryman, and Tim Futing Liao. *The Sage encyclopedia of social science research methods*. Sage Publications, 2003.

[9] Judith Eckle-Kohler, Tri-Duc Nghiem, and Iryna Gurevych. Automatically assigning research methods to journal articles in the domain of social sciences. In *Proceedings of the 76th ASIS&T Annual Meeting: Beyond the Cloud: Rethinking Information Boundaries*, page 44. American Society for Information Science, 2013.

[10] Joseph Turian, Lev Ratinov, and Yoshua Bengio. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 384–394, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[11] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[12] Ariel S. Schwartz and Marti A. Hearst. A simple algorithm for identifying abbreviation definitions in biomedical text. In *Pacific Symposium on Biocomputing*, pages 451–462, 2003.

[13] Kyo Kageura and Bin Umino. Methods of automatic term recognition: A review. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 3(2):259–289, 1996.

[14] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics, April 2017.

i