8

KNOWLEDGE EXTRACTION FROM SCHOLARLY PUBLICATIONS: THE GESIS CONTRIBUTION TO THE RICH CONTEXT COMPETITION

Wolfgang Otto, Andrea Zielinski, Behnam Ghavimi, Dimitar Dimitrov, Narges Tavakolpoursaleh, Karam Abdulahhad, Katarina Boland and Stefan Dietze

Introduction

GESIS – Leibniz Institute for the Social Sciences (GESIS)[1] is the largest European research and infrastructure provider for the social sciences and offers research data, services and infrastructures supporting all stages of the scientific process. The Knowledge Technologies for the Social Sciences (WTS)[2] department is responsible for developing all digital services and research data infrastructures at GESIS and aims to provide integrated access to social science data and services. Next to traditional social science research data, such as surveys and census data, an emerging focus is to build data infrastructures able to exploit novel forms of social science research data, such as large Web crawls and archives.

Research at WTS[3] addresses areas such as information retrieval, information extraction and natural language processing (NLP), semantic technologies and human–computer interaction and aims to ensure access and use of social science research data in accordance with the FAIR principles, for instance, through interlinking of research data, established vocabularies and knowledge graphs and by facilitating semantic search across distinct platforms and datasets. Due to the increasing importance of Web and W3C standards as well as Web-based research data platforms, in addition to traditional research data portals, findability and interoperability of research data across the Web constitutes one current challenge. In the context of Web-scale reuse of social science resources, the extraction of structured data about scholarly entities such as datasets and methods from unstructured and semi-structured text, as found in scientific publications or resource metadata, is crucial in order to be able to uniquely identify social science resources and to understand their inherent relations.

Previous work at WTS/GESIS addressing such challenges applies NLP and machine learning techniques to, for instance, extract and disambiguate mentions of datasets[4] (Boland et al., 2012; Ghavimi et al., 2016)), authors (Backes, 2018a, 2018b) or software tools (Boland and Krüger, 2019) from scientific publications or to extract and fuse scholarly data from large-scale Web crawls (Sahoo et al., 2017; Yu et al., 2019). Resulting pipelines and data are used to empower scholarly search engines such as *GESIS-wide search*[5] (Hienert et al., 2019) which provides federated search for scholarly resources (datasets, publications, etc.) across a range of GESIS information systems, or the *GESIS DataSearch* platform[6] (Krämer et al., 2018), which enables search across a vast number of social science research datasets mined from the Web.

Given the strong overlap of our research and development profile with the recent initiatives of the Coleridge Initiative to evolve this research field through the Rich Context Competition (RCC),[7] we are enthusiastic about having participated in the competition and are looking forward to continuing this collaboration towards providing sound frameworks and tools which automate the process of interlinking and retrieving scientific resources.

The central tasks in the RCC are the extraction and disambiguation of mentions of datasets and research methods as well as the classification of scholarly articles into a discrete set of research fields. After the first phase, each team received feedback from the organizers of the RCC consisting of a quantitative and qualitative evaluation. Whereas

the quantitative results of our initial contribution throughout the first phase showed significant room for improvement, the qualitative assessment, conducted by four judges on a sample of 10 documents, underlined the potential of our approach.

This chapter describes our approaches, techniques, and additional data used to address all three competition tasks. As described below, we decided to follow a module-based approach where each module or the entire pipeline can be reused. The rest of the chapter is organised as follows. We begin by providing an overview of our approach, background data and preprocessing steps, and then describe our approaches in more detail, including results towards each of the tasks. Finally, we discuss our results and provide an overview of future work.

Approach, Data and Preprocessing

This section describes the external data sources we used as well as our preprocessing steps.

Approach Overview and Initial Evaluation Feedback

The central tasks in the RCC are the extraction of dataset mentions from text. Even so, we considered the discovery of research methods and research fields important. To this end, we decided to follow a module-based approach. Users could choose to use each specific module alone or as part of a data-processing pipeline. Figure 8.1 depicts the modules developed and their dependencies. Here, the upper three modules (in grey) describe the preprocessing steps (see Chapter 2). The lower four modules (in red) are used to generate the output in a predefined format as specified by the competition.

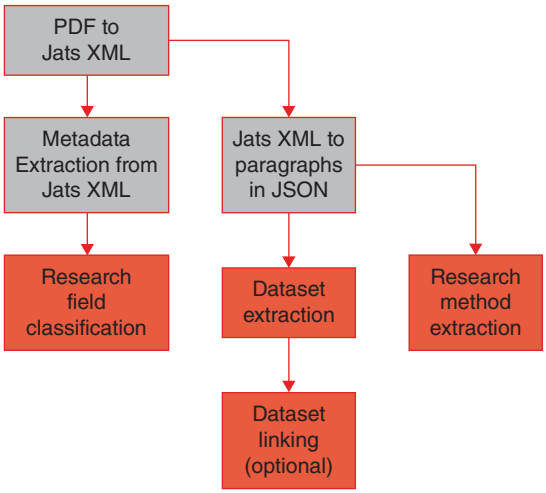

Figure 8.1 An overview of the individual software modules described in this chapter and their dependencies: our preprocessing pipeline (grey); the three main tasks of the RCC (red)

The preprocessing step consists of extracting metadata and raw text from PDF documents. The output of this step is then used by the software modules responsible for tackling the individual sub-tasks. These sub-tasks are to discover research datasets, methods and fields (see below). First, a named entity recognition (NER) module is used to find dataset mentions. This module uses a supervised approach trained on a weakly labelled corpus. In the next step, we combine all recognized mentions for each publication and compare these mentions to the metadata from the list of datasets given by the competition. For this linking step the mentions and year information located in the same sentence are used. The corresponding sentence and extracted information are saved for debugging and potential usage in future pipeline components. The task of identifying research methods is solved using a named entity recognition and linking module with incorporated word embeddings and lexical resources. To identify research fields, we trained a classifier on openly available abstracts and metadata from the domain of social sciences crawled from the Social Science Open Access Repository[8] (SSOAR). We tried different classifiers and selected the best-performing one, a classifier based on fastText,[9] that is, a neural-net-based approach with high performance (Joulin et al., 2017).

After the first phase, each team received feedback from the organizers of the RCC. The feedback was twofold, a quantitative and qualitative evaluation. Unfortunately, the quantitative assessment showed our algorithm for dataset mention retrieval did not perform well regarding precision and recall metrics. However, our approach was found convincing regarding the quality of results. The qualitative feedback was based on a random sample of 10 documents given to four judges. The judges were asked to manually extract dataset mentions. Then the overlap between their dataset extractions and the output of our algorithm was calculated. Other factors that judges took into consideration were specificity, uniqueness, and multiple occurrences of dataset mentions. As for the extraction of research methods and fields, no ground truth was provided; these tasks were evaluated against the judges' expert knowledge. Similarly, to the extraction of dataset mentions, specificity and uniqueness were considered for these two tasks. The feedback our team received was overall positive.

External Data Sources

To develop our algorithms, we utilized two external data sources. For the discovery of research methods and fields, we resorted to data from SSOAR.[10] GESIS – Leibniz Institute for the Social Sciences maintains SSOAR by collecting and archiving literature of relevance to the social sciences.

In SSOAR, full texts are indexed using controlled social science vocabulary (Thesaurus,[11] Classification[12]) 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[13] interface.

Another external source we have used to discover research methods is the ACL Anthology Reference Corpus (Bird et al., 2008). ACL ARC is a corpus of scholarly publications about computational linguistics. The corpus consists of a total of 22,878 articles.

Preprocessing

Although the organizers of the RCC offered plain texts for each publication, we decided to build our own preprocessing pipeline. The extraction of text from PDF files is still an error-prone process. To handle de-hyphenation and paragraph segmentation during extraction time and benefit from automatic metadata extraction (i.e. title, author, abstracts and references) we decided to use a third-party extraction tool. Cermine[14] (Tkaczyk et al., 2015) transforms the files into XML documents using the Journal Article Tag Suite (Jats).[15] For the competition we identified two interesting elements of the Jats XML format, namely, <front> and <body>. The <front> element contains the metadata of the publication, whereas the <body> contains the main textual and graphic content of the publication. As a last step of the preprocessing, we removed all linebreaks from the publication. The output of this step is a list of metadata fields and values, as shown in Table 8.1 for each publication paragraph.

Table 8.1 Example preprocessing output 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

Dataset Extraction

Task Description

In the scientific literature, datasets are cited to reference, for example, the data on which an analysis is performed or on which a particular result or claim is based. In this competition, we focus on extracting and disambiguating dataset mentions from social science

publications to a list of given dataset references. Identifying dataset mentions in literature is a challenging problem due to the huge number of styles for citing datasets. Although there are proposed standards for dataset citation in full texts, researchers still ignore or neglect such standards (see, for example, Altman and King, 2007). Furthermore, in many research publications, a correct citation of datasets is often missing (Boland et al., 2012). The following two sentences exemplify the problem of the usage of an abbreviation to make a reference to an existing dataset. Example 1 illustrates the use of abbreviations that are known mainly in the author's research domain. Example 2 illustrates the ambiguity of abbreviations. In this case, *WHO* identifies a dataset published by the World Health Organization and does not refer to the institution itself.

> **Example 1**: P-values are reported for the one-tail paired t-test on *Allbus* (dataset) and *ISSP* (dataset).*

> **Example 2**: We used *WHO data* (dataset) from 2001 to estimate the spreading degree of AIDS in Uganda.

We treat the problem of detecting dataset mentions in full text as an NER task.

Formal Problem Definition

Let D denote a set of existing datasets d and the knowledge base 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).

Challenges

We focus on the extraction of dataset mentions in the body of the full text of scientific publications. There are three types of dataset mentions: (i) the full name of a dataset ('National Health and Nutrition Examination Survey'); (ii) an abbreviation ('NHaNES'); and (iii) a vague reference (e.g. 'the monthly statistic'). With all three types, the NER task faces special challenges. In the first case, the dataset name used can vary in different publications. For instance, one publication cites the dataset as 'National Health and Nutrition Examination Survey', while another could use the words 'Health and Nutrition Survey'. In the case where abbreviations are used, a disambiguation problem occurs, for example, in 'WHO data'. WHO may describe the World Health Organization or the White House Office. If an abbreviation is used after the dataset name has been written in full, the mapping between these different spellings in one text is referred to as coreference resolution. The biggest challenge is again the lack of annotated training data. In the following we describe how we dealt with this lack of ground truth data.

Phase 1 Approach

Missing ground truth data was the main problem to be dealt with during this competition. Supervised learning methods for dataset mention extraction from texts are not applicable without the identification of external training data or the creation of useful labelled training data from information given by the competition. Because of the lack of existing training data for the task of dataset mention extraction we resorted to the list of dataset mentions and publication pairs provided and reannotated the particular sentences in the publication text. A list of dataset-identifying words was provided by the competition for some of the known links between publications and datasets. These words represent the evidence of the linkage between publication and datasets and were extracted from the publication text. In the course of reannotation, we searched for each of the identifying words in the corresponding publication texts. For each match, we annotated the occurrence in our raw text and used these annotations as ground truth. As described in the preprocessing section, our units for processing the publication text are paragraphs. The reannotated corpus consists of a list of paragraphs for each publication with stand-off annotations identifying the mentions of datasets (i.e. position of the start and end characters and the entity type for each mention: *dataset*). This reannotation was then used to train spaCy's neural-network-based NER model.[16] We created a holdout set of 1000 publications and a training set of size 4000. Afterwards, we trained our model with the paragraph as a sampling unit. In the training set, 0.45% of the paragraphs contained mentions. For each positive training example, we added one negative sample that contains no known dataset mentions and is randomly selected. 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: precision and recall, each for strict and for partial match. While the strict match metrics are standard evaluation metrics, the partial match metrics are their relaxed variants in which the degree to which dataset mentions have to match can vary. Consider the following partial match example: 'National Health and Nutrition Examination Survey' is the extracted dataset mention, while 'National Health and Nutrition Examination Survey (NHANES)' is the true dataset mention. In contrast to the strict version of the metrics, this overlapping match is considered a match for the partial version. The scores describe whether a model is able to find the correct positions of dataset mentions in the texts, even if the start and end positions of the characters are not the same, but the ranges overlap.

Table 8.2 Performance of phase 1 approach for dataset extraction

Metric	Value
Precision (partial match)	0.93
Recall (partial match)	0.95
Precision (strict match)	0.80
Recall (strict match)	0.81

Table 8.2 shows the results of the dataset mention extraction on the holdout set. The model can achieve high strict precision and recall values. As expected, the results are even better for the partial version of the metrics. This means that even if we could not match the dataset mention in a text exactly, we can find the right context with very high precision.

Phase 2 Approach

In the second phase of the competition, 5000 additional publications were provided by RCC. We extended our approach to consider the list with dataset names supplied by the organizers and reannotated the complete corpus of 15,000 publications in the same manner as in phase 1 to obtain training data. This time we split the data into 80% for training and 20% for test.

Evaluation

We resorted to the same evaluation metrics as in phase 1. However, we calculated precision and recall on the full text of the publication and not on the paragraphs as in the first phase.

Table 8.3 shows the results achieved by our model. We observe lower precision and recall values. Compared to phase 1, there is also a smaller difference between the precision and recall values for the strict and partial version of the metrics.

Table 8.3 Performance of phase 2 approach for dataset extraction

Metric	Value
Precision (partial match)	0.51
Recall (partial match)	0.90
Precision (strict match)	0.49
Recall (strict match)	0.87

Research Method Extraction

Task Description

Inspired by recent work by Nasar et al. (2018), 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*. The reason is that the related semantic entity types might provide clues or might be directly related to the research method itself.

For example, in order to achieve a certain research goal, an experiment is used in which a certain combination of *methods* is applied to a *dataset*. The methods can be specified as concepts or indirectly through the use of certain *software*. The result is then quantified with a *performance* using a specific measure.

Example 3: *P-values* (measurement) are reported for the *one-tail paired t-test* (method) on *Allbus* (dataset) and *ISSP* (dataset).

We selected the entity types *research method*, *research theory*, *research tool* and *research measurement* as the target research method related entity types (see Table 8.4). This decision is based on an examination of the Sage ontology given by the RCC as a sample of how research method terms might look.

Table 8.4 Entity types of relevance for the research method extraction task

Entity type	Corresponding Sage type	Examples
Research Method	SAGE-METHOD	bootstrapping, active interviews
Research Measurement	SAGE- MEASURE	latent variables, phi coefficient, Z-score
Research Theory	SAGE-THEORY	Frankfurt school, feminism, actor network theory
Research Tool	SAGE-TOOL	SPSS, R statistical package

Formal Problem Definition

The task of named entity recognition and linking is to (i) identify the mentions, *m*, of research-related entities in a sentence and (ii) link them, if possible, to a reference knowledge base, *K*, (i.e. the Sage Thesaurus)[17] or (iii) assign a type to each entity (e.g. a *research method*) selected from a set of predefined types.

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.[18] 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 the *linking problem*. Note that part of this problem also results from PDF-to-text conversion, which is error-prone. Dealing with incomplete knowledge bases, that is, *handling out-of-vocabulary 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, resulting in insightful co-occurrence statistics, providing additional detail directly related to entity resolution, and helping to assess the *relevance of terms* by means of a score.

Our Approach

Our research method extraction tool builds on Stanford's CoreNLP and Named Entity Recognition System.[19] The information extraction process follows the workflow depicted in Figure 8.2, using separate modules for preprocessing, classification, linking and term filtering.

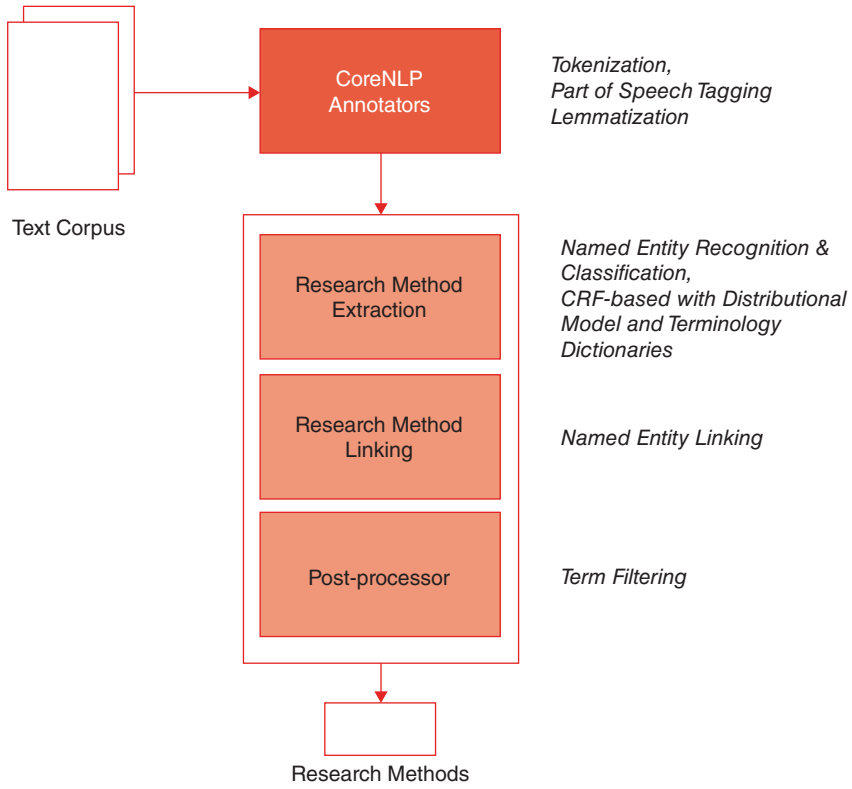

Figure 8.2 Overview of the entity extraction pipeline

We envision the task of finding entities in scientific publications as a sequence labelling 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 conditional random field (CRF) based machine learning classifier with major semantic classes (see Table 8.4), using training material from the ACL RD-TEC 2.0 dataset (QasemiZadeh and Schumann, 2016). Apart from this, we follow a domain adaptation approach inspired by Agerri and Rigau (2016) and ingest semantic background knowledge extracted from external scientific corpora, in particular the ACL Anthology (Bird et al., 2008; Gildea et al., 2018). We perform entity linking by means of a new gazetteer based on the *Sage Encyclopedia of Social Research Methods* (Lewis-Beck et al., 2003), 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 termhood and unithood, while scoring is achieved by calculating a relevance score based on TF-IDF (see Table 8.6).

Our research experiments are based on publications from SSOAR[20] as well as the train and test data of the RCC corpus.[21] Our work extends previous work on this topic (see Eckle-Kohler et al., 2013) in various ways. First, we do not limit our study to abstracts, but use the entire full text. 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 topical information from word embeddings trained on a scientific corpus as an additional feature to our NER model. For this, we use agglomerative clustering of the word embeddings to identify topical groups of words. The cluster number of each word is used as additional sequential input feature for our CRF model. Semantic representations of words are a successful extension of common features, resulting in higher NER performance (Turian et al., 2010) and can be trained offline. In this work, the word vectors were learned based on 22,878 documents of the scientific ACL Anthology Reference Corpus[22] using Gensim[23] with the skip-gram model (see Mikolov et al., 2013) and a preclustering algorithm.[24]

Features

The features incorporated into the linear chain CRF are shown in Table 8.5. 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, orthographics, gazetteer, part-of-speech (POS) tags, along with word clustering have been used.

Table 8.5 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

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').[25] 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 SSOAR. More concretely, we carried out the following steps to extend Sage as an offline 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 term variants such as abbreviations, synonyms, related terms from SSOAR (semi-automatic)
3 Computation of term and document frequency scores for SSOAR (automatic).

Extracting Term Variants Such as Abbreviations, Synonyms and Related Terms

A total of 26,082 candidate terms have been recognized and classified by our pipeline and manually inspected to find synonyms and related words that could be linked to Sage, and to build a post-filter for incorrectly classified terms. Moreover, abbreviations have been extracted using the algorithm of Schwartz and Hearst (2003). In this way, a named entity gazetteer could be built and is used at run-time. It comprises 1111 terms from Sage and 447 terms from the glossary of statistical terms,[26] 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 offline 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 8.6.

Table 8.6 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

Definition of a Relevance Score

The 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') (Kageura and Umino, 1996). Regarding *unithood*, the NER model implicitly contains heuristics about legal POS tag sequences for candidate terms, consisting of at least one noun (NN), preceded or followed by modifiers such as adjectives (JJ), participles (VB*) or cardinal numbers (CD), complemented by word-shape features.

In order to find out if the candidate term also fulfils 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 are 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 full text papers from SSOAR and 10 documents from the RCC, all randomly selected from a holdout corpus. 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: the judges reviewed 10 random publications and generated qualitative scores for each document. In this evaluation, the research method extraction tool received the overall best results of all competitors for this task.[27]

Research Field Classification

Task Description

The goal of this task is to identify the research fields covered in the social science publications. In general, two approaches could be applied to this task. One is the extraction of relevant terms from the publications. This approach sees the task as a keyword extraction task and considers the terms detected as descriptive terms regarding the research field. The second approach is to learn to classify publications' research fields with the use of annotated data in a supervised manner. The benefit of the second approach is that the classification scheme to describe the research field can be defined by domain experts. The disadvantage of supervised trained classifiers for this task is the lack of applicable training data. Furthermore, it must be ensured that the training data is comparable to the texts the research field classifier should be applied on.

Formal Problem Definition

Let P denote a set of publications of size n, A a set of corresponding abstracts of the same size, and L a set of k defined class labels describing research fields. The task of research field classification is to select, for each publication $p_i \in P$, based on the information contained in the corresponding abstract $a_i \in A$, a set of n labels $\{C_i = \varnothing \cap c_1 \dots c_n \mid c_n \in L\}$.

n denotes the number of labels from L describing the research field a_i and can vary for each publication P_i. If there is no label l_k representing the information given by the abstract a_i, the set of class labels is the empty set $\varnothing$.

Our Approach

Since we did not receive any gold standard for this task during the competition we decided to make use of external resources. We decided to use an external labelled dataset

to train a text classifier which is able to predict one or more research labels for a given abstract of a publication.

Because publications given throughout the competition belong to the domain of social sciences we considered training data from the same domain. Namely from SSOAR. The advantages are twofold. On the one hand, we could rely on professional annotations in a given classification scheme covering the social sciences and related areas. On the other hand, the source is openly available.[28]

The SSOAR annotated data contains four different annotation schemes for research-field-related information. Having reviewed these schemes, we decided to use the Classification Social Science (classoz) annotation scheme. The number of classes in each schema, coverage of each classification, and the distribution of data in each schema affected our decision. An exhaustive description of the data used was given earlier in this chapter.

Preprocessing and Model Architecture

SSOAR is a multilingual repository. Therefore, the available abstracts may vary in language, and the language of the abstract may differ from the language of the article itself. We selected all English abstracts with valid classification as our dataset, mainly because of the language of the RCC corpus. However, it should be noted that the multilingual SSOAR abstract corpus has a skewed distribution of languages, with English and German as the main languages. We counted 22,453 English abstracts with valid classification after filtering. Due to the unequal distribution of labels in the dataset, we needed to guarantee 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 out of 154 classification labels representing research fields. To create train and test sets, 22,453 SSOAR publications with their assigned labels were split randomly. We used a train/validation/test split of 70/10/20. We decided to train a text classifier based on a fastText (Joulin et al., 2017) model. The used implementation was written by the authors of this paper.[29]. The arguments to use this model were the speed compared to a more complex neural net architecture and that the performance was comparable to the state of the art (e.g. Wang et al., 2018). The model was trained with learning rate 1.0 for 150 epochs. Also, the negative sampling parameter was set to 25.

Evaluation

Figure 8.3 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 multilabel classification, this allows us to evaluate our model from different perspectives. As illustrated in Figure 8.3, the intersection of the micro-precision and the micro-recall curves is at the threshold of 0.1, where the highest micro F_1 score is achieved. By increasing the

threshold from this point, the micro-precision score is increasing, but the micro-recall is falling. By decreasing the threshold, these trends are inverted. Also, the default threshold of 0.5 does not look promising. In spite of a micro-precision score of about 0.75, we have a problem with the very high number of items without any prediction. In respect of this observation it is advantageous to select a lower threshold in a productive environment. The 'without prediction' curve shows for a given threshold the share of publications in the test set without any prediction. If the selected threshold value is high, the number of publications for which the model cannot predict a research field increases. For example, a selected threshold value of 0.55 leads to 40% unclassified publications in the test set. The 'one correct' curve indicates the quality of the publication-wise prediction. It shows the share of all publications in the test set where at least one of the predicted research field labels can be found in the ground truth data. For instance, if a threshold of 0.1 is selected for 75% of the publications in the test set, at least one of the model predictions is correct. This value decreases with increasing threshold similar to the recall metric. The final micro F_1 value on the test set for our model and a selected threshold of 0.1 is 0.56 (precision 0.55, recall 0.56).

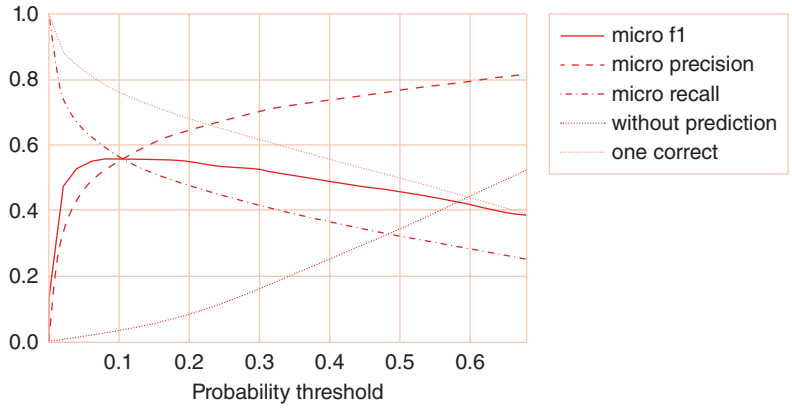

Figure 8.3 Performance for different selected probability thresholds (validation set)

Discussion and Limitations

Dataset Extraction

For the dataset extraction task, the proposed methods are only tested on social-science-related data. The performance measures we have introduced are based on a holdout dataset from our automatically created dataset. The recall may be biased, given that our training and test datasets are biased towards known datasets, and datasets not yet part of our reference set are not considered.

The results of the second phase presented during the RCC workshop[30] show that our approach performs well compared to the other finalist teams, with the highest precision (52.2%; second, 47.0%) and second-best recall (20.5%; best, 34.8%). With respect to F_1, our approach provides the second-best-performing system for this task (29.5%; best, 40.0%). The results on the manually created holdout set underline that our system performs better with respect to precision compared to the other finalist teams. Given that our models are supervised through a corpus of social science publications, we anticipate limited generalizability across other disciplines and plan to investigate this aspect as part of future work. In this context, the focus of our training data on survey data, also reflected in dataset titles such as *Current Population Survey*, could have biased the model to detect the survey as a specific type of research datasets better than other subtypes such as text corpora in the NLP community. In general, however, our approach to using a weakly labelled corpus created from a list of dataset names could be applied in other research domains.

Research Method Extraction

We consider the extraction of research methods from full text as a particularly challenging task because the sample vocabulary given by the RCC organizers covers a large thematic variety of areas. The task itself was defined as the identification of research methods associated with a specific publication, which in turn are drawn from a specific research field. Since no training data has been provided, we created and annotated a new corpus for the task and trained a CRF model, adding lexical resources. The qualitative reviews during the two phases of the competition attested that this approach works fine.

Research Field Classification

Our supervised machine learning approach to the research field classification task performs well on the dataset created from social science publication metadata. A micro F_1 measure of over 55% seems to indicate reasonable performance, considering the small dataset with 44 labels and a mean number of keywords of three terms per publication. As one example of multilabel classification with a comparable size of labels we would like to mention the classification of texts in the domain of medicine presented in Wang et al. (2018). The models tested by the authors on the task of multilabel prediction from 50 different labels led to micro F_1 values between 53% and 62%. Considering the evaluation approach, focused on publications from the social sciences, the generalizability across other disciplines remains unclear and requires further research. Even though the classification scheme used may cover neighbouring disciplines (e.g. medicine), the numbers of samples of the training data covering research fields other than the social sciences is limited. Our pragmatic approach of basing our classifications on the abstracts of the publications makes it applicable even in scenarios where the full text of publications is not accessible.

Conclusion

This chapter has provided an overview of our solutions submitted to the Rich Context Competition 2018. With the aim of improving search, discovery and interpretability of scholarly resources, we address three distinct tasks all concerned with extracting structured information about research resources from scientific publications, namely the extraction of dataset mentions, the extraction of mentions of research methods and the classification of research fields.

In order to address all the aforementioned challenges, our pipelines make use of a range of preprocessing techniques together with state-of-the-art NLP methods as well as supervised machine learning approaches tailored towards the specific nature of scholarly publications as well as the dedicated tasks. In addition, background datasets have been used to facilitate supervision of methods at larger scale.

Our results indicate both significant opportunities for automating the three tasks but also their challenging nature, in particular given the lack of publicly available gold standards for training and testing. Aggregating and publishing such data has been identified as an important activity for future work, and is a prerequisite for significantly advancing state-of-the-art methods.

Acknowledgements

This work has been partially funded by the Deutsche Forschungsgemeinschaft (DFG) under grant number MA 3964/7–1. Wolfgang Otto acknowledges the enabling support provided by the Indo-German Joint Research Project titled 'Design of a Sciento-text Computational Framework for Retrieval and Contextual Recommendations of High-Quality Scholarly Articles' (grant no. DST/INT/FRG/DAAD/P–28/2017) for this work.

References

Agerri, R. and Rigau, G. (2016) Robust multilingual named entity recognition with shallow semi-supervised features. *Artificial Intelligence*, 238, 63–82.

Altman, M. and King, G. (2007) A proposed standard for the scholarly citation of quantitative data. *D-lib Magazine*, 13(3/4).

Backes, T. (2018a) Effective unsupervised author disambiguation with relative frequencies. In J. Chen, M. A. Gonçalves, J. M. Allen, et al. (eds), *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*. New York: ACM, pp. 203–212.

Backes, T. (2018b) The impact of name-matching and blocking on author disambiguation. In A. Cuzzocrea, J. Allan, N. W. Paton, et al. (eds), *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. New York: ACM, pp. 803–812. Available at: http://dblp.uni-trier.de/db/conf/cikm/cikm2018.html#Backes18.

Bird, S., Dale, R., Dorr, B. J. et al. (2008) 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 '08). Marrakech: . European Language Resources Association..

Boland, K. and Krüger, F. (2019) Distant supervision for silver label generation of software mentions in social scientific publications. In *Proceedings of the 4th Joint Workshop on Bibliometric-Enhanced Information Retrieval and Natural Language Processing for Digital Librarie*s. Paris: CEUR-WS, pp. 15–27.

Boland, K., Ritze, D., Eckert, K. and Mathiak, B. (2012) Identifying references to datasets in publications. In P. Zaphiris, G. Buchanan, E. Rasmussen and F. Loizides (eds), *Theory and Practice of Digital Libraries*. Berlin: Springer, pp. 150–161.

Eckle-Kohler, J., Nghiem, T.-D. and Gurevych, I. (2013) 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*. Montreal: American Society for Information Science, p. 44.

Finkel, J. R, Grenager, T. and Manning, C. (2005) Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*. New Brunswick, NJ: Association for Computational Linguistics, pp. 363–370.

Ghavimi, B., Mayr, P., Lange, C. et al. (2016) A semi-automatic approach for detecting dataset references in social science texts. *Information Services & Use* 36(3–4), 171–187.

Gildea, D., Kan, M.-Y., Madnani, N. et al. (2018) The ACL Anthology: Current state and future directions. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS). Melbourne:* Association for Computational Linguistics pp. 23–28.

Hienert, D., Kern, D., Boland, K. et al. (2019) A digital library for research data and related information in the social sciences. In M. Bonn, D. Wu, J. S. Downie and A. Martaus (eds), *Proceedings of the 18th ACM/IEEE Joint Conference on Digital Libraries*. *Champaign, IL: IEEE Press*, pp. 148–157. Available at: http://dblp.uni-trier.de/db/conf/jcdl/jcdl2019.html#HienertKBZM19

Joulin, A., Grave, E., Bojanowski, P. et al. (2017) 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*. Valencia: Association for Computational Linguistics, pp. 427–431.

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

Krämer, T., Klas, C.-P. and Hausstein, B. (2018) A data discovery index for the social sciences. *Scientific Data*, 5, 180064.

Lewis-Beck, M., Bryman, A. E. and Liao, T. F. (2003) *The Sage Encyclopedia of Social Science Research Methods*. Thousand Oaks, CA: Sage.

Mikolov, T., Sutskever, I., Chen, K. et al. (2013) Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26, pp. 3111–3119.

Nasar, Z., Jaffry, S. W. and Malik, M. K. (2018) Information extraction from scientific articles: A survey. *Scientometrics*, 117(3), 1931–1990.

QasemiZadeh, B. and Schumann, A.-K. (2016) The ACL RD-TEC 2.0: A language resource for evaluating term extraction and entity recognition methods. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC '16)*. Portorož, Slovenia*: European L*anguage Resources Association, pp. 1862–1868.

Sahoo, P., Gadiraju, U., Yu, R., Saha, S. and Dietze, S. (2017) Analysing structured scholarly data embedded in web pages. In A. González-Beltrán, F. Osborne, S. Peroni and S. Vahdati (eds), *Semantics, Analytics, Visualization: Enhancing Scholarly Data* (Lecture Notes in Computer Science 9792). Cham: Springer, pp. 90–100.

Schwartz, A. S. and Hearst, M. A. (2003) A simple algorithm for identifying abbreviation definitions in biomedical text. In *Pacific Symposium on Biocomputing, 2003*. River Edge, NJ: World Scientific, pp. 451–462.

Tkaczyk, D., Szostek, P., Fedoryszak, M. et al. (2015) CERMINE: Automatic extraction of structured metadata from scientific literature. *International Journal on Document Analysis and Recognition*, 18(4), 317–335.

Turian, J., Ratinov, L. and Bengio, Y. (2010) Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics, pp. 384–394. Available at: http://dl.acm.org/citation.cfm?id=1858681.1858721

Wang, G., Li, C., Wang, W. et al. (2018) Joint embedding of words and labels for text classification. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne: Association for Computational Linguistics. doi: 10.18653/v1/p18–1216

Yu, R., Gadiraju, U., Fetahu, B. et al. (2019) KnowMore – knowledge base augmentation with structured web markup. *Semantic Web*, 10(1), 159–180. Available at: http://dblp.uni-trier.de/db/journals/semweb/semweb10.html#YuGFLRD19

Notes

1 https://www.gesis.org/en/institute
2 https://www.gesis.org/en/institute/departments/knowledge-technologies-for-the-social-sciences/
3 https://www.gesis.org/en/research/applied-computer-science/labs/wts-research-labs
4 https://www.gesis.org/en/research/external-funding-projects/archive/infolis-i-and-ii
5 https://search.gesis.org
6 https://datasearch.gesis.org/
7 https://coleridgeinitiative.org/richcontextcompetition
8 https://www.ssoar.info
9 https://fasttext.cc/
10 https://www.gesis.org/ssoar/home
11 https://www.gesis.org/en/services/research/tools/thesaurus-for-the-social-sciences
12 https://www.gesis.org/angebot/recherchieren/tools-zur-recherche/klassifikation-sozialwissenschaften (in German)
13 http://www.openarchives.org
14 https://github.com/CeON/CERMINE
15 https://jats.nlm.nih.gov
16 https://spacy.io
17 http://methods.sagepub.com
18 Apart from those used in traditional NER systems such as *person*, *location*, or *organization* with abundant training data, as covered in the Stanford NER system (Finkel et al., 2005).
19 https://nlp.stanford.edu/projects/project-ner.shtml
20 https://www.ssoar.info
21 https://coleridgeinitiative.org/richcontextcompetition, with a total of 5000 English documents
22 https://acl-arc.comp.nus.edu.sg/
23 https://radimrehurek.com/gensim/

24 Word embeddings are trained with a skip-gram model using embedding size equal to 100, word window equal to 5, minimal occurrences of a word for it to be considered 10. Word embeddings are clustered using agglomerative clustering with number of clusters set to 500, 600, 700. Word linkage with Euclidean distance is used to minimize the variance within the clusters.
25 A glossary of statistical terms as provided at https://www.statistics.com/resources/glossary/ has been added as well.
26 Based on https://www.statistics.com/resources/glossary
27 Rank 1, 2, 2, 1, 1 for judges 1–5.
28 A script to download the SSOAR metadata can be found at github/research-field-classifier
29 https://fasttext.cc/
30 The workshop agenda can be found at https://coleridgeinitiative.org/richcontextcompetition/workshopagenda. The results of the finalists are presented at https://youtu.be/PE3nFrEkwoU?t=9865.