2020

Acknowledgement Entity Recognition in CORD-19 Papers

Jian Wu
Old Dominion University

Pei Wang
Old Dominion University

Xin Wei
Old Dominion University

Sarah Rajtmajer
Pennsylvania State University

C. Lee Giles
Pennsylvania State University

See next page for additional authors

Original Publication Citation

Wu, J., Wang, P., Wei, X., Rajtmajer, S., Giles, C. L., & Griffin, C. (2020). Acknowledgement entity recognition in CORD-19 papers. In *Proceedings of the First Workshop on Scholarly Document Processing* (pp. 10-19). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.sdp-1.3

Acknowledgement Entity Recognition in CORD-19 Papers

Jian Wu
Computer Science
Old Dominion University
Norfolk, VA, USA
jwu@cs.odu.edu

Pei Wang, Xin Wei
Computer Science
Old Dominion University
Norfolk, VA, USA
{pwang001,xwei001}@odu.edu

Sarah Michele Rajtmajer, C. Lee Giles
Information Sciences and Technology
Pennsylvania State University
University Park, PA, USA

Christopher Griffin
Applied Research Laboratory
Pennsylvania State University
University Park, PA, USA

Abstract

Acknowledgements are ubiquitous in scholarly papers. Existing acknowledgement entity recognition methods assume all named entities are acknowledged. Here, we examine the nuances between acknowledged and named entities by analyzing sentence structure. We develop an acknowledgement extraction system, ACKEXTRACT based on open-source text mining software and evaluate our method using manually labeled data. ACKEXTRACT uses the PDF of a scholarly paper as input and outputs acknowledgement entities. Results show an overall performance of $F_1 = 0.92$. We built a supplementary database by linking CORD-19 papers with acknowledgement entities extracted by ACKEXTRACT including persons and organizations and find that only up to 50–60% of named entities are actually acknowledged. We further analyze chronological trends of acknowledgement entities in CORD-19 papers. All codes and labeled data are publicly available at https://github.com/lamps-lab/ackextract.

1 Introduction

Acknowledgements have been an institutionalized part of research publications for some time (Blaise, 2001). Acknowledgement statements show the authors' public gratitude and recognition to individuals, organizations, and grants for various contributions. Acknowledged individuals and organizations have been under-presented in author ranking and citation impact analysis mostly due to their presumed sub-authorship contribution. A recent survey found that discipline, academic rank, and gender have a significant effect on coauthorship disagreement rate (Smith et al., 2019), leading to non-author collaborators receiving less attention. Recently, the presence of non-author collaborators in the biomedical and social sciences (Paul-Hus et al., 2017) showed that non-author collaborators are not rare and their presence varies significantly by disciplines.

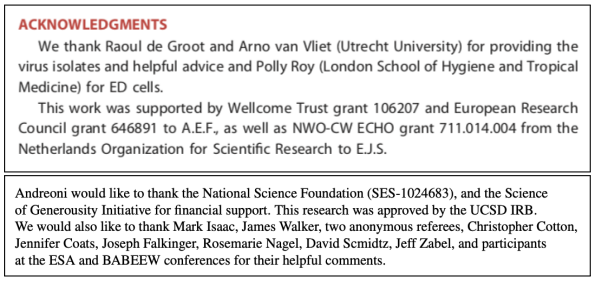
Acknowledgements can be classified depending on the nature of the contribution. Song et al. (2020) classified sentences in acknowledgement sections into 6 categories: declaration, financial, peer interactive communication and technical support, presentation, general acknowledgement, and general statement. They can also be classified based on the type of entities such as individual, organization, and grant. Since 2008, funding acknowledgements have been indexed by Web of Science. However, there is still no dedicated software to accurately recognize acknowledged *people* and *organizations* and generate a centralized acknowledgement database. Early works on acknowledgements were based on datasets manually extracted from specific journals, which was not scalable. Building such a large database can support further study of acknowledgements at a larger scale.

There are several scenarios that make the acknowledgement entity recognition (AER) task challenging. The upper panel of Figure 1 shows examples of sentences appearing in an isolated acknowledgement section. The "Utrecht University" is mentioned but should not be counted as an acknowledgement entity because it is just the affiliation of "Arno van Vliet" who is acknowledged. Acknowledgement statements can also appear at footnotes (Figure 1 *Bottom*), mixed with other footnote and/or body text. Author names may also appear in the statements, such as "Andreoni" in this example, and should be excluded.

Existing works on AER leverage off-the-shelf named entity recognition (NER) packages, such

Proceedings of the First Workshop on Scholarly Document Processing, pages 10–19

https://doi.org/10.18653/v1/P17



ACKNOWLEDGMENTS

We thank Raoul de Groot and Arno van Vliet (Utrecht University) for providing the virus isolates and helpful advice and Polly Roy (London School of Hygiene and Tropical Medicine) for ED cells.

This work was supported by Wellcome Trust grant 106207 and European Research Council grant 646891 to A.E.F., as well as NWO-CW ECHO grant 711.014.004 from the Netherlands Organization for Scientific Research to E.J.S.

Andreoni would like to thank the National Science Foundation (SES-1024683), and the Science of Generousity Initiative for financial support. This research was approved by the UCSD IRB. We would also like to thank Mark Isaac, James Walker, two anonymous referees, Christopher Cotton, Jennifer Coats, Joseph Falkinger, Rosemarie Nagel, David Scmidtz, Jeff Zabel, and participants at the ESA and BABEEW conferences for their helpful comments.

Figure 1: *Upper:* Acknowledgement statements appear in an isolated section (Stewart et al., 2018). *Bottom:* acknowledgement statements appear in a footnote (Andreoni and Gee, 2015).

as the Natural Language Toolkits (NLTK) (Bird, 2006) e.g., Khabsa et al. (2012); Paul-Hus et al. (2020), followed by simple semi-manual data cleansing, resulting in a fraction of entities that are *mentioned* but *not actually acknowledged*. In this paper, we design an automatic AER system called ACKEXTRACT that further classifies extraction results from open source NER packages that recognize *people* and *organizations* and distinguish entities that are *actually acknowledged*. The extractor finds acknowledgement statements from isolated sections and other locations such as footnotes, which is common for papers in social and behavioral sciences. Our contributions are:

1. Develop ACKEXTRACT as open-source software to automatically extract acknowledgement entities from research papers.
2. Apply ACKEXTRACT on the CORD-19 dataset and supplement the dataset with a corpus of classified acknowledgement entities for further studies.
3. Use the CORD-19 dataset as a case study to demonstrate that acknowledgement studies without classifying named entities, can significantly overestimate the number of entities that are actually acknowledged because many people and organizations are mentioned but not explicitly acknowledged.

2 Related Works

Early work on acknowledgement extraction was manually applied, which was labor-intensive. Cronin et al. (1993) extracted a total of 9561 peer interactive communication (PIC) names from a total of 4200 research sociology articles, most were persons' names. They also defined the following six categories of acknowledgement: moral support, financial support, editorial support, presentational support, instrumental/technical support, and conceptual support, or PIC (Cronin et al., 1992).

Councill et al. (2005) used a hybrid method for automatic AER from research papers and automatically created an acknowledgement index(Giles and Councill, 2004). The algorithm first used a heuristic method for identifying acknowledgement passages. It then uses an SVM model for identifying lines containing acknowledgement sentences outside labeled acknowledgement sections. A regular expression was used to extract entity names from acknowledging text. This method achieved an overall precision of about 0.785 and a recall of 0.896 on CiteSeer papers (Giles et al., 1998). The algorithm does not distinguish entity types.

Khabsa et al. (2012) leveraged OpenCalais[1] and AlchemyAPI[2], free services at that time, to extract named entities from acknowledgement sections and built ACKSEER, a search engine for acknowledgement entities. They merged outputs of both NER APIs and generated a list and disambiguated entity mentions using the longest common subsequence (LCS) algorithm. The ground truth contains 200 top-cited CiteSeerX papers in which 130 had acknowledgement sections. They achieved 92.3% and 91.6% precision and recall for acknowledgement section extraction but did not evaluate entity extraction.

Recent studies of acknowledgements tend to use results from off-the-shelf NER packages with simple filters, assuming that *named entities* were *acknowledged entities*. For example, Paul-Hus et al. (2020) uses the Stanford NER module in NLTK to extract persons. Song et al. (2020) also directly use people and organizations recognized by the Stanford CoreNLP (Manning et al., 2014). These works achieved a high recall by recognizing most name entities in the acknowledgements but ignored their relations to the papers where they appear, resulting in a fraction of entities that are mentioned but not actually acknowledged. Song et al. (2020) consider grammar structure such as verb tense and voice and sentence patterns when labeling sentences to their six categories. For example, "was funded" is followed by an "organization". However, they only label sentences and do not annotate them down to

[1] https://web.archive.org/web/20081023021111/http://opencalais.com/calaisAPI#

[2] https://web.archive.org/web/20090921044923/http://www.alchemyapi.com/

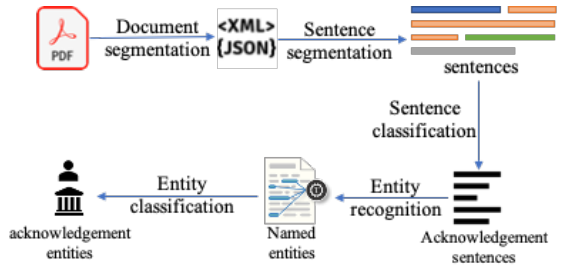


Figure 2: Architecture of ACKEXTRACT.

the entity level. Our system examines the relationship between entities and the current work, with the purpose of discriminating *acknowledgement entities* from *named entities*. In this system, we focus on *people* and *organizations*.

Recently, Dai et al. (2019) proposed GrantExtractor, a pipeline system to extract grant support information from scientific papers. A model combining BiLSTM-CRF and pattern matching was used to extract entities of grant numbers and agencies from funding sentences, which are identified using heuristic methods. The system achieves a micro-F_1 up to 0.90 in extracting grant pairs (agency, number).

Kayal et al. (2019) proposed an *ensemble* approach called FUNDINGFINDER for extracting funding information from text. The authors construct feature vectors for candidate entities using whether the entities are recognized by four NER implementation: Stanford (Conditional Random Field model), LingPipe (Hidden Markov model), OpenNLP (Maximum Entropy model), and Elsevier's Fingerprint Engine. The F_1-measure for funding body is only 0.68 ± 0.3.

Our method is different from existing methods in threefold. (1) It is built on top of state-of-the-art neural NER methods, which results in a relatively high recall. (2) It uses a heuristic method to filter out entities that are just mentioned but not acknowledged. (3) It extracts both organizations and people.

3 Dataset

On March 16th, 2020, Allen Institute of Artificial Intelligence, released the first version of the COVID-19 Open Research Dataset (CORD-19) (Wang et al., 2020), in collaboration with several other institutions. The dataset contains metadata and segmented full text of research articles selected by searching a list of keywords about coronavirus, SARS-CoV, MERS, and other related terms from four digital libraries including WHO, PubMed Central, BioRxiv, and MedRxiv. The initial dataset contained about 28k papers and was updated weekly with papers from new sources and the latest publication. We used the dataset released on April 10, 2020 containing over 59,312 metadata records, among which 54,756 have the full text in JSON format. The CORD-19 papers were generated by processing PDFs using the S2ORD pipeline (Lo et al., 2019), in which GROBID (Lopez, 2009) was employed for document segmentation and metadata extraction.

The full text in JSON files is directly used for sentence segmentation. However, we observe that GROBID extraction results are not perfect. In particular, we estimate the fraction of acknowledgement sections omitted. We also estimate the number of acknowledgement entities omitted in the data release due to the extraction error of GROBID. To do this, we downloaded 45,916 full-text PDF papers collected by the Internet Archive (IA) because the CORD-19 dataset does not include PDF files[3]. We found 13,103 CORD-19 papers in the IA dataset.

4 Acknowledgement Extraction

4.1 Overview

The architecture of our acknowledgement entity extraction system is depicted in Figure 2. The system can be divided into the following modules.

1. **Document segmentation.** CORD-19 provides full text as JSON files, but in general, most research articles are published in PDF, so our first step is converting a PDF document to text and segment it into sections. We use GROBID that has shown superior performance over many other document extraction methods (Lipinski et al., 2013). The output is an XML file in TEI schema.
2. **Sentence segmentation.** Paragraphs are segmented into sentences. We compare several sentence segmentation software packages and choose Stanza (Qi et al., 2020) because of its relatively high accuracy.
3. **Sentence classification.** Sentences are classified into acknowledgement and non-acknowledgement statements. The result is a set of acknowledgement statements inside or outside the acknowledgement sections.

[3] https://archive.org/download/covid19_fatcat_20200410

4. **Entity recognition.** Named entities are extracted from acknowledgement statements. We compare four commonly used NER software packages and choose Stanza because of its relatively high performance. In this work, we focus on *person* and *organization*.
5. **Entity classification.** In this module we classify named entities by analyzing sentence structures, aiming at discriminating named entities that are actually acknowledged, rather than just mentioned. We demonstrate that triple extraction packages such as REVERB and OLLIE fail to handle acknowledgement statements with multiple entities in objects in our dataset. The results are acknowledgement entities including *people* or *organizations*.

4.2 Document Segmentation

The majority of scholarly papers are published in PDF format, which are not readily readable by text processors. Several attempts have been made to convert PDF into text (Bast and Korzen, 2017) and segment the document into section and sub-section levels. GROBID is a machine learning library for extracting, parsing, and re-structuring raw documents into TEI encoded documents. Other similar methods have been recently developed such as OCR++ (Singh et al., 2016) and Science Parse[4]. Lipinski et al. (2013) compared 7 metadata extraction methods and found that GROBID (version 0.4.0) achieved superior performance over the others. GROBID trained a cascading of conditional random field (CRF) models on PubMed and computer science papers. The recent version (0.6.0) has a set of powerful functionalities such as extracting and parsing headers and segmenting full-text extraction. GROBID supports a batch mode and an API service mode, the latter enables large scale document processing on multi-core servers such as in CiteSeerX (Wu et al., 2015). A benchmarking result for version 0.6.0 shows that the section title parsing achieves and $F_1 = 0.70$ under the strict matching criteria and $F_1 = 0.75$ under the soft matching criteria[5]. Singh et al. (2016) claims OCR++ achieves better performance than GROBID in several fields evaluated on computer science papers. However, the lack of a service mode API and multi-domain adaptability limits its usability. Science-Parse only extracts key metadata such as title, authors, year, and venue. Therefore, we adopt GROBID to convert PDF documents into XML files.

Depending on the structure and provenance of PDFs, GROBID may miss acknowledgements in certain papers. To estimate the fraction of papers in which acknowledgements were missed by GROBID, we visually inspected a random sample of 200 papers from the CORD-19 dataset, and found that only 146 papers (73%) contain acknowledgement statements, out of which GROBID successfully extracted all acknowledgement statements from 120 papers (82%). For the remaining 26 papers that GROBID failed to parse, 17 papers are in sections, 9 papers are in footnotes. We developed a heuristic method that can extract acknowledgement statements from all 120 papers with acknowledgement statements output by GROBID.

4.3 Sentence Segmentation

The acknowledgement sections or statements extracted above are paragraphs, which needs to be segmented (or tokenized) into sentences. we compared four software packages for sentence segmentation including NLTK (Bird, 2006), Stanza (Qi et al., 2020), Gensim (Řehůřek and Sojka, 2010), and the Pragmatic Segmenter[6].

NLTK includes a sentence tokenization method `sent_tokenize()`, which uses an unsupervised algorithm to build a model for abbreviated words, collocations, and words that start sentences; and then uses that model to find sentence boundaries. **Stanza** is a Python natural language analysis package developed by the Stanford NLP group. Sentence segmentation is modeled as a tagging problem over character sequences, where the neural model predicts whether a given character is the end of a sentence. The `split_sentences()` function in **Gensim** package splits a text and returns list of sentences from a given text string using unsupervised pattern recognition. The **Pragmatic Segmenter** is a rule-based sentence boundary detection gem that works out-of-the-box across many languages.

To compare the above methods, we created a ground truth corpus by randomly selecting acknowledgment sections or statements from 47 papers and manually segmenting them, resulting in 100 sentences. Table 1 shows the comparison results for four methods. The precision is calculated

[4] https://github.com/allenai/spv2

[5] https://grobid.readthedocs.io/en/latest/Benchmarking-pmc/

[6] https://github.com/diasks2/pragmatic_segmenter

Method	Precision	Recall	F_1
Gensim	0.65	0.64	0.65
NLTK	0.72	0.69	0.70
Pragmatic	0.86	0.76	0.81
Stanza	**0.81**	**0.88**	**0.84**

Table 1: Sentence segmentation performance.

by dividing the number of correctly segmented sentences by the total number of sentences segmented. The recall is calculated by dividing the number of correctly segmented sentences by the total number of manually segmented sentences. Stanza outperforms the other three, achieving an $F_1 = 0.84$.

4.4 Sentence Classification

Not all sentences in acknowledgement sections express acknowledgement, such as the following sentence, `The funders played no role in the study or preparation of the manuscript` Song et al. (2020). In this module, we classify sentences into acknowledgement and non-acknowledgement statements. We developed a set of regular expressions that match both verbs (e.g., thank, gratitude to, indebted to), adjectives (e.g., grateful to), and nouns (e.g., helpful comments, useful feedback) to cover as many cases as possible. To evaluate this method, we manually selected 100 sentences, including 50 positive and negative samples from the sentences obtained in Section 4.3. Our results show that 96 out of 100 sentences were classified correctly, resulting accuracy of 0.96.

4.5 Entity Recognition

In this step, named entities are extracted using state-of-the-art NER software packages, including NLTK Bird (2006), Stanford CoreNLP Manning et al. (2014), spaCy Honnibal and Montani (2017), and Stanza (Qi et al., 2020). Stanza is a Python library offering fully neural pre-trained models that provide state-of-the-art performance on many raw text processing tasks when it was released. The NER model adopted the contextualized sequence tagger in Akbik et al. (2018). The architecture includes a Bi-LSTM character-level language model, followed by a one-layer Bi-LSTM sequence tagger with a conditional random field (CRF) encoder. Although Stanza was developed based on Stanford CoreNLP, they exhibit differential performances in our NER task.

The ground truth is built by randomly selecting

Method	Entity	Precision	Recall	F_1
NLTK	Person	0.45	0.68	0.55
	Org.	0.59	0.77	0.67
spaCy	Person	0.74	0.88	0.80
	Org.	0.63	0.74	0.68
Stanford-CoreNLP	Person	0.88	0.87	0.87
	Org.	0.68	0.80	0.73
Stanza	Person	**0.89**	**0.93**	**0.91**
	Org.	**0.60**	**0.89**	**0.72**

Table 2: Comparison of NER software package.

100 acknowledgement paragraphs from sections, footnotes, and body text, and manually annotating *person* and *organization* entities, without discriminating whether they are acknowledged or not. This results in 146 *person* and 209 *organization* entities. The comparison results indicate that overall Stanza outperforms the other three, achieving $F_1 = 0.91$ for *person* and $F_1 = 0.72$ for *organization*. Especially, the recall of Stanza is 9% higher than Stanford CoreNLP (Table 2).

4.6 Entity Classification

As we showed, not all named entities are acknowledged, such as the "'Utrecht University" in Figure 1. Therefore, it is necessary to build a classifier to discriminate acknowledgement entities – entities that are thanked by the paper or the authors, from named entities.

The majority of acknowledgement statements in academic articles have a relatively standard subject-predicate-object (SPO) structure. They use a limited number of words or phrases, such as "thank", "acknowledge", "are grateful", "is supported", and "is funded" as predicates. However, the object can contain multiple named entities, some of which are used as attributes of the others. In rare cases, certain sentences may not have subjects and predicates, such as the first sentence in Figure 4.

Our approach can be divided into three steps. Two representative examples are illustrated in Figure 3. The pseudo-code is shown in Algorithm 1.

Step 1: We resolve the type of voice (active or passive), subject, and predicate of a sentence using dependency parsing by Stanza. This is because named entities can appear as subjective or objective parts. We then locate all named entities. The semantic meaning of a predicate and its type of voice can be used to determine whether entities acknowledged are in the objective part or subjective part. In most cases the target entities are objects.

Step 2: A sentence with multiple entities in the objective part is split into shorter sentences, called "subsentences", so that each subsentence is associated with only up to one named entity. This is done by first splitting the sentence by "and". For each subsentence, if the subject and predicate are missing, we fill them up using the subject and predicate of the original sentence. The object in each subsentence does not necessarily contain an entity. For example, in the right panel of Figure 3, because "expertise" is not a named entity, it is replaced by "none" in the subsentence. The SPO relations that do not contain named entities are removed.

There are two scenarios. In the first scenario, a sentence or a subsentence may contain a list of entities, with only the first being acknowledged. In the third example of Figure 4, the named entities are `['Qi Yang', 'Morehouse School of Medicine', 'Atlanta', 'GA']` but only the first entity is acknowledged. The rest entities, which are recognized as *organizations* or *locations*, are used for supplementing more information. In this scenario, only the first named entity is extracted.

Step 3: In the second scenario, acknowledged entities are connected by commas or "and", such as in `We thank Shirley Hauta, Yurij Popowych, Elaine van Moorlehem and Yan Zhou for help in the virus production`. In this scenario, entities in this structure have the same type, indicating that they play similar roles. This parallel pattern can be captured by regular expressions. The entities resolved in Step 2 and 3 will be merged to form the final set.

The method to find the parallel structure is as follows. First, check each entity whether its type is *person*. If so, the entities are substituted with integer indexes. The sentence becomes `The authors would like to thank 0, 1 and 2, and the kind support of Bayer Animal Health GmbH and Virbac Group.` If there are 3 or more consecutive numbers in this form, this is the parallel pattern, which is captured by regular expressions. The pattern also allows text between names (Figure 5). Next, the numbers in this part will be extracted and mapped to corresponding entities. In the example above, the numbers `[0,1,2]` correspond to the index of the entities `[Norbert Mencke, Lourdes Mottier, David McGahieand]`. Similar pattern recognition are performed for *organization*. The process is depicted in Figure 5.

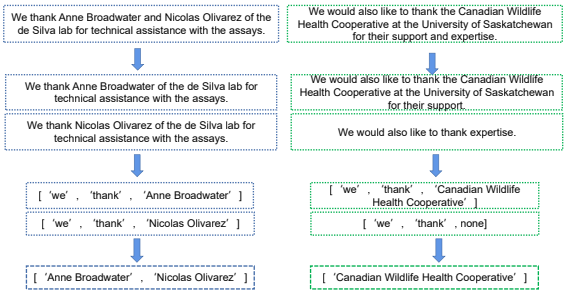


Figure 3: Process mapping multiple subject-predicate-object relations to acknowledgement entities with two representative examples. "None" means the object does not contain named entities.

Method	Precision	Recall	F_1
Stanza	0.57	0.91	0.70
RB(without step 3)	0.94	0.78	0.85
RB	**0.94**	**0.90**	**0.92**

Table 3: Performance of Stanza and our relation-based (RB) classifier. Precision and recall show overall results by combining *person* and *organization*.

We investigated open information extraction methods such as REVERB (Fader et al., 2011) and OLLIE (Mausam et al., 2012). We found that they only work for sentences with relatively simple structures such as `The authors wish to thank Ming-Wei Guo for reagents and technical assistance`, but fail with more complicated sentences with long objective part or parallel subsentences (Figure 4). We also investigated the semantic role labeling (SRL) library in AllenNLP toolkit (Shi and Lin, 2019). For the third sentence in Figure 4, the AllenNLP SRL library resolves the entire string after "thank" as an argument, but fails to distinguish entity types. Our method can handle all statements in Figure 4.

As a baseline, we use Stanza to extract all named entities and compare its performance with our relation-based (RB) classifier. The ground truth is compiled using the same corpus described in Section 4.5 except that only acknowledgement entities (as opposed to all named entities) are labeled positive. The results (Table 3) show that Stanza achieves high recall but poor precision, indicating that a significant fraction ($\sim 40\%$) of named entities are not acknowledged. In contrast, our classifier (RB) achieves a precision of 0.94, with a small loss of recall, achieving an $F_1 = 0.92$.

One limitation of the RB classifier is that it relies on text quality. Sentences are expected to follow

To Aulio Costa Zambenedetti, for the art work, Nilson Fidêncio, Silvio Marques, Tania Schepainski and Sibelli Tanjoni de Souza, for technical assistance.

The authors also thank the Program for Technological Development in Tools for Health-PDTIS FIOCRUZ, for the use of its facilities (Platform RPT09H -Real Time PCR -Instituto Carlos Chagas/Fiocruz-PR).

The authors wish to thank Ms. Qi Yang, DNA sequencing lab, Morehouse School of Medicine, Atlanta, GA, and the Research Cores at Morehouse School of Medicine supported through the Research Centers at Minority Institutions (RCMI) Program, NIH/NCRR/RCMI Grant G12-RR03034.

Figure 4: Sentence examples, from which REVERB or OLLIE failed to extract acknowledgement entities underlined, which can be identified by our method.

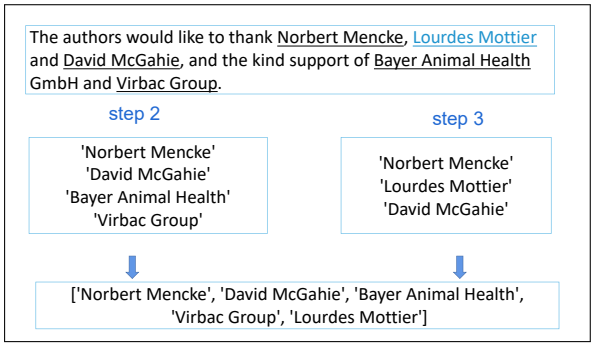


Figure 5: Merge the results given by step 2 and step 3 and obtain the final results.

the editorial convention, such that entity names are clearly segmented by period. If two sentences are not properly delimited, e.g., missing a period, the classifier may make incorrect predictions.

5 Data Analysis

The CORD-19 dataset contains PMC and PDF folders. We found that almost all papers in the PMC folders are included in the PDF folders. For consistency, we only work on papers in the PDF folders, containing 37,612 unique PDF papers[7]. Using ACKEXTRACT, we extracted a total of 102,965 named entities from 21,469 CORD-19 papers. The fraction of papers with named entities (21469/37612≈57%) is roughly consistent with the fraction obtained from the 200 samples (120/200≈60% in Section 4.2). Among all named entities, 58,742 are acknowledgement entities. These numbers suggest that using our model, only about 50% of named entities are acknowledged. Using only named entities, acknowledgement studies could significantly overestimate the

[7]We excluded 2003 duplicate papers with exactly the same SHA1 values.

Algorithm 1: Relation-Based Classifier

```
Function find_entity():
    pre-process with punctuation and format
    find candidate entities entity_list by Stanza
    for x ∈ entity_list do
        if x ∈ subject part (differs based on predicate) then
            x = none
    entity_list.remove(none)
    return entity_list
Input: sentence
Output: entity_list_all
find subject part, predicate and the object part
if "predicate value" ∈ reg expression 1 then
    if "and" ∈ sentence then
        split into subsentences by 'and'
        for each subsentence do
            entity_list ← find_entity(subsentence)
            if entity_list ≠ none then
                entity_list_all ← append.entity_list[0]
    else if "and" ∉ sentence then
        entity_list ← find_entity(subsentence)
        entity_list_all ← entity_list[0]
else if "predicate value" ∈ reg expression 2 then
    do the same in if part but focus on entities in subject
for x ∈ candidate list by Stanza do
    if x ∈ sentence&x.type = type then
        orig_list ← append x
        sent ← replace x by index(x)
    if reg format ∈ sent then
        temp ← find reg expression in sent
        numlist ← find numbers in temp
        list ← index orig_list by numlist
combine list and entity_list_all
```

number of acknowledgement entities by relying on entities recognized by NER software packages without further classification. Here, we analyze our results and study some trends of acknowledgement entities in the CORD-19 dataset.

The top 10 acknowledged organizations (Table 4) are all funding agencies. Overall NIH (NIAID is an institute of NIH) is acknowledged the most, but funding agencies in other countries are also acknowledged a lot in CORD-19 papers.

Figure 6 shows the numbers of CORD-19 papers with vs. without acknowledgement (*person* or *organization*) recognized from 1970 to 2020. The huge leap around 2002 was due to the wave of coronavirus studies during the SARS period. The small drop-down in 2020 was due to data incompleteness. The plot indicates that the fraction of

Organization	Count
National Institutes of Health or NIH	1414
National Natural Science Foundation of China or NSFC	615
National Institute of Allergy & Infectious Diseases or NIAID	209
National Science Foundation or NSF	183
Ministry of Health	160
Wellcome Trust	146
Deutsche Forschungsgemeinschaft	119
Ministry of Education	119
National Science Council	104
Public Health Service	85

Table 4: Top 10 acknowledged organizations identified in our CORD-19 dataset.

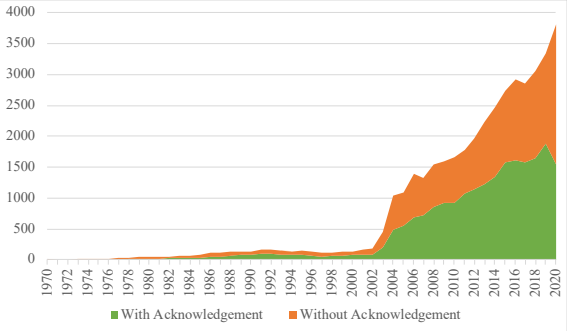


Figure 6: Numbers of CORD-19 papers with vs. without acknowledgement recognized from 1970 to 2020.

papers with acknowledgement has been increasing gradually over the last 20 years. Figure 7 shows the numbers of the top 10 acknowledged organizations from 1983 to 2020. The figure indicates that the number of acknowledgements to NIH has been gradually decreasing over the past 10 years while the acknowledgements to NSF is roughly constant. In contrast, the number has been gradually increasing from NSFC (a Chinese funding agency). Note that the distribution of acknowledged organizations has a long tail and organizations behind the top 10 actually dominate the total number. However, the top 10 organizations are the biggest research agencies and the trend to some extent reflects strategic shifts of funding support.

6 Conclusion and Future Work

Here, we extended the work of Khabsa et al. (2012) and built an acknowledgement extraction framework denoted as ACKEXTRACT for research articles. ACKEXTRACT is based on heuristic methods and state-of-the-art text mining libraries (e.g., GROBID and Stanza) but features a classifier that discriminates acknowledgement entities from named entities by analyzing the multiple subject-predicate-object relations in a sentence. Our approach successfully recognizes acknowledgement entities that cannot be recognized by OIE packages such as REVERB, OLLIE, and the AllenNLP SRL library.

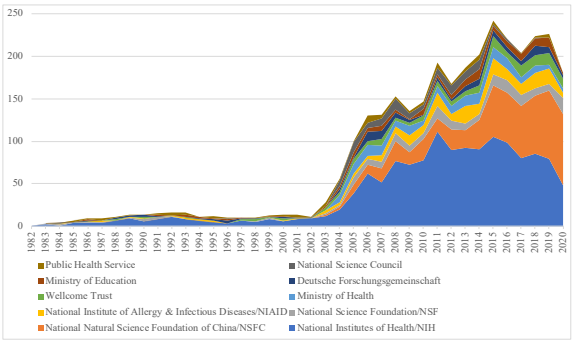


Figure 7: Trend of acknowledged organizations of top 10 organizations from 1983 to 2020.

This method is applied to the CORD-19 dataset released on April 10, 2020, processing one PDF document in 5 seconds on average. Our results indicate that only 50–60% named entities are acknowledged. The rest are mentioned to provide additional information (e.g., affiliation or location) about acknowledgement entities. Working on clean data, our method achieves an overall $F_1 = 0.92$ for *person* and *organization* entities. The trend analysis of the CORD-19 papers verifies that more and more papers include acknowledgement entities since 2002, when the SARS outbreak happened. The trend also reveals that the overall number of acknowledgements to NIH is gradually decreasing over the past 10 years, while more papers acknowledge NSFC, a Chinese funding agency. One caveat of our method is that organizations in different countries are not distinguished. For example, many countries have agencies called "Ministry of Health". In the future, we plan to build learning-based models for sentence classification and entity classification. The code and data of this project have been released on GitHub at: https://github.com/lamps-lab/ackextract.

Acknowledgments

This work was partially supported by the Defense Advanced Research Projects Agency (DARPA). under cooperative agreement No. W911NF-19-2-0272. The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

James Andreoni and Laura K. Gee. 2015. Gunning for efficiency with third party enforcement in threshold public goods. *Experimental Economics*, 18(1):154–171.

Hannah Bast and Claudius Korzen. 2017. A benchmark and evaluation for text extraction from PDF. In *2017 ACM/IEEE Joint Conference on Digital Libraries, JCDL 2017, Toronto, ON, Canada, June 19-23, 2017*, pages 99–108. IEEE Computer Society.

Steven Bird. 2006. NLTK: the natural language toolkit. In *ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006*. The Association for Computer Linguistics.

Cronin Blaise. 2001. Acknowledgement trends in the research literature of information science. 57(3):427–433.

Isaac G. Councill, C. Lee Giles, Hui Han, and Eren Manavoglu. 2005. Automatic acknowledgement indexing: Expanding the semantics of contribution in the citeseer digital library. In *Proceedings of the 3rd International Conference on Knowledge Capture*, K-CAP '05, pages 19–26, New York, NY, USA. ACM.

Blaise Cronin, Gail McKenzie, Lourdes Rubio, and Sherrill Weaver-Wozniak. 1993. Accounting for influence: Acknowledgments in contemporary sociology. *Journal of the American Society for Information Science*, 44(7):406–412.

Blaise Cronin, Gail McKenzie, and Michael Stiffler. 1992. Patterns of acknowledgement. *Journal of Documentation*.

S. Dai, Y. Ding, Z. Zhang, W. Zuo, X. Huang, and S. Zhu. 2019. Grantextractor: Accurate grant support information extraction from biomedical fulltext based on bi-lstm-crf. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pages 1–1.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1535–1545.

C. Lee Giles, Kurt D. Bollacker, and Steve Lawrence. 1998. CiteSeer: An automatic citation indexing system. In *Proceedings of the 3rd ACM International Conference on Digital Libraries, June 23-26, 1998, Pittsburgh, PA, USA*, pages 89–98.

C Lee Giles and Isaac G Councill. 2004. Who gets acknowledged: Measuring scientific contributions through automatic acknowledgment indexing. *Proceedings of the National Academy of Sciences*, 101(51):17599–17604.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Subhradeep Kayal, Zubair Afzal, George Tsatsaronis, Marius Doornenbal, Sophia Katrenko, and Michelle Gregory. 2019. A framework to automatically extract funding information from text. In *Machine Learning, Optimization, and Data Science*, pages 317–328, Cham. Springer International Publishing.

Madian Khabsa, Pucktada Treeratpituk, and C. Lee Giles. 2012. Ackseer: a repository and search engine for automatically extracted acknowledgments from digital libraries. In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '12, Washington, DC, USA, June 10-14, 2012*, pages 185–194. ACM.

Mario Lipinski, Kevin Yao, Corinna Breitinger, Joeran Beel, and Bela Gipp. 2013. Evaluation of header metadata extraction approaches and tools for scientific pdf documents. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '13, pages 385–386, New York, NY, USA. ACM.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Dan S. Weld. 2019. S2orc: The semantic scholar open research corpus.

Patrice Lopez. 2009. Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *Proceedings of the 13th European Conference on Research and Advanced Technology for Digital Libraries*, ECDL'09, pages 473–474, Berlin, Heidelberg. Springer-Verlag.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Mausam, Michael Schmitz, Robert Bart, Stephen Soderland, and Oren Etzioni. 2012. Open language learning for information extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12,

pages 523–534, Stroudsburg, PA, USA. Association for Computational Linguistics.

Adèle Paul-Hus, Adrián A. Díaz-Faes, Maxime Sainte-Marie, Nadine Desrochers, Rodrigo Costas, and Vincent Larivière. 2017. Beyond funding: Acknowledgement patterns in biomedical, natural and social sciences. *PLOS ONE*, 12(10):1–14.

Adèle Paul-Hus, Philippe Mongeon, Maxime Sainte-Marie, and Vincent Larivière. 2020. Who are the acknowledgees? an analysis of gender and academic status. *Quantitative Science Studies*, 0(0):1–17.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *CoRR*, abs/2003.07082.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

Peng Shi and Jimmy Lin. 2019. Simple BERT models for relation extraction and semantic role labeling. *CoRR*, abs/1904.05255.

Mayank Singh, Barnopriyo Barua, Priyank Palod, Manvi Garg, Sidhartha Satapathy, Samuel Bushi, Kumar Ayush, Krishna Sai Rohith, Tulasi Gamidi, Pawan Goyal, and Animesh Mukherjee. 2016. OCR++: A robust framework for information extraction from scholarly articles. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 3390–3400. ACL.

E. Smith, B. Williams-Jones, Z. Master, V. Larivière, C. R. Sugimoto, A. Paul-Hus, M. Shi, and D. B. Resnik. 2019. Misconduct and misbehavior related to authorship disagreements in collaborative science. *Sci Eng Ethics*.

Min Song, Keun Young Kang, Tatsawan Timakum, and Xinyuan Zhang. 2020. Examining influential factors for acknowledgements classification using supervised learning. *PLOS ONE*, 15(2):1–21.

H. Stewart, K. Brown, A. M. Dinan, N. Irigoyen, E. J. Snijder, and A. E. Firth. 2018. Transcriptional and translational landscape of equine torovirus. *J Virol*, 92(17).

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, Paul Mooney, Dewey Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: the covid-19 open research dataset. *CoRR*, abs/2004.10706.

Jian Wu, Jason Killian, Huaiyu Yang, Kyle Williams, Sagnik Ray Choudhury, Suppawong Tuarob, Cornelia Caragea, and C. Lee Giles. 2015. Pdfmef: A multi-entity knowledge extraction framework for scholarly documents and semantic search. In *Proceedings of the 8th International Conference on Knowledge Capture*, K-CAP 2015, pages 13:1–13:8, New York, NY, USA. ACM.