Evaluating Reference String Extraction Using Line-Based Conditional Random Fields: A Case Study with German Language Publications

Martin Körner[1](0000-0001-8951-4849), Behnam Ghavimi[2](0000-0002-4627-5371), Philipp Mayr[2](0000-0002-6656-1658), Heinrich Hartmann[3](0000-0002-3929-2421), and Steffen Staab[1](0000-0002-0780-4154)

[1] Institute for Web Science and Technologies, University of Koblenz-Landau, Germany
{mkoerner,staab}@uni-koblenz.de
[2] GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany
{behnam.ghavimi,philipp.mayr}@gesis.org
[3] Independent, Munich, Germany
heinrich@heinrichhartmann.com

Abstract. The extraction of individual reference strings from the reference section of scientific publications is an important step in the citation extraction pipeline. Current approaches divide this task into two steps by first detecting the reference section areas and then grouping the text lines in such areas into reference strings. We propose a classification model that considers every line in a publication as a potential part of a reference string. By applying line-based conditional random fields rather than constructing the graphical model based on individual words, dependencies and patterns that are typical in reference sections provide strong features while the overall complexity of the model is reduced. We evaluated our novel approach RefExt against various state-of-the-art tools (CERMINE, GROBID, and ParsCit) and a gold standard which consists of 100 German language full text publications from the social sciences. The evaluation demonstrates that we are able to outperform state-of-the-art tools which rely on the identification of reference section areas.

Keywords: Reference Extraction · Citations · Conditional Random Fields · German Language Papers

1 Introduction

Citation data shows the link between efforts of individual researchers, topics, and research fields. Despite the widely acknowledged benefits, the open availability of citation data is unsatisfactory. Some commercial companies such as Elsevier and Google do have access to citation data and utilize them to supply their users with effective information retrieval features, recommendation systems, and other

knowledge discovery processes. Yet, the majority of smaller information retrieval systems, such as Sowiport [1] for the social sciences, lack comprehensive citation data.

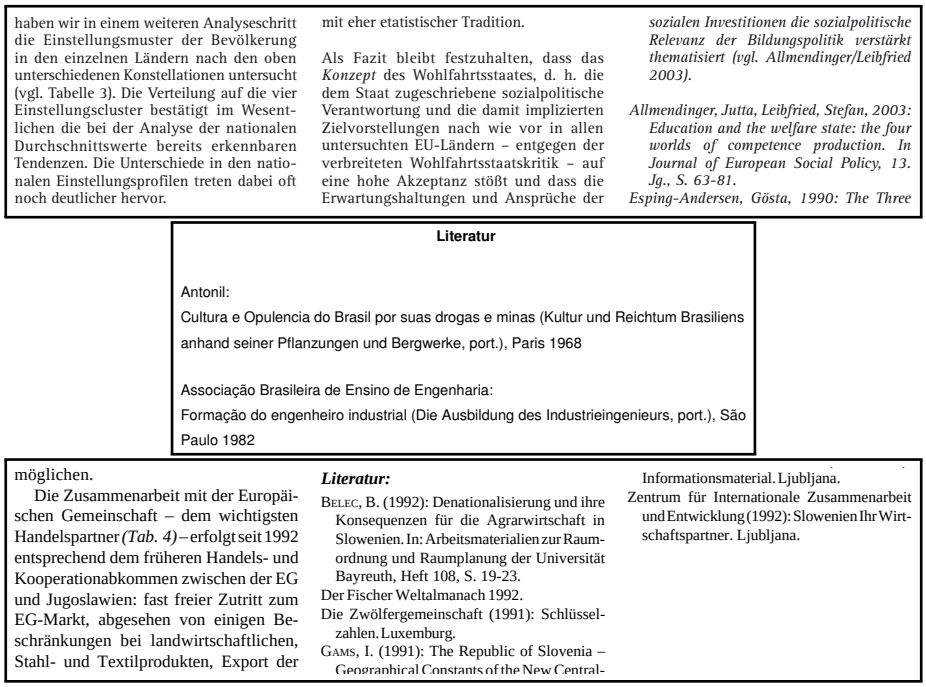
Recent activities like the "OpenCitations Project" or the "Initiative for Open Citations" aim to open up this field and improve the current situation. The "Extraction of Citations from PDF Documents" (EXCITE) project[4] at GESIS and University of Koblenz-Landau is in line with these initiatives and aims to make more citation data available to researchers with a particular focus on the German language social sciences. The shortage of citation data for the international and German social sciences is well known to researchers in the field and has itself often been subject to academic studies [2]. In order to open up citation data in the social sciences, the EXCITE project develops a set of algorithms for the extraction of citation and reference information from PDF documents and the matching of reference strings against bibliographic databases.

In this paper, we will consider the earlier steps in the extraction process that result in individual reference strings. There are several factors that result in the difficulty of the reference extraction task. One such factor is the high number of possible reference styles. According to Zotero[5], there exist more than four hundred difference citation styles in the social sciences alone. Further, there exists a large variety of layouts for publications including different section headings, headers, footers, and varying numbers of text columns. Figure 1 shows three challenging examples where the reference section does not contain a heading, where the reference strings contain a line break after the author names, and where reference strings strongly differ in their length, respectively.

Current solutions that perform reference string extraction have in common that they first identify the reference section and then, in a separate step, segment the reference section into individual reference strings. Thereby, errors that are made during the classification of reference sections directly impact the accuracy of the reference string extraction. For example, if a paragraph that contains reference strings was not recognized as part of the reference section, its reference strings will not be considered in the following step. To prevent this, our approach does not extract reference strings from an area that is first identified as the reference section. Instead, this indicator of a possible reference section is considered as only one of many features in a machine learning model that directly classifies the text lines as reference strings given the full text of the research paper. Other features are based on the text layout and the content of a given text line. A key observation here is that a text line usually does not contain information of more than one reference string. This allows the model to operate not on a word level but on a text line level.

The performance of our approach was evaluated using a novel gold standard for publications in the German language social sciences. To allow for a fair comparison, existing methods were retrained on the same data set that is used in

[4] https://west.uni-koblenz.de/en/research/excite
[5] https://www.zotero.org/styles/



haben wir in einem weiteren Analyseschritt die Einstellungsmuster der Bevölkerung in den einzelnen Ländern nach den oben unterschiedenen Konstellationen untersucht (vgl. Tabelle 3). Die Verteilung auf die vier Einstellungscluster bestätigt im Wesentlichen die bei der Analyse der nationalen Durchschnittswerte bereits erkennbaren Tendenzen. Die Unterschiede in den nationalen Einstellungsprofilen treten dabei oft noch deutlicher hervor.

mit eher etatistischer Tradition.

Als Fazit bleibt festzuhalten, dass das *Konzept* des Wohlfahrtsstaates, d. h. die dem Staat zugeschriebene sozialpolitische Verantwortung und die damit implizierten Zielvorstellungen nach wie vor in allen untersuchten EU-Ländern – entgegen der verbreiteten Wohlfahrtsstaatskritik – auf eine hohe Akzeptanz stößt und dass die Erwartungshaltungen und Ansprüche der

sozialen Investitionen die sozialpolitische Relevanz der Bildungspolitik verstärkt thematisiert (vgl. Allmendinger/Leibfried 2003).

Allmendinger, Jutta, Leibfried, Stefan, 2003: Education and the welfare state: the four worlds of competence production. In Journal of European Social Policy, 13. Jg., S. 63-81.

Esping-Andersen, Gösta, 1990: The Three

Literatur

Antonil:

Cultura e Opulencia do Brasil por suas drogas e minas (Kultur und Reichtum Brasiliens anhand seiner Pflanzungen und Bergwerke, port.), Paris 1968

Associação Brasileira de Ensino de Engenharia:

Formação do engenheiro industrial (Die Ausbildung des Industrieingenieurs, port.), São Paulo 1982

möglichen.

Die Zusammenarbeit mit der Europäischen Gemeinschaft – dem wichtigsten Handelspartner *(Tab. 4)* – erfolgt seit 1992 entsprechend dem früheren Handels- und Kooperationabkommen zwischen der EG und Jugoslawien: fast freier Zutritt zum EG-Markt, abgesehen von einigen Beschränkungen bei landwirtschaftlichen, Stahl- und Textilprodukten, Export der

Literatur:

Belec, B. (1992): Denationalisierung und ihre Konsequenzen für die Agrarwirtschaft in Slowenien. In: Arbeitsmaterialien zur Raumordnung und Raumplanung der Universität Bayreuth, Heft 108, S. 19-23.

Der Fischer Weltalmanach 1992.

Die Zwölfergemeinschaft (1991): Schlüsselzahlen. Luxemburg.

Gams, I. (1991): The Republic of Slovenia – Geographical Constants of the New Central-

Informationsmaterial. Ljubljana.

Zentrum für Internationale Zusammenarbeit und Entwicklung (1992): Slowenien Ihr Wirtschaftspartner. Ljubljana.

Fig. 1. Examples for difficult reference sections. The publications are part of the evaluation dataset and have the SSOAR-IDs[6] 35306, 43525, and 48511, respectively.

our approach. As a result, the evaluation also provides insights into how well existing methods adapt to publications in the German language social sciences.

The remainder of this paper is structured as follows. In Section 2, we present the related work in the area of reference string extraction. Section 3 introduces a novel approach[7] to reference string extraction that does not rely on the detection of reference zones. This approach is evaluated in Section 4 using a new gold standard for reference string extraction in the area of German language social sciences. Section 5 contains a summary and possible future work.

2 Related Work

There exists a considerable amount of literature about the extraction of bibliographic information from the reference section of scientific publications [4–10].

Reviewing this literature shows that there are two categories of approaches. One group concentrates on the reference string segmentation task by assuming

[6] Documents available from `http://www.ssoar.info/ssoar/handle/document/<ID>` by replacing `<ID>` with the corresponding SSOAR-ID.

[7] This approach was described in a preprint by Körner [3].

the reference strings to be given [4–6]. The other group considers the reference string extraction from an article in the PDF or text format [7–10]. Further, all reference string extraction approaches follow two common steps.

The first step identifies the text areas of the publication that contain the reference strings. Councill, Giles, and Kan [8] as well as Wu et al. [9] use a set of regular expressions to locate the beginning and end of reference sections. Tkaczyk et al. [10] apply a layout analysis on publications given as PDF files which results in textual areas that are grouped into zones. These zones are then classified as "metadata", "body", "references", or "other" using a trained Support Vector Machines (SVMs) model [10]. Lopez [7] trains a conditional random field (CRF) [11] model that performs a segmentation of textual areas into zones similar to Tkaczyk et al. [10].

In a second step, the lines in the identified areas are grouped into individual reference strings. Councill, Giles, and Kan [8] as well as Wu et al. [9] apply regular expressions to detect possible markers of reference strings such as numbers or identifiers surrounded by brackets. If such markers are found, the lines are grouped accordingly. If no markers are found, the lines are grouped based on the line length, ending punctuation, and strings that appear to be author name lists [8]. Tkaczyk et al. [10] use the k-means learning algorithm to perform a clustering into two groups: The first lines of reference strings and all other lines. The features for this clustering include layout information such as the distance to the previous line and textual information such as a line ending with a period [10]. As with the reference area detection, Lopez [7] learn a CRF model for this task. This model uses an input format that is different from the one that is used for their first CRF model. Tokens are split at white spaces and for each token, a list of features is created. Such features include layout information such as the font size and font weight as well as textual features such as the capitalization of the token and whether the token resembles a year, location, or name.

3 Approach

As previously discussed, a typical problem of existing reference string extraction approaches is a wrong classification of textual areas during the first step (see Section 2). Another key insight is that reference strings commonly start in a new line. This is used in the previously described k-means clustering algorithm by Tkaczyk et al. [10] and provides potential advantages over a word-based approach. For example, a line-based model drastically reduces the number of assigned target variables while still allowing the expression of relevant features. Further, it can capture patterns that repeat every few lines more naturally than a word-based model which focuses on a more local context.

These two insights are leveraged by applying a line-based classification model on the whole publication. For this, a possible set of labels consists of `B-REF`, `I-REF`, and `O` where `B-REF` denotes the first line of a reference string, `I-REF` a line of a reference string which is not the first line, and `O` any other line. This is based on the Beginning-Intermediate-Other (BIO) notation [12]. For our

evaluation, we assigned one of the three labels to every text line in a publication. Having such a labeling, it is then possible to automatically extract the reference strings by concatenating a line labeled with `B-REF` together with the following lines labeled with `I-REF` until reachting a line that is labeled with `B-REF` or `O`.

Our model RefExt[8] uses both textual and layout features. In our evaluation we used textual features that signalize whether a line only consists of a number, starts with a capitalized letter, ends with a number, ends with a period, ends with a comma, or contains a year, a year surrounded by braces, a page range, an ampersand character, a quotations mark, a colon, a slash, or opening and closing braces. Another type of textual features counts the occurrences of numbers, words, periods, commas, and words that only consist of one capitalized letter. Further, we used four layout features. One signalizes whether the current line is indented when compared to the previous line. Another detects a gap between the current and previous line that is larger than a predefined value. The third layout feature is assigned to a line that contains less characters than the previous one. The last layout feature signalizes the position of a given line in the whole document. For this, the current line number is divided by the total number of lines. A more detailed description of the used features, together with all evaluation results, is provided on GitHub[9].

One advantage of using CRFs is that features do not have to be independent from each other due to the modeled conditional probability distribution [13]. Another advantage is the possibility to include contextual information. To do so, a CRF model with a high Markov order can be applied. This is feasible due to the line-based approach and the resulting lower number of random variables in the model when compared to a word-based approach.

4 Evaluation

The gold standard that is used in the following evaluation is based on 100 German publications from the SSOAR repository[10] in the PDF format. Since this evaluation focuses on the reference string extraction, documents that consist of scanned pages or that do not contain a reference section were excluded beforehand from the otherwise random selection. The resulting papers contain an average of 54 reference strings with a total number of 5,355 reference strings. Figure 2 gives an overview of the publication types and publication years of the gold standard documents. Resulting from the fact that existing reference string extraction tools use different input formats for their training procedures and also show differences in text-encoding of the resulting reference strings, a number of annotation file formats were created and manually inspected. This resulting gold standard is available on GitHub[11].

[8] `https://github.com/exciteproject/refext`
[9] `https://github.com/exciteproject/amsd2017`
[10] `http://www.ssoar.info/`
[11] `https://github.com/exciteproject/ssoar-gold-standard`

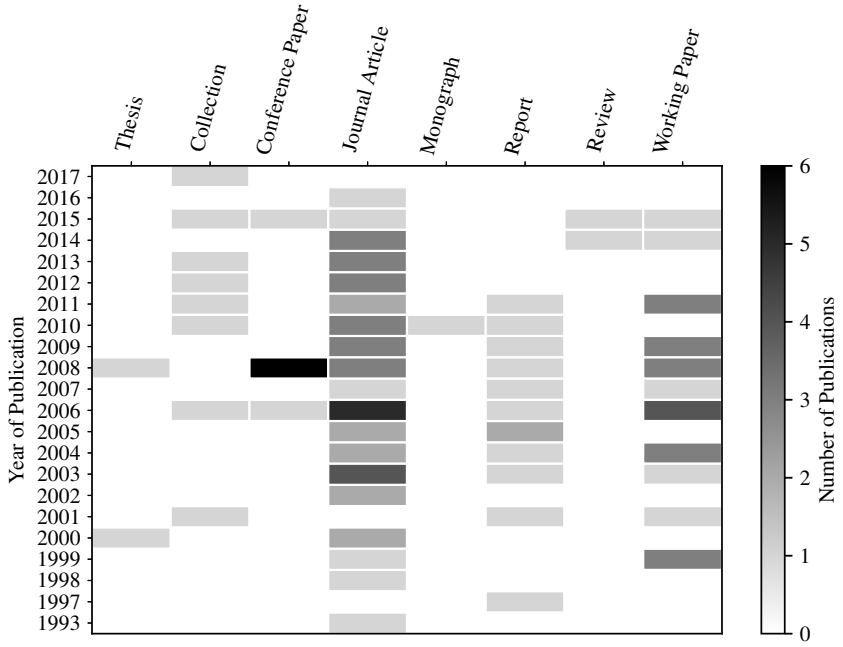


Fig. 2. Distribution of papers in the gold standard based on the publication type and year of publication.

Since most existing citation information extraction tools focus on English language publications, the possibility to adapt the tool to German language publications is crucial. Two tools that allow such retraining are CERMINE [10] and GROBID [7]. For ParsCit [8], an older version allows the adaption of the regular expressions that detect reference section headings and other relevant headings such as appendices. Other tools that do not allow a retraining, such as PDFX [14] and pdfextract[12], were excluded due to their low performance on German language publications. The evaluation considers the performance on a line level based on the BIO notation in Table 1 and on a reference string level in Table 2 using the metrics macro precision, macro recall, and macro F1-score. To compute macro metrics, a metric is first calculated on the individual publications and then averaged over all publications. Thereby, publications with a large amount of reference strings do not have a bigger impact on the final metric than publications with only a few reference strings. Further, in order to reduce the influence of the chosen split into training and testing data, the evaluation was performed using 10-fold cross-validation where each fold contains ten randomly chosen papers of the gold standard for testing and the remaining ninety papers for training. In the result tables, CERMINE, GROBID, and ParsCit are abbreviated with CER, GRO, and Pars, respectively. The suffixes D and T sig-

[12] `https://www.crossref.org/labs/pdfextract`

Table 1. Macro-metrics of BIO-annotated reference lines using 10-fold cross-validation on 100 German social science publications.

Metric	CER-D	CER-T	Pars-D	Pars-M	GRO-D	GRO-T	RefExt-T
`B-REF` Precision	0.719	0.734	0.683	0.769	0.692	0.871	**0.916**
`B-REF` Recall	0.600	0.557	0.620	0.688	0.789	0.865	**0.952**
`B-REF` F1-Score	0.616	0.589	0.616	0.689	0.712	0.861	**0.922**
`I-REF` Precision	0.729	0.755	0.577	0.678	0.664	0.857	**0.882**
`I-REF` Recall	0.340	0.313	0.809	0.843	0.839	0.871	**0.944**
`I-REF` F1-Score	0.432	0.415	0.647	0.716	0.703	0.855	**0.902**

Table 2. Macro-metrics of reference string extraction using 10-fold cross-validation on 100 German social science publications.

Metric	CER-D	CER-T	Pars-D	Pars-M	GRO-D	GRO-T	RefExt-T
Precision	0.296	0.303	0.558	0.617	0.627	0.847	**0.879**
Recall	0.233	0.220	0.552	0.595	0.718	0.839	**0.906**
F1-Score	0.245	0.235	0.542	0.590	0.650	0.837	**0.885**

nalize whether the tool was using its default model or a model that was trained on the ninety publications from the gold standard. Pars-M represents version 101101 of ParsCit with modified regular expressions that match against German language section headings such as "Literatur" and "Anhang". Pars-D uses the latest ParsCit version as of May 31, 2017 which instead uses a trained model for detecting reference sections. We were not able to retrain this model with the given source code and documentation. Further, we used CERMINE version 1.13 and GROBID version 0.4.1. Our approach, RefExt version 0.1.0, is based on the CRF models of MALLET [15] as well as the PDF text extraction and reading order detection of CERMINE. For the finite state transducer of MALLET we applied states in the "three-quarter" order. Thereby, the network contains a weight between every pair of adjacent labels as well as between the label and its corresponding features. Further, we applied a set of conjunctions that allow the usage of features of the previous two and following two lines. We found this network structure to perform similar to more complex structures while providing a reduced training time and a lower risk of overfitting. The learning was performed using the label log-likelihood with L1-regularization with a weight of 20.

The results show that RefExt is able to outperform the other tools on our gold standard. Over the 100 documents divided into ten folds, there were two publications[13] for which RefExt had a recall of zero in terms of reference strings. In both cases, the year number appeared at the end of the reference strings which is uncommon for the training corpus. In addition, one of the reference

[13] `http://www.ssoar.info/ssoar/handle/document/32521` and `http://www.ssoar.info/ssoar/handle/document/43525`

styles includes a line break after the listed authors in a reference string[14] which is also unusual. GROBID had a recall of zero in seven publications in terms of reference strings. Interestingly, the two publications that were problematic for RefExt had a recall of 1.0 and 0.662 in GROBID, respectively. Thereby, a combined approach might be worthwhile.

5 Summary and Future Work

We have presented a novel approach to reference string extraction using line-based CRFs. The evaluation demonstrated that this approach outperforms existing tools when trained on the same amount of annotated data it the area of German language social sciences. Yet, there are several aspects that require further efforts. Having a precision and recall of around 0.9 is not sufficient for the usage in a productive system and it remains to be evaluated how the performance improves when extending the training data. Improvements might also be possible by adding more domain-specific features. Examples for such features are last name dictionaries or words that commonly appear in German language reference strings such as "Hrsg." and "Zeitschrift". Further, a number of journals in the German social sciences such as "Totalitarismus und Demokratie"[15] and "Südosteuropäische Hefte"[16] use a citation style where references are not grouped in a separate reference section but instead appear in the footnotes. This could present an interesting use case of our approach.

Acknowledgements. This work has been funded by Deutsche Forschungsgemeinschaft (DFG) as part of the project "Extraction of Citations from PDF Documents (EXCITE)" under grant numbers MA 3964/8-1 and STA 572/14-1. We would like to thank Dominika Tkaczyk for her support regarding the CERMINE tool as well as Alexandra Bormann, Jan Hübner, and Daniel Kostić for contributing to the gold standard that was used in this research.

References

1. Hienert, D., Sawitzki, F., Mayr, P.: Digital Library Research in Action – Supporting Information Retrieval in Sowiport. D-Lib Magazine **21**(3/4) (2015)
2. Moed, H.F.: Citation analysis in research evaluation. Volume 9. (2005)
3. Körner, M.: Reference String Extraction Using Line-Based Conditional Random Fields. ArXiv e-prints (2017)
4. Peng, F., McCallum, A.: Information extraction from research papers using conditional random fields. Information Processing & Management **42**(4) (2006) 963–979
5. Cortez, E., da Silva, A.S., Gonçalves, M.A., Mesquita, F., de Moura, E.S.: Flux-cim: flexible unsupervised extraction of citation metadata. In: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries, ACM (2007) 215–224

[14] Shown as the first example in Figure 1.

[15] `http://www.hait.tu-dresden.de/td/home.asp`

[16] `http://suedosteuropaeische-hefte.org/`

6. Groza, T., Grimnes, G.A., Handschuh, S.: Reference information extraction and processing using conditional random fields. Information Technology and Libraries (Online) **31**(2) (2012) 6
7. Lopez, P.: Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In: International Conference on Theory and Practice of Digital Libraries, Springer (2009) 473–474
8. Councill, I.G., Giles, C.L., Kan, M.Y.: ParsCit: An open-source CRF Reference String Parsing Package. In: Proceedings of LREC. Volume 2008. (2008) 661–667
9. Wu, J., Williams, K., Chen, H.H., Khabsa, M., Caragea, C., Ororbia, A., Jordan, D., Giles, C.L.: Citeseerx: Ai in a digital library search engine. In: AAAI. (2014) 2930–2937
10. Tkaczyk, D., Szostek, P., Fedoryszak, M., Dendek, P.J., Bolikowski, Ł.: Cermine: automatic extraction of structured metadata from scientific literature. International Journal on Document Analysis and Recognition (IJDAR) **18**(4) (2015) 317–335
11. Lafferty, J., McCallum, A., Pereira, F., et al.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the eighteenth international conference on machine learning, ICML. Volume 1. (2001) 282–289
12. Houngbo, H., Mercer, R.E.: Method Mention Extraction from Scientific Research Papers. In: COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, 8-15 December 2012, Mumbai, India. (2012) 1211–1222
13. Koller, D., Friedman, N.: Probabilistic graphical models: principles and techniques. MIT press (2009)
14. Constantin, A., Pettifer, S., Voronkov, A.: Pdfx: fully-automated pdf-to-xml conversion of scientific literature. In: Proceedings of the 2013 ACM symposium on Document engineering, ACM (2013) 177–180
15. McCallum, A.K.: Mallet: A machine learning for language toolkit. (2002)