Extracting bibliographic references from footnotes with EXcite-docker

Christian Boulanger[1,*], Anastasiia Iurshina[2]

[1]Max Planck Institute for Legal History and Legal Theory, Frankfurt a.M.
[2]University of Stuttgart

Abstract

The paper presents a project that aims at providing a user-friendly way for the domain-specific extraction and segmentation of references from PDF documents containing scholarship from the humanities and social sciences. The software builds on code developed by the EXcite project, adding a server and improved web interface for producing gold standard with which to train the extraction and segmentation models. The paper notes that the model trained with EXcite's gold standard, similarly to comparable software, is optimized for documents in which bibliographic references are given in a bibliography section at the end of the document. The results are much worse in the case of documents where the full or partial references are in footnotes. Searching for ways of improving the performance, we compare the accuracy of a model trained with a small set of annotated documents with references in footnotes with that of the default EXcite model and that of a model trained with a combined dataset. Preliminary results suggest that a specialized footnote model provides better accuracy as compared to a model trained with a combined dataset. We conclude with the roadmap to further improve the accuracy of the model.

Keywords

Digital Humanities, Scholarly literature, Citations, Footnotes, Reference mining, Reference extraction, Reference segmentation

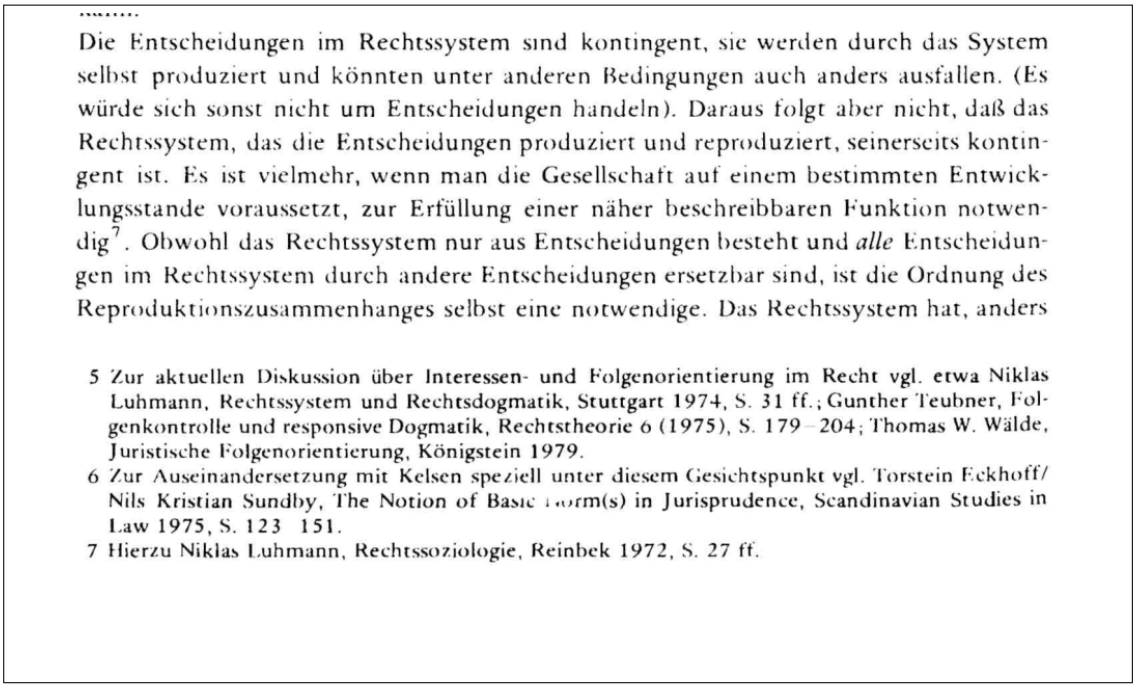
1. Introduction

Current Open Source software for metadata mining, in particular for reference extraction, specializes in documents in which bibliographic references are listed in a bibliography section at the end of the document. However, in legal scholarship, the humanities and parts of the social sciences, the cited literature is referenced in footnotes, which contain the full (or sometime only fragmentary) information on the source of the citation (see Fig. 1).

As a result, the accuracy of reference extraction of the available tools is very low.[1] In this paper, we describe ongoing work on a tool ("excite-docker") that has been built on existing work in the EXcite project and is being used for reference mining in (socio-)legal studies. We compare the accuracy of a reference extraction model trained with a small set of annotated

*Corresponding author.
boulanger@lhlt.mpg.de (C. Boulanger); anastasiia.iurshina@ipvs.uni-stuttgart.de (A. Iurshina)
https://lhlt.mpg.de/boulanger (C. Boulanger)
0000-0001-6928-3246 (C. Boulanger); 0000-0002-1231-2314 (A. Iurshina)

[1]This can be seen when trying to extract references from a paper with references in footnotes using the webservices of GROBID https://cloud.science-miner.com/grobid/ and EXcite https://excite.informatik.uni-stuttgart.de/excite.



Die Entscheidungen im Rechtssystem sind kontingent, sie werden durch das System selbst produziert und könnten unter anderen Bedingungen auch anders ausfallen. (Es würde sich sonst nicht um Entscheidungen handeln). Daraus folgt aber nicht, daß das Rechtssystem, das die Entscheidungen produziert und reproduziert, seinerseits kontingent ist. Es ist vielmehr, wenn man die Gesellschaft auf einem bestimmten Entwicklungsstande voraussetzt, zur Erfüllung einer näher beschreibbaren Funktion notwendig[7]. Obwohl das Rechtssystem nur aus Entscheidungen besteht und *alle* Entscheidungen im Rechtssystem durch andere Entscheidungen ersetzbar sind, ist die Ordnung des Reproduktionszusammenhanges selbst eine notwendige. Das Rechtssystem hat, anders

5 Zur aktuellen Diskussion über Interessen- und Folgenorientierung im Recht vgl. etwa Niklas Luhmann, Rechtssystem und Rechtsdogmatik, Stuttgart 1974, S. 31 ff.; Gunther Teubner, Folgenkontrolle und responsive Dogmatik, Rechtstheorie 6 (1975), S. 179–204; Thomas W. Wälde, Juristische Folgenorientierung, Königstein 1979.

6 Zur Auseinandersetzung mit Kelsen speziell unter diesem Gesichtspunkt vgl. Torstein Eckhoff/Nils Kristian Sundby, The Notion of Basic Norm(s) in Jurisprudence, Scandinavian Studies in Law 1975, S. 123 151.

7 Hierzu Niklas Luhmann, Rechtssoziologie, Reinbek 1972, S. 27 ff.

Figure 1: A typical example of bibliographic references in a footnote

documents with references in footnotes with that of the default EXcite model, and that of a model trained with a combined dataset. Preliminary results suggest that a specialized footnote model provides better accuracy as compared to a model trained with a combined dataset of footnote **and** bibliography data. We conclude with the roadmap to further improve the accuracy of the model.

2. Related Work

2.1. Software for extraction of citation data

For a long time, technologies for extraction of bibliographic metadata from scholarly articles, including citations, have been the domain of commercial services using closed-source technologies.[2]. For a couple of years now, Free and Open Source Software (FOSS) projects have emerged that develop these technologies and allow their use unencumbered by license fees or usage restrictions. Projects going back into the late 2000s and early 2010s, such as GROBID[1] or CERMINE[2], have concentrated on English language papers in the natural sciences, which have - in all their variety - relatively similar document structures and citation patterns.

Since the focus on this kind of literature yielded sub-optimal results for the extraction of citations from German language social science literature, the EXcite project (https://excite.

[2]The most important data source, in particular for bibliometric analyses, has been the Web of Science (https://webofscience.com), which is extremely expensive and restrictive in how its data can be used. Other examples are Scopus, Dimensions or the discontinued Microsoft Academic Graph

informatik.uni-stuttgart.de) has developed a set of algorithms for information extraction and matching.[3] The results have been promising for German social science literature.[4, 5]. There are, however, two areas in which the code as released in 2019 (see https://github.com/exciteproject/) are in need of improvement from the perspective of this paper. First of all, the corpus of documents that has been used to train EXcite's default model contains only very few documents with references in the footnotes (see https://github.com/exciteproject/EXgoldstandard. Not surprisingly, this results in an extremely low accuracy of reference identification. In fact, the number of false positives far outweighs the number of correctly identified citations. On the other hand, the tools in their published form were not readily usable without intimate knowledge of the code.

2.2. Use case: the Legal Theory Graph Project

These challenges became apparent when one of the authors was looking for reference extraction software to use in a Digital Humanities project at the Max Planck Institut for Legal History and Legal Theory. The "Legal Theory Graph Project" aims at producing machine-readable data on legal theory scholarship since 1945 (with a focus on socio-legal theory), mapping scholars, institutions, publications and citations in a network graph.[3] Data for this graph is collected, among other things, by harvesting metadata and using text and reference mining techniques.

In accordance with the stated goals of the Max Planck Society,[4], the project adheres to an Open Source and Open Access approach, which means that both the data produced as well as the technologies used should be freely available. This rules out to rely on proprietary data sources and technologies. In any case, the coverage of scholarship in law, the social sciences and the humanities is very limited in the case of both the commercial services and the newly emerging Open Access sources of bibliographic and bibliometric data[6, 7]. This is particularly true for publications which are not journal articles or which are not identified by a Digital Object Identifier (DOI).

In law, the social sciences and the humanities, however, a large part of research is published as books, as chapters in edited volumes or as articles in journals which do not assign DOIs (yet). Most importantly, all of the bibliographic and bibliometric data sources mainly contain data on English language scholarship of the last 20-30 years, whereas the Legal Theory Graph Project has a focus on German scholarship and includes a much larger time period. Some of this data is covered by Google Scholar[6, 7], but Google Scholar has no API and actively prevents data mining. For these reasons, the data had to be self-produced using freely available technologies. EXcite seemed to be the best candidate for the reference extraction part of the project's technical workflow (see Fig. 2).

2.3. Collaborative project: excite-docker

The authors collaborated to achieve the following tasks:

- working on EXcite's codebase to make it more readily usable for non-technical users,

[3]See https://www.lhlt.mpg.de/2514927/03-boulanger-legal-theory-graph?c=2367449
[4]See https://openaccess.mpg.de/Berlin-Declaration

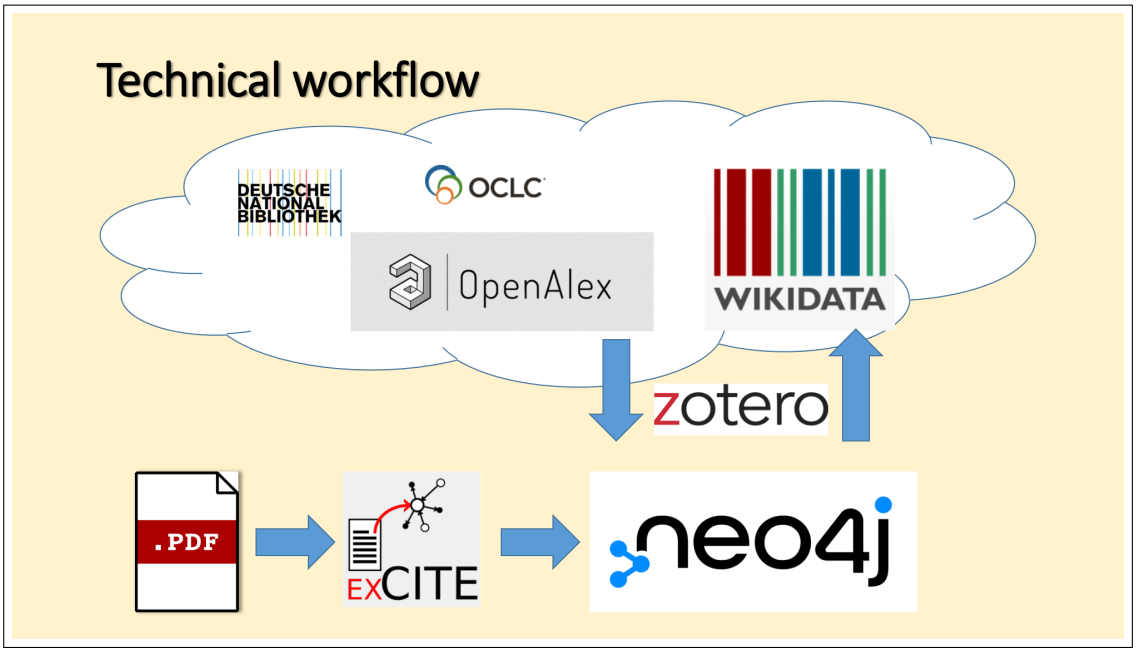


Figure 2: The workflow of the Legal Theory Graph Project

including for support staff producing annotations that can serve as training data for the reference extraction models;

- improving the algorithm's accuracy for the intended target literature
- adding an evaluation workflow, to be able to document improvements in accuracy in a reproducible way

The result is the GitHub repository https://github.com/cboulanger/excite-docker[5]. The code in the repository builds a Docker image which can be run on any platform that support Docker. The docker container provides a web app (see figure 3) that can be used to produce annotations for the learning algorithm and a command line interface to run the commands for training of the machine learning models, citation data extraction and accuracy evaluation.

The web application has been used to annotate text extracted from PDFs from the German Journal of Law and Society (1980-) (https://zfrsoz.info), which will be part of the Legal Theory Graph. Many papers published in this journal from the 1980 use footnote citations. In addition, the citation style is very inconsistent. This data is ideal for stress-testing extraction models, in fact, initial extractions with the default excite model resulted in large amounts of unusable results.

[5]The original code has been forked from https://git.gesis.org/hosseiam/excite-docker but has been rewritten in large parts. The main ML algorithms are largely unchanged and have been written by the EXcite team members. All improvements to the extraction algorithms and the evaluation algorithms are by A. Iurshina, who has also ported the code to Python 3. C. Boulanger has rewritten the web application for document annotation and has added the CLI.

Table 1
Average accuracy for different configurations

Configuration	Extraction Acc	Segmentation Acc
Default	0.24	0.37
Footnotes only	**0.26**	0.37
Combined	0.22	**0.47**

3. Data

As training data, we used EXcite[5] gold dataset[6] as well as manually annotated papers from the German Journal of Law and Society. EXcite's dataset contains 125 annotated articles in German language (2652 reference strings) and 100 articles in English (2838 reference strings in different languages). A small fraction of these references is in footnotes, most, however, are in the reference section.

The dataset with the German Journal of Law and Society papers contains 20 documents (970 reference strings).

For evaluation, only socio-legal papers with references in footnotes were used because they represent the examples of the data we want to see improvements on. Since the dataset contains only 20 papers, we used 5 papers as the test set (454 reference strings).

4. Results

EXParser[3] was trained and evaluated in three configurations:

- Trained on EXcite gold and tested on the test split of socio-legal papers (default)
- Trained on the training split of socio-legal papers and tested on the test split of socio-legal papers (footnotes only)
- Trained on EXcite gold combined with the training split of socio-legal papers and tested on the test split of socio-legal papers (combined)

Table 1 presents evaluation results for three configurations. Accuracy[7] was calculated as follows: for extraction, we found the longest common sequence between each extracted reference string and the ground truth file and divided the length of this sequence by the length of the extracted line. For segmentation, we evaluated for each reference string, the proportion of correctly classified tokens (title, author, etc).

[6] https://github.com/exciteproject/EXgoldstandard

[7] We relied on accuracy as the evaluation metric, not on F1-score, as F1-score is not very suitable for evaluation of reference extraction. Besides, our goal was not compare with EXparser but to see if there is an improvement with adding more data.

5. Conclusion and Future Work

As Tkaczyk et al. argue, "tuning [extraction] models to the task-specific data results in the increase in the quality" [8]. In the case of reference-in-footnotes vs. references-in-bibliography, this is to be expected, since the structure of the pages and the text as a whole is very different. The results of this initial test makes us skeptical whether training the extraction model with a large diversity of documents will improve the accuracy of the extraction. Instead, it suggests that models that are more fine-tuned to the type of the document (footnote vs. bibliography) are the way to go, and that an extraction workflow should either try different models for the best outcome or - since this is time-intensive- use heuristics to select the model that suits the type of the paper. The data allows this conclusion only for reference *extraction*. Concerning reference *segmentation*, the combined model did better. The difference can be explained by the fact that even though citation styles vary widely[8], structurally the differences are much smaller compared to the difference between references-in-footnotes vs. references-in-bibliography.

In any case, a lot of work is still ahead. We have far too few annotated reference-in-footnotes documents needed to improve the accuracy of the specialized footnotes model. Preparing such documents is a time-consuming task and would profit from the collaboration with other projects that want to extract citation data from similar type of scholarly literature. Alternatively, synthetic training data could be produced from existing citation data on journals that use the references-in-footnotes style, although such data is yet to be located. We are also looking into ways of increasing the accuracy using word embeddings, part-of-speech-tagging or dictionary-based approaches that would help to identify text that is expected to appear in references in the target literature and to further increase the accuracy of segmentation. Another direction to explore, given that we will have enough data, is replacing manual feature extraction, on which EXParser relies heavily, with deep-leaning-based models or with a combination of manually and automatically-extracted features.

References

[1] P. Lopez, GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications, in: M. Agosti, J. Borbinha, S. Kapidakis, C. Papatheodorou, G. Tsakonas (Eds.), Research and Advanced Technology for Digital Libraries, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2009, pp. 473–474. doi:10.1007/978-3-642-04346-8_62.

[2] D. Tkaczyk, P. Szostek, M. Fedoryszak, P. J. Dendek, L. Bolikowski, CERMINE: automatic extraction of structured metadata from scientific literature, International Journal on Document Analysis and Recognition (IJDAR) 18 (2015) 317–335. URL: https://doi.org/10.1007/s10032-015-0249-8. doi:10.1007/s10032-015-0249-8.

[3] A. Hosseini, B. Ghavimi, Z. Boukhers, P. Mayr, Excite – a toolchain to extract, match and publish open literature references, 2019. doi:10.1109/JCDL.2019.00105.

[4] B. Ghavimi, W. Otto, P. Mayr, An Evaluation of the Effect of Reference Strings and Segmentation on Citation Matching, in: A. Doucet, A. Isaac, K. Golub, T. Aalberg, A. Jatowt

[8] See, for example https://citationstyles.org/ for an open source attempt to make this variety manageable for software

(Eds.), Digital Libraries for Open Knowledge, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2019, pp. 365–369. doi:10.1007/978-3-030-30760-8_35.

[5] Z. Boukhers, S. Ambhore, S. Staab, An end-to-end approach for extracting and segmenting high-variance references from pdf documents, 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL) (2019) 186–195.

[6] A.-W. Harzing, S. Alakangas, Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison, Scientometrics 106 (2016) 787–804. URL: https://doi.org/10.1007/s11192-015-1798-9. doi:10.1007/s11192-015-1798-9.

[7] A. Martín-Martín, M. Thelwall, E. Orduna-Malea, E. Delgado López-Cózar, Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations, Scientometrics 126 (2021) 871–906. URL: https://doi.org/10.1007/s11192-020-03690-4. doi:10.1007/s11192-020-03690-4.

[8] D. Tkaczyk, A. Collins, P. Sheridan, J. Beel, Machine Learning vs. Rules and Out-of-the-Box vs. Retrained: An Evaluation of Open-Source Bibliographic Reference and Citation Parsers, in: Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, JCDL '18, Association for Computing Machinery, New York, NY, USA, 2018, pp. 99–108. doi:10.1145/3197026.3197048.

Figure 3: Web application for the annotation of documents