EXCITE – A toolchain to extract, match and publish open literature references

Azam Hosseini
GESIS – Leibniz Institute for the Social Sciences
azam.hosseini@gesis.org

Behnam Ghavimi
GESIS – Leibniz Institute for the Social Sciences
behnam.ghavimi@gesis.org

Zeyd Boukhers
University of Koblenz-Landau
boukhers@uni-koblenz.de

Philipp Mayr
GESIS – Leibniz Institute for the Social Sciences
philipp.mayr@gesis.org

ABSTRACT

This demo paper presents a generic toolchain to extract, segment and match literature references from full text PDF files in the project EXCITE. The aim of EXCITE is extracting and matching citations from social science publications and making more citation data available to researchers. Each single step in the EXCITE pipeline and the open source tools used to accomplish the tasks are explained. The public demo system which integrates all components of the toolchain under an user-friendly interface is put forward and illustrated. As a final step, a special component is introduced which is capable to ingest the extracted and matched references into the Open Citation Corpus.

KEYWORDS

Reference Extraction, Reference Matching, Open Citations, Demo

1 INTRODUCTION

Despite the widely acknowledged benefits of citation data, the open access to references/citations is still insufficient. Some commercial companies such as Clarivate Analytics, Elsevier or Google possess citation data in large-scale and use them to provide services for their users. On the other side, 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 [5]. The accessibility of information in the social sciences lags behind other fields (e.g. the natural sciences) where more citation data is available.

Recently, some initiatives and projects e.g. the "Open Citations" project or the "Initiative for Open Citations" focus on publishing citation data openly[1]. The "Extraction of Citations from PDF Documents" - EXCITE[2] project is one of these projects. The aim of EXCITE is extracting and matching citations from social science publications [4] and making more citation data available to researchers. EXCITE is focusing on social science publications in German language but is introducing a generic toolchain which can be used and trained for any domain. All tools in the EXCITE project are made available to other researchers. This demo paper introduces the EXCITE toolchain.

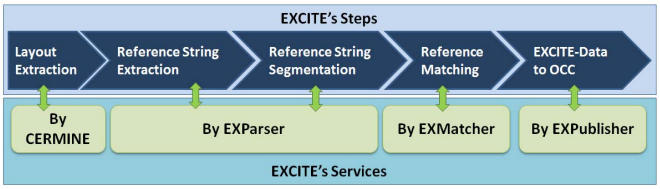

Figure 1: An overview of processing steps and tools in the project EXCITE

2 EXCITE TOOLCHAIN

A number of algorithms are developed in the EXCITE project for extracting references from PDF full texts and matching them against bibliographic databases (see overview in Figure 1). The extraction of references is implemented as a four-steps process:

(1) Extraction of text from PDF files by CERMINE[3],
(2) Identification of reference strings and segmentation of references into its constituent fields such as author, title, etc. by Exparser [1][4],
(3) Matching of references against bibliographic databases by EXmatcher [2][5],
(4) Export and publication of references to reusable formats by conversion of the generated reference information to the json format with OCC ontology [6].

For the matching task in EXCITE, different target databases are utilized: a) sowiport [3], b) GESIS Search[6] and c) Crossref[7]. The EXCITE corpus (PDF files to be processed in the EXCITE project) contains SSOAR[8] documents (approx. 35k), Springer Online Journals collection (approx. 80k), and sowiport full text papers (approx. 116K). The extracted citation data from the EXCITE corpus will be integrated into GESIS Search and OCC (OpenCitations[9] Corpus). EXCITE toolchain is not depended to any citation style or

[1] https://i4oc.org/
[2] http://excite.west.uni-koblenz.de/website/

[3] https://github.com/CeON/CERMINE
[4] https://github.com/exciteproject/Exparser
[5] https://github.com/exciteproject/EXmatcher
[6] https://search.gesis.org/
[7] https://search.crossref.org
[8] https://www.ssoar.info
[9] http://opencitations.net

978-1-7281-1547-4/19/$31.00 ©2019 IEEE
DOI 10.1109/JCDL.2019.00105

language but the current system is trained by using the manually assessed EXCITE gold standard [10] (including German and English languages). The Excite toolchain code are openly available and can be adapted to new domains and languages easily.

3 DEMO SYSTEM

The EXCITE demo system *http://excite.west.uni-koblenz.de/excite* is a web interface which is deployed to integrate different parts of the EXCITE toolchain. The used web framework is Flask which integrates Python modules within web functionality such as RESTful web service. For delegating long lasting tasks, Celery is used in the EXCITE web architecture. Celery is a task queue based on distributed message passing. It enables systems to process batch jobs in the way that each defined worker performs a task in the queue and when the task is done next one will be picked. Codes in other programming languages (e.g., CERMINE in JAVA) with standard I/O format can be easily executed by a Python module. Therefore, they can be integrated in a toolchain by inserting their related tasks in Celery queue. With Celery different modules of the EXCITE toolchain can be applied on a bunch of PDF files asynchronously. There are two main functions in the demo. First: uploading single PDF files; second: running the EXCITE toolchain and checking the generated results (see Figure 2).

1 Uploading files The first step is uploading files to the server. A unique random code will be generated as soon as a user submits a file. The code will be displayed on the demo page. The code also can be sent via email (if the user entered the email in a form). This code is necessary for tracking the results of the toolchain for the submitted file. Users can check the result on the demo page by follow up code. The result will be shown in separated tabs.

2 Running EXCITE toolchain After uploading a file, the "EXCITE toolchain" will be started automatically. There are three main steps in this process: *Layout extraction:* Extracting the layout from a PDF will be started by calling a Java module base on CERMINE. The output of this step will be a "Layout file" which contains text content of each PDF file and it's related layout information such as weight and height of each line. *Reference and Segment Extraction:* In this step Exparser will be called for extracting references from the layout file. Exparser is a python code based on CRF algorithm and does reference strings extraction and segmentation in one step to reduce error rate. The output will be provided in these different formats: plain text, xml and BibTex format. *Reference Matching:* In this step EXmatcher will be called for matching references against corresponding items in the defined target bibliographical databases. The input of EXmatcher is reference strings and segments generated in the previous step. The output will be matched document ids and the probability for each match. This algorithm is build based on the combination of a blocking technique (SOLR is used for indexing) and a SVM classifier. Figure 2 demonstrates the EXCITE demo.

4 OUTLOOK

In the remaining project time of EXCITE we will provide a Docker file containing the EXCITE toolchain to make the tools more efficiently and conveniently re-usable. The EXCITE toolchain will also be accessible as a web service API to allow third-parties to extract citation data from arbitrary publications. The EXpublisher module[11] makes sure that extracted references are enriched with the information of matched items. Afterwards, the information is converted to json format with OCC ontology and ingested into the Open Citation Corpus. OpenCitations makes this data available for users by providing dump data and also a SPARQL endpoint. Each entity (e.g., responsible agents, and reference strings) has a unique OCC identifier. For example, 'be/01101'[12] is an identifier for a reference string in OCC. The identifier contains two parts which are connected with a slash symbol. The first part, 'be' defines the type of data which is a bibliographic entity (i.e., reference string). The second part (i.e., digit part - 01101) is the main identifier. All entities in OCC with main identifier starting with '0110' are generated by the EXCITE project.

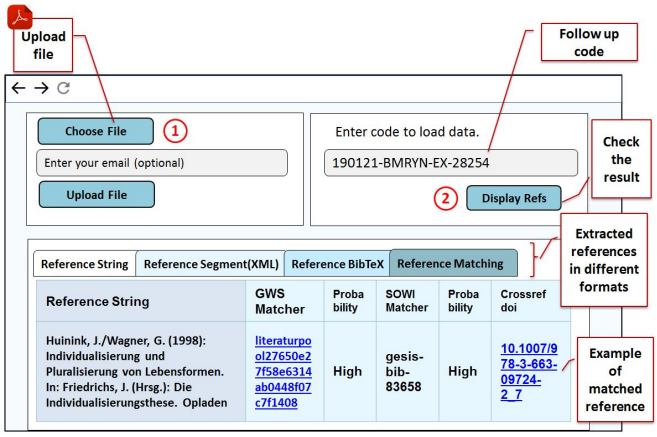

Figure 2: EXCITE Demo system

ACKNOWLEDGMENTS

This work has been funded by Deutsche Forschungsgemeinschaft (DFG) under grant numbers MA 3964/8-1 and STA 572/14-1.

REFERENCES

[1] Zeyd Boukhers, Shriharsh Ambhore, and Steffen Staab. 2019. An End-to-end Approach for Extracting and Segmenting High-Variance References from PDF Documents. In *Proc. of JCDL 2019*. ACM.
[2] Behnam Ghavimi, Wolfgang Otto, and Philipp Mayr. 2019. EXmatcher: Combining Features Based on Reference Strings and Segments to Enhance Citation Matching. Manuscript submitted for publication.
[3] Daniel Hienert, Frank Sawitzki, and Philipp Mayr. 2015. Digital Library Research in Action - Supporting Information Retrieval in Sowiport. *D-Lib Magazine* 21, 3/4 (2015). https://doi.org/10.1045/march2015-hienert
[4] Martin Körner, Behnam Ghavimi, Philipp Mayr, Heinrich Hartmann, and Steffen Staab. 2017. Evaluating Reference String Extraction Using Line-Based Conditional Random Fields: A Case Study with German Language Publications. In *New Trends in Databases and Information Systems*. Vol. 767. Springer International Publishing, 137–145.
[5] Henk F. Moed. 2005. *Citation Analysis in Research Evaluation (Information Science & Knowledge Management)*. Springer-Verlag, Berlin, Heidelberg.
[6] Silvio Peroni and David Shotton. 2018. The OpenCitations Data Model. (2 2018).

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

[11] https://github.com/exciteproject/EXpublisher
[12] http://opencitations.net/corpus/be/01101.html