Preface to the Understanding Literature References in Academic Full Text workshop at JCDL 2022

Anastasiia Iurshina[1], Muhammad Ahsan Shahid[2], Tobias Backes[2], Philipp Mayr[2] and Steffen Staab[1]

[1]University of Stuttgart
[2]GESIS – Leibniz-Institute for the Social Sciences

Abstract

This preface describes the Understanding Literature References in Academic Full Text (ULITE) workshop. ULITE was held as a virtual event on June 24, 2022. It was co-located with the Joint Conference on Digital Libraries (JCDL 2022).

1. Introduction

The goal of the ULITE workshop[1] at JCDL 2022 is to engage communities interested in the broad topic of literature reference understanding and automatic processing of scientific fulltext publications. Our workshop has a focus on working with open infrastructures/tools and offering the extracted information as open data for reuse. Our view is to expose people from one community to the work of the respective other community and to foster fruitful interaction across communities.

The target audience of the workshop are researchers and practitioners, junior and senior, from Natural Language Processing (NLP) and Information Extraction as well as Information Retrieval and Bibliometrics/Scientometrics. These could be IR/NLP researchers interested in potential new application areas for their work as well as researchers and practitioners working with bibliometric data and interested in how IR/NLP methods can make use of such data.

2. Overview of the papers

Four submissions were accepted as research papers at ULITE workshop. In addition to this, we had 4 invited talks. For two of the talks, papers were submitted. All 6 papers are included in CEUR-WS proceeding. The slides of the presentations can be found on the workshop website.

Anastasiia.Iurshina@ipvs.uni-stuttgart.de (A. Iurshina); Ahsan.Shahid@gesis.org (M. A. Shahid); Tobias.Backes@gesis.org (T. Backes); Philipp.Mayr@gesis.org (P. Mayr); Steffen.Staab@ipvs.uni-stuttgart.de (S. Staab)
0000-0002-1231-2314 (A. Iurshina); 0000-0002-1231-2314 (S. Staab)

© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

[1]https://exciteproject.github.io/ULITE-ws/

CEUR-WS.org/Vol-3220/preface.pdf

2.1. Keynote

We had one keynote speaker:

Silvio Perroni (University of Bologna, Italy) **OpenCitations: a short introduction**: In this paper, Silvio introduced a brief history of open citations, their main characteristics and use in the context of OpenCitations, a scholarly infrastructure organisation dedicated to open scholarship and the publication of open bibliographic and citation data using Semantic Web technologies.

2.2. Research papers

Four research papers were presented at ULITE.

- Frederik Arnold and Robert Jäschke:
  A Game with Complex Rules: Literature References in Literary Studies
- Christian Boulanger and Anastasiia Iurshina:
  Extracting bibliographic references from footnotes with EXcite-docker
- Bastian Birkeneder, Philipp Aufenvenne, Christian Haase, Philipp Mayr and Malte Steinbrink:
  Extracting literature references in German Speaking Geography – the GEOcite project
- Tarek Saier, Meng Luan and Michael Färber:
  A Blocking-Based Approach to Enhance Large-Scale Reference Linking

2.3. Invited talks

Four invited talks were given, for two of them papers were submitted:

- Arcangelo Massari and Ivan Heibi:
  How to structure citations data and bibliographic metadata in the OpenCitations accepted format
- Silvia Eunice Gutiérrez De la Torre, Julián Equihua, Andreas Niekler and Manuel Burghardt:
  Into the bibliography jungle: using random forests to predict dissertations' reference section

The two talks without papers:

- Bikash Joshi
  "Inline Citation Extraction from Scientific Manuscripts"
- Swati Sanagar
  "Finest Tool for Bibliography Reference Matching to Article and Deduplication"

3. Workshop outcome

The main outcome of the joined discussion between participants is the decision to join forces in creating a multi-domain golden standard dataset for literature references extraction and segmentation. It is clear that the lack of annotated data is one of the most serious limitations for the progress in the task of automatic reference extraction and segmentation. As annotating of the data is a very time-consuming and laborious process, it is difficult for one team to obtain enough data. However, by combining several smaller datasets, we can create one of the substantial size. In addition to the size, as the participants come from very different domains (law, literature, geography etc), the format of the annotated articles would be very diverse.