An Evaluation of Large Language Models in the Context of Converting Plain Text Citations to Machine-Readable Formats

Abstract

BibTeX is a popular citation management system that is used across academic disciplines. There are a great many tools online that can convert BibTeX citations to styles such as APA, MLA, and Chicago. However, there are much fewer tools that can convert plain text citations to BibTeX. One promising approach to this problem is to use large language models to convert plain text citations to BibTeX. We present a dataset of [[change to actual number]] plain text citations and their BibTeX equivalents across a variety of citation styles. Further, we evaluate the performance of several large language models on creating BibTeX entries from plain-text citations. Finally, we discuss potential economic and social benefits of reducing the time spent on this process.

Background

The academic research librarian is often one who is most often concerned with the intricacies of citation styles and their importance. Hensley As libraries continue to become more involved with publishing through initiatives such as the Library Publishing Coalition, the necessity for creating robust, reliable infrastructure for automating these publications increases. *Library Publishing Coalition | Academic & Research Libraries Engaged in Scholarly Publishing* The author's host institution, Northwestern University Libraries, is a member of this coalition and has a vested interest in the development of such infrastructure.

The first journal published by Northwestern University Libraries in partnership with the Center for Applied Transgender Studies, *The Bulletin of Applied Transgender Studies*, required several days work to produce a new issue due to the manual nature of the citation conversion process to BibTeX. *Bulletin of Applied Transgender Studies | A Publication of the Center for Applied Transgender Studies* With the integration of out-of-the-box LLMs into our workflow, we have been able to reduce the time spent on this process to a matter of minutes.

Data Collection Methodology

Data was collected using a combination of OpenAlex and Habanero, orchestrated using a python script. We made use of the OpenAlex API to retrieve a random DOI from the OpenAlex database, checked if they were valid DOIs that were in English, and then used the Habanero API to retrieve the BibTeX and citation for that DOI. We then used the Habanero API to choose a random citation style from the following list:

- apa

- harvard3
- elsevier-harvard
- mla
- ecoscience
- chicago author date
- council-of-science-editors

These citation styles were chosen due to their popularity and the fact that they are supported by the Habanero Content Negotiation API. We then used the Habanero Content Negotiation API to retrieve the citation in plain text and BibTeX format.

The output of this process was a pandas dataframe with the following columns:

- DOI
- BibTeX Citation
- Plain Text Citation
- Plain Text Citation Style

The DOI can be used as the unique identifier for each row in the dataframe. The BibTeX Citation and Plain Text Citation columns can be used as input and output values for training LLMs. The Plain Text Citation Style column can be used to filter the data by citation style.

LLM Performance Evaluation Methodology

Results

Data and Code Availability

Data and code are available for (re)use under the CC0 license. "CC0" Data is available in CSV format in Northwestern University's Institutional Repository, Arch.[1] Code is available on GitHub. [2]

Bulletin of Applied Transgender Studies | A Publication of the Center for Applied Transgender Studies. https://bulletin.appliedtransstudies.org/. Accessed 27 Oct. 2023.

"CC0." *Creative Commons*, https://creativecommons.org/public-domain/cc0/. Accessed 4 Jan. 2024.

Hensley, Merinda Kaye. "Citation Management Software: Features and Futures." *Reference & User Services Quarterly*, vol. 50, no. 3, 2011, pp. 204–08, https://www.jstor.org/stable/41241164.

Library Publishing Coalition | Academic & Research Libraries Engaged in Scholarly Publishing. https://librarypublishing.org/. Accessed 4 Jan. 2024.

[1] https://arch.library.northwestern.edu/

[2] https://github.com/aerithnetzer/Use_of_Citations_in_LLMs