Citation Detective: a Public Dataset to Improve and Quantify Wikipedia Citation Quality at Scale

Ai-Jou Chou
National Chiao Tung University
Taiwan
ajchou@cs.nctu.edu.tw

Guilherme Gonçalves
Google
Ireland
guilherme.p.gonc@gmail.com

Sam Walton
Wikimedia Foundation
United Kingdom
swalton@wikimedia.org

Miriam Redi
Wikimedia Foundation
United Kingdom
miriam@wikimedia.org

ABSTRACT

Machine learning models designed to improve citation quality in Wikipedia, such as text-based classifiers detecting sentences needing citations ("Citation Need" models), have received a lot of attention from both the scientific and the Wikimedia communities. However, due to their highly technical nature, the accessibility of such models is limited, and their usage generally restricted to machine learning researchers and practitioners. To fill this gap, we present *Citation Detective*, a system designed to periodically run Citation Need models on a large number of articles in English Wikipedia, and release public, usable, monthly data dumps exposing sentences classified as missing citations. By making Citation Need models usable to the broader public, *Citation Detective* opens up new opportunities for research and applications. We provide an example of a research direction enabled by *Citation Detective*, by conducting a large-scale analysis of citation quality in Wikipedia, showing that article citation quality is positively correlated with article quality, and that articles in Medicine and Biology are the most well sourced in English Wikipedia. The *Citation Detective* data and source code will be made publicly available and are being integrated with community tools for citation improvement such as *Citation Hunt*.

KEYWORDS

datasets, neural networks, Wikipedia, data dumps

ACM Reference Format:
Ai-Jou Chou, Guilherme Gonçalves, Sam Walton, and Miriam Redi. 2020. *Citation Detective*: a Public Dataset to Improve and Quantify Wikipedia Citation Quality at Scale. In *Proceedings of The Web Conference (Wiki Workshop'20)*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Wiki Workshop'20, April 2020, Taipei, Taiwan

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The core content policy of Verifiability[1] is one of the key mechanisms that Wikipedia communities adopt to monitor the quality of its content. The policy requires any information which is likely to be challenged to be backed by a citation to a reliable source.

One of the methods by which Wikipedia's editor communities flag verifiability issues is by tagging material with a *[Citation Needed]* flag. This flag can apply to one or more sentences of text, alerting readers and fellow editors that the preceding content is missing a citation to a reliable source. Articles with any content tagged with *[Citation Needed]* are added to maintenance categories for review. On the English Wikipedia, as of February 2020, this category contains more than 380,000 articles.[2]

Wikipedia's editor communities have created tools and workflows to address the backlog of unsourced content on the encyclopedia, particularly to aid in navigating and filtering the list. One such tool is *Citation Hunt*[3], a microcontribution tool that presents users with a single sentence or paragraph ending in a *[Citation Needed]* flag, allowing filtering of the selected articles by article topic. The user is asked to find a reliable source of information which could verify the content, and add it to the article. In this way, Wikipedia editors can address reference gaps one entry at a time, search for unsourced content by topic, or even use the tool as a simple entry point for new contributors, such as in the 1Lib1Ref campaign.[4]

At the time of writing there is no simple way for Wikipedia's editor communities to monitor citation needs at scale across the encyclopedia, nor to find cases of content missing a citation without prior addition of a *[Citation Needed]* flag. The true extent of the encyclopedia's unsourced content is therefore currently unknown.

A recent research work aimed to fill this gap by designing machine learning classifiers able to detect sentences needing citations in Wikipedia [7]: through a qualitative analysis of the citation guidelines in Wikipedia, the authors created a *taxonomy of reasons* why inline citations are required in Wikipedia, and then designed and open-sourced text-based classifiers to determine *if* a sentence needs a citation ("Citation Need" model), and *why*.

While the "Citation Need" model is a first step towards understanding Wikipedia citation quality at scale, its usability is limited

[1] https://en.wikipedia.org/wiki/Wikipedia:Verifiability
[2] https://en.wikipedia.org/wiki/Category:All_articles_with_unsourced_statements
[3] https://tools.wmflabs.org/citationhunt
[4] https://meta.wikimedia.org/wiki/The_Wikipedia_Library/1Lib1Ref/Resources

to those researchers and practitioners who are familiar with classifiers based on natural language processing. To overcome this issue, in this paper, we present a system called *Citation Detective*, which makes these models readily usable by the broader Wikipedia and research community. *Citation Detective* "productionizes" the Citation Need model by applying the classifier to a large number of articles from the English Wikipedia and periodically releasing a public dataset of unsourced statements on the encyclopedia.

The *Citation Detective* dataset enables a number of research works and applications. First, it enables to understand Wikipedia citation quality at scale, by allowing to quantify and track the proportion of unsourced and well-sourced content in Wikipedia articles. To show the potential of such a dataset for this task, in this paper we provide a large-scale analysis of the encyclopedia's citation coverage, exploring the data along dimensions of topic, quality, and popularity. Second, the dataset produced by *Citation Detective* can easily be integrated with tools such as *Citation Hunt* to improve community workflows, by surfacing unsourced content with no prior *[Citation Needed]* tag. At the time of writing, the *Citation Hunt* tool is being extended to accommodate sentence suggestions from the *Citation Detective* dataset.[5]

In this paper we provide an overview of the relevant research, data, and tools, a summary of our work on the *Citation Detective* dataset, and an analysis of the state of citation coverage on Wikipedia.

2 BACKGROUND AND STATE OF THE ART

This paper is closely related to the body of research and tools supporting efforts to improve citation coverage and quality in Wikipedia.

The Wikipedia editor community monitors the quality of information and citations through various mechanisms, including templates such as *[Citation Needed]* or *[Unreferenced]*. However, recent studies estimate that many articles might still have a small number of references or no references at all, and that readers rarely verify statements by clicking on inline citations [5, 6].

Tools such as *Citation Hunt* provide user friendly interfaces to help contributors fixing sentences which are missing reliable sources, and initiatives such as The Wikipedia Library[6] help editors find the right sources to cite. To further support researchers and editors in this task, the Wikimedia Foundation has recently released structured datasets to aid navigation of the citation space in Wikipedia. These datasets include a list of all citations with identifiers in Wikipedia, for all articles in all languages [3] and its extended version containing all citations with identifiers tagged with topics and accessibility labels [8].

Some recent publications have focused on machine-assisted recommendations for citation quality improvement. These efforts include source recommendations for outdated citations [2], and automatic detection of the *citation span*, namely the portion of a paragraph which is covered by an inline citation [1]. Redi et al. [7] designed a set of classifiers based on natural language processing that, given a sentence, can automatically detect whether it needs a citation ("Citation Need" classifier), and why ("Citation Reason").

[5]See prototype at: https://tools.wmflabs.org/aiko-citationhunt
[6]https://en.wikipedia.org/wiki/Wikipedia:The_Wikipedia_Library

In this paper, we extend the work in [7] in two ways. First, we design a framework to make the Citation Need model available to the public, by creating a system that periodically classifies a large number of sentences in English Wikipedia with the Citation Need model, and releases a dump of the sentences which are classified as needing citations. Second, we provide an analysis of citation quality in English Wikipedia by applying the Citation Need model at scale on a sample of articles.

3 CITATION DETECTIVE

We present here *Citation Detective*, a system that applies the Citation Need models to a large number of articles in English Wikipedia, producing a dataset which contains sentences detected as missing citations with their associated metadata (such as article name and revision id).

3.1 System Workflow

The workflow of producing the *Citation Detective* database includes the following steps.

3.1.1 Generating a List of Pages. Given an `article_sample_rate`, we query the page table from Wikipedia SQL replicas[7] to generate a `page_id` list. The Page ID is a unique identifier for Wikipedia articles preserved across edits and renames for pages in Wikipedia. The result of this step is a random sample of articles from English Wikipedia (which can be replicated for any other Wikipedia).

3.1.2 Retrieving Page Content. The page list is passed to the MediaWiki API[8] to retrieve the page content. For each page in the list, we query the MediaWiki API to get the title, revision ID, and content (text) of the article.

3.1.3 Constructing Model Input Data. An input instance for the Citation Need model is a made of (1) set of FasTtext [4] word vectors representing the each word in a sentence, and (2) the average word vector for all the words in the section title where the sentence lies (see [7] for more details). The public code repository for the Citation Need model[9] provides pre-defined dictionaries of words and section titles based on FastText. In this step, we aim to extract individual sentences and their section titles, and transform them into Fasttext embeddings using the sentence dictionary and the section dictionary provided along with the Citation Need model.

First, we broke an article into sections by the highest level section titles, and we discard sections that do not need citations such as "See also", "References", "External links". Then, we split a section paragraphs, and further divide it into sentences using NLTK's sentence tokenizer. Next, we split a sentence into words, and transform each word into its embedding by matching it with a key in the sentence dictionary. Similarly, we trasnform the section title into a section embedding using the section dictionary. If a words or a section title is not included in a dictionary, it will be assigned an average word embedding corresponding to unknown words (following the procedure in [7]). At the end of this step, we have, for each article, a set of sentences converted into word vectors and ready to be used as input data for the Citation Need model.

[7]https://www.mediawiki.org/wiki/Manual:Page_table
[8]https://www.mediawiki.org/wiki/API:Main_page
[9]https://github.com/mirrys/citation-needed-paper

3.1.4 Citation Need Model Prediction. We throw the word embeddings and section embeddings to the Citation Need model for predicting a score y in the range $[0, 1]$: the higher the score, the more likely that a sentence needs a citation.

3.1.5 Storing Data into Database. In the last step, we store into a SQL database each sentence with a score higher than $\hat{y} >= 0.5$: the text of the sentence, the text of the paragraph which contains the sentence, the section title, the revision ID of the article, and the predicted citation need score. The schema is shown in Table 1.

Field	Type	Description
id	integer	Primary key
sentence	string	The text of the sentence
paragraph	string	The text of the paragraph
section	string	The section title
rev_id	integer	The revision ID of the article
score	float	The predicted citation need score

Table 1: Schema of *Citation Detective*

3.2 System Implementation Details

In this section we briefly introduce the implementation details for *Citation Detective* and the important design decisions learnt from the technology transfer.

When processing the text corpus of Wikipedia articles, we need to parse Wikitext, also known as Wiki markup or Wikicode, which consists of special syntax and templates for inserting images, hyperlinks, tables, etc. `mwparserfromhell`[10] provides an easy-to-use and powerful parser for Wikicode so that we can simply get section titles and filter out infobox, images, tables that are not necessary to throw in the model. While we need to process the data for the Citation Model, in the *Citation Detective* database, we eventually store sentences in the original, unprocessed Wikicode format, which means sentences may contain any Wiki markups such as templates and links. This design decision is to ensure other tools can consume the data more easily. Tools just have to look for that text in the Wikicode at the specified revision. While plain text format is easy for humans to read, matching it with its corresponding Wikicode is non trivial for machine-assisted tools and other stakeholders.

Since the system is meant to work on a large number of articles in Wikipedia, efficiency is an important issue. In practice, to classify sentences at scale, we leverage multiple processes on a given machine. We observed that one of the bottlenecks of the system spend is the time needed by the Wikipedia API to query the content of the articles. Therefore, we use a pool of worker processes to parallelize the task, distributing the querying task across processes. On the other hand, in order to parallelize the model prediction, we load the Citation Need model in a separate process and shared it with the other worker processes. Worker processes communicate with the server process via a proxy and perform prediction tasks across processes. In the experiment, the multiprocess version is 3.3x speedup compared to the single-process version. For the first version of *Citation Detective*, we set the `article_sample_rate` at 0.2.

[10]https://github.com/earwig/mwparserfromhell

Number of sentences	4,120,432
Number of Articles	483,460
Sentences per Article	9.8

Table 2: Summary of data for Citation Quality Analysis

3.3 Database Release and Update

The Citation Detective database is now available on the Wikimedia Toolforge[11] as the public SQL database `citationdetective_p`. Every time we update the database, the Citation Detective takes a random 2% sample of articles in English Wikipedia, namely around 120 thousand articles, resulting in around 380 thousand sentences in the database which are classified as needing citations. Access to the database from outside the Toolforge environment is not currently possible, but is under investigation for the future.

4 ANALYZING CITATION QUALITY AT SCALE

In this Section, we provide an example of use-case for systems like *Citation Detective*: quantifying the quality of citations in Wikipedia at scale. We use the Citation Need models to quantify citation quality on hundreds of thousands of articles from English Wikipedia, we analyze the relation between article quality and citation quality, and break down these statistics by article topic.

4.1 Data Collection

To perform this analysis, we first need data about articles, their sentences, and their citation need. Since, at the time of writing, the *Citation Detective* system is still under refinement, we create a one-off dataset for this experiment. We sample 7% of the articles in English Wikipedia, and then randomly sample 10 sentences for each article. We report in Table 2 a summary of the data used for these experiments.

4.1.1 Extracting Article Quality, Topic, Popularity, and Reference Quality Labels. We then extract basic article properties. First, we use the ORES scoring platform[12] to extract the articles' *topic* category (e.g. Science, Economics, etc) and level of *quality* (Stub, Good Article, etc.). We also use the Pageviews API[13] to get the number of total *views* received by each article during the month of May 2019. Finally, we check which articles in our data have been marked by editors as "Missing Sources", i.e. they appear in the category *"All articles needing additional references"*[14]. We will use these manual labels as grountruth to validate article's citation quality.

4.1.2 Computing Article's Citation Quality. Using the Citation Need model, we then compute article citation quality, namely the proportion of "well sourced" sentences in an article. To do so, we classify all sentences with the model, and label each sentence with a binary Citation Need label y according to the model output: $y = [\hat{y}]$, where $[\cdot]$ is the rounding function and $\hat{y}$ is the output of the Citation Need model. When $y = 1$, the sentence needs a citation, when $y = 0$, the sentence doesn't need one. Next, we aggregate sentence-level Citation Need labels to calculate the article citation quality Q. Q is

[11]https://tools.wmflabs.org/
[12]https://ores.wikimedia.org
[13]https://wikimedia.org/api/rest_v1/
[14]See complete list for English Wikipedia at this query: https://quarry.wmflabs.org/query/34358

the proportion of sentences needing citations that already have a citation in the text.

$$Q = \frac{1}{p} \sum_{i \in P} c_i, \tag{1}$$

where p is the number of sentences needing citations for a given article, i.e. having $y = 1$; c_i reflects the presence of a citation in the original text of the sentence i: $c = 0$ if the sentence doesn't have an inline citation in the original text or $c = 1$ if the sentence has an inline citation in the original text; P is the set of p sentences needing citations in the article according to the Citation Need model.

When $Q = 0$ the quality is very low, as none of the sentences classified by the model as needing citations actually have a citation in the original text.

4.2 Results

We report here a set of summary results of articles' citation quality analysis, broken down by articles' characteristics.

4.2.1 Citation Quality Score VS Manual Reference Quality Annotations. To validate the accuracy of the citation quality score, we look at the average citation quality for articles that have been marked by editors as "Missing Sources" (our groundtruth), and compare it with the average Q for all other articles. We find that the average citation quality score across all articles is 0.66: namely, in average, 66% of the sentences in an article that are marked as missing citations already have an inline citation in the original text. This percentage drops for articles marked as "Missing Sources": the average Q for those articles is 0.49, thus showing that the Citation Quality score can correctly expose those articles which require more attention because of low quality references.

4.2.2 Citation Quality Score VS Article Quality and Popularity. To further investigate the accuracy of the citation quality score, we correlate, for each article, the citation quality score Q with the article quality score previously computed through ORES. We observe a strong Pearson correlation (statistically significant with $p - value < 0.05$) between these 2 quantities ($\rho = 0.34$). We also compute the correlation between citation quality and article popularity, finding a significant correlation of $\rho = 0.09$. Although weaker than the correlation between citation quality and article quality, this positive correlation is probably due to the fact that very popular articles tend also to be of high quality (there is a significant correlation of $\rho = 0.14$ between article quality and popularity).

4.2.3 Breakdown of Citation Quality by Topic. Finally, we break down citation quality by article topic. We compute the average citation quality for all articles belonging to a given topic, and report the results in Figure 4.2.3. We find that the most well sourced articles ($Q > 0.85$) belong to the Medicine and Biology topics. "Language and Literature", the topic category hosting most biographies, also ranks among the top well-sourced topics. We find that articles in Mathematics and Physics tend to be marked as poorly sourced. This is probably due to the fact that these articles don't report many inline citations, as the proof of the scientific claims is in the formulas/equations that follow, and these articles tend to have a few references cited in general.

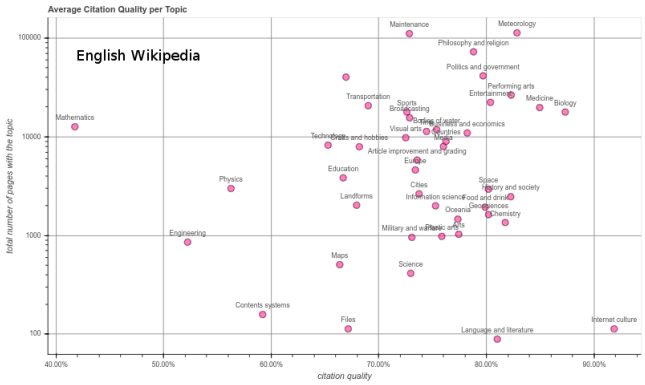

Figure 1: Average article citation quality score by article topic. X axes corresponds the average Q for all articles in a given topic, and Y axes corresponds to the number of articles for a given topic in the sample drawn for the analysis.

5 CONCLUSIONS

We presented a framework to analyze, monitor and improve citation quality at scale. We designed *Citation Detective*, a system that applies Citation Need models to a large number of articles in English Wikipedia, and periodically released data dumps exposing unsourced sentences in Wikipedia. To give an example of the potential applications of the *Citation Detective* data, we provided a large-scale analysis of citation quality in Wikipedia, showing that citation quality is positively correlated with article quality, and that articles in Medicine and Biology are the most well sourced in English Wikipedia.

This analysis provides an initial overview of the potential applications of *Citation Detective*, and is a limited view on the overall picture, both within the English Wikipedia, and across the other (nearly 300) language Wikipedia projects. Future work on this project could broaden this dataset to include a higher percentage (or even all) of the English Wikipedia's content. We may also consider selecting articles non-randomly, such as ensuring the dataset contains all highly-viewed or high quality articles. Additionally, the Citation Need model is capable of analysing other language projects, for which additional datasets could be made available.

The data is presently only available within the Toolforge environment due to technical limitations. In future work we aim to make the database more accessible, such as through the *Quarry* database querying service.[15]

REFERENCES

[1] Besnik Fetahu, Katja Markert, and Avishek Anand. 2017. Fine Grained Citation Span for References in Wikipedia. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*. 1990–1999. https://aclanthology.info/papers/D17-1212/d17-1212

[2] Besnik Fetahu, Katja Markert, Wolfgang Nejdl, and Avishek Anand. 2016. Finding News Citations for Wikipedia. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016*. 337–346. https://doi.org/10.1145/2983323.2983808

[15] https://meta.wikimedia.org/wiki/Research:Quarry

[3] Aaron Halfaker, Bahodir Mansurov, Miriam Redi, and Dario Taraborelli. 2019. Citations with identifiers in Wikipedia. (12 2019). https://doi.org/10.6084/m9.figshare.1299540

[4] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. FastText.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651* (2016).

[5] Włodzimierz Lewoniewski, Krzysztof Węcel, and Witold Abramowicz. 2017. Analysis of References Across Wikipedia Languages. *Communications in Computer and Information Science Information and Software Technologies* (2017), 561–573. https://doi.org/10.1007/978-3-319-67642-5_47

[6] Tiziano Piccardi, Miriam Redi, Giovanni Colavizza, and Robert West. 2020. Quantifying Engagement with Citations on Wikipedia. *arXiv preprint arXiv:2001.08614* (2020).

[7] Miriam Redi, Besnik Fetahu, Jonathan Morgan, and Dario Taraborelli. 2019. Citation Needed: A Taxonomy and algorithmic assessment of Wikipedia's verifiability. In *Proc. International Conference on World Wide Web*.

[8] Miriam Redi and Dario Taraborelli. 2018. Accessibility and topics of citations with identifiers in Wikipedia. (7 2018). https://doi.org/10.6084/m9.figshare.6819710.v1