Wiki Workshop (10th edition) – May 11, 2023

Longitudinal Assessment of Reference Quality on Wikipedia

Aitolkyn Baigutanova
KAIST

Jaehyeon Myung
KAIST

Diego Saez-Trumper
Wikimedia Foundation

Ai-Jou Chou
Wikimedia Foundation

Miriam Redi
Wikimedia Foundation

Changwook Jung
KAIST

Meeyoung Cha
IBS & KAIST

Abstract

In this work, we share key findings from our recent study (Baigutanova et al., 2023) that will be published at WWW 2023. We present a methodology to measure Wikipedia's content verifiability through its reference quality by defining two metrics. The first is the *reference need (RN)* index, which represents the proportion of citation-missing sentences among those that require a citation. The second is the *reference risk (RR)*, which measures the proportion of non-authoritative sources according to the classification in the perennial source list, a community-driven reference labeling. These two indicators help assess the status quo of reference quality on Wikipedia over a decade from 2010 to 2020.[1] Our large-scale temporal analysis shows how the ratios of unreliable sources and citation-missing sentences have decreased over time. We further present our findings on the Wikipedia community's efforts to maintain the list of risky sources to eliminate such references.

Keywords: Wikipedia, Verifiability, the Web, NLP, Fake News

Methods

Reference Need (RN)

We describe the Citation Detective tool, which uses machine learning to compute the RN score for a given article revision, in the full version of our paper (Baigutanova et al., 2023). The tool classifies all sentences in a revision and labels each sentence with a binary Citation Need (Redi et al., 2019) 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 does not need one. We compute each revision's RN score by aggregating sentence-level Citation Need labels:

$$RN = 1 - \frac{1}{|P|} \sum_{i \in P} c_i, \tag{1}$$

where P is the set of sentences needing citations for a given article; c_i reflects the presence of a citation in the original text of the sentence i: $c = 0$ if the sentence does not have an inline citation in the original text or $c = 1$ if the sentence has an inline citation in the original text.

Reference Risk (RR)

The Wikipedia editing community maintains a classification of the reliability of the sources that have been frequently questioned, which is known as the *perennial sources list*[2]. Our research utilizes blacklisted and deprecated categories of this classification as risky sources, as they are suggested to be prohibited in general. Using the public Wikipedia XML dumps, we ran a regular expression to extract risky references in article revisions. Then, the revision's RR score is computed as the proportion of sentences containing unreliable references in that revision:

$$RR = \frac{x}{N}, \tag{2}$$

where N is the total number of citations in a given revision; x is the number of unreliable references. Revisions not including any reference are omitted in this analysis.

Data

We built three datasets from the English edition of Wikipedia. (i) **Random** dataset includes 3,177,963 revisions of randomly sampled 20K pages. (ii) **Top** dataset includes 23,802,067 revisions of 10K pages that received the highest total page views in the English Wikipedia within the analyzed period, as computed by Wikimedia's Pageviews API.[3] Every editing revision is logged with the following metadata: revision id, timestamp, user id, prior revision count of the editing user, user type anonymous or not, bot or not, page id, revision byte size difference compared to the prior revision, and revision type minor or not. As the scope of this study is limited to understanding the role of human editors in maintaining the reference quality of Wikipedia articles, we filtered out edits made by bots in the further analysis.

[1] For data and code, refer to `https://github.com/aitolkyn99/wiki_reference_quality`

[2] `https://en.wikipedia.org/wiki/Wikipedia:Reliable_sources/Perennial_sources`

[3] `https://wikitech.wikimedia.org/wiki/Analytics/AQS/Pageviews`

We built the third dataset to examine the lifespan of deprecated and blacklisted domains. We froze the date to January 2022 and obtained the history of all references listed in the perennial sources list used until that point. (iii) **Reference History** dataset consists of 4,203,467 occurrences of references that are still existing and that are removed. The dataset consists of the following information for each occurrence of a reference: the page id, the timestamp when the reference was added, the timestamp when it was removed, the domain of the reference, the category of the domain, and the timestamp when the corresponding domain was classified as deprecated or blacklisted in the perennial source list if applies. The removal timestamp is blank if a reference was added but not yet removed.

Results

Evolution of Reference Quality

Tracking the RN and RR scores allows us to examine how reference quality has evolved over the past decade. The evolution of the reference need score is shown in the left plot of Figure 1. First, the average reference need score per article went down gradually over the last ten years, dropping by around 20% in both *Top* and *Random* datasets. This demonstrates that a greater proportion of Wikipedia pages now include citations or more than 60% of citation-requiring sentences accompany a reference.

The evolution of the reference risk score is shown in the right plot of Figure 1. The risk score has remained below 1% throughout the analyzed period. While the score only started to decrease in 2018 for the *Random* dataset, the *Top* dataset saw a decline starting in 2016. The decrease in the RR score coincided with the introduction of the perennial sources list in 2018. This might suggest that the collaborative effort of Wikipedia editors enabled them to address newly registered non-authoritative sources, resulting in a decrease in the following years. We observe that the RR scores across the two datasets have increasingly diverged over the past few years.

Lifespan of Risky Sources

To explore the role of community-driven work in the evolution of reference quality, we examine whether classifying sources in the perennial source list as ”deprecated” or ”blacklisted” motivates editors to remove existing risky references. We calculate the lifespan of risky references as the time elapsed between their addition and removal using the *Reference History* dataset. We analyze the lifespan of references within a year before and after their classification in the perennial sources list, as the list was established in 2018.

Figure 2 shows the median lifespan (or the number of days a reference survives before being removed by future edits) of risky references decreased by more than threefold once they were added to the perennial list by editors. Additionally, the lifespan of risky references at the 75th percentile decreases by approximately two months. These results indicate that labeling of perennial sources encouraged editors to remove unreliable references quickly if they were labeled undesirable. There was no definite consensus among deprecated sources regarding the domains of ”Daily Mail” and ”The Sun.” Their status was the subject of multiple discussions, so they were excluded from our main analysis.

Discussion/Conclusions

The RN index gradually decreased over the past decade, indicating that more articles now accompany references. This trend results from an increasing volume of community initiatives to improve citation coverage, including the exceptional work done by editors and the success of tools to ensure Wikipedia’s verifiability. These efforts improve Wikipedia itself and, in return, result in a higher quality encyclopedia for humans and machines.

Our results may be considered a lower bound of the reference risk value because the perennial sources list only covers a small fraction of potentially unreliable sources. Unfortunately, using external reliability indexes and fact-checking systems is difficult in the Wikipedia context, given that existing lists are country-specific or not generic enough to cover the diversity of topics and sources. Creating a global index of source reliability would improve this estimate, support targeted interventions in specific content areas, and expose potential disinformation attacks from malicious users. Together with other efforts to build trust around the world, our scientific community could support such a global effort to improve and keep an eye on the quality of Wikipedia’s sources that directly affect the services people use. Systems that help automatically flag the presence of newly added unreliable sources could help editors monitor reference quality, and this paper provides a foundational methodology to build such support tools.

References

[Baigutanova et al.2023] Aitolkyn Baigutanova, Jaehyeon Myung, Diego Saez-Trumper, Ai-Jou Chou, Miriam Redi, Changwook Jung, and Meeyoung Cha. 2023. Longitudinal assessment of reference quality on wikipedia. In *proc. of the WWW.*

[Redi et al.2019] Miriam Redi, Besnik Fetahu, Jonathan Morgan, and Dario Taraborelli. 2019. Citation needed: A taxonomy and algorithmic assessment of wikipedia’s verifiability. In *proc. of the WWW.*

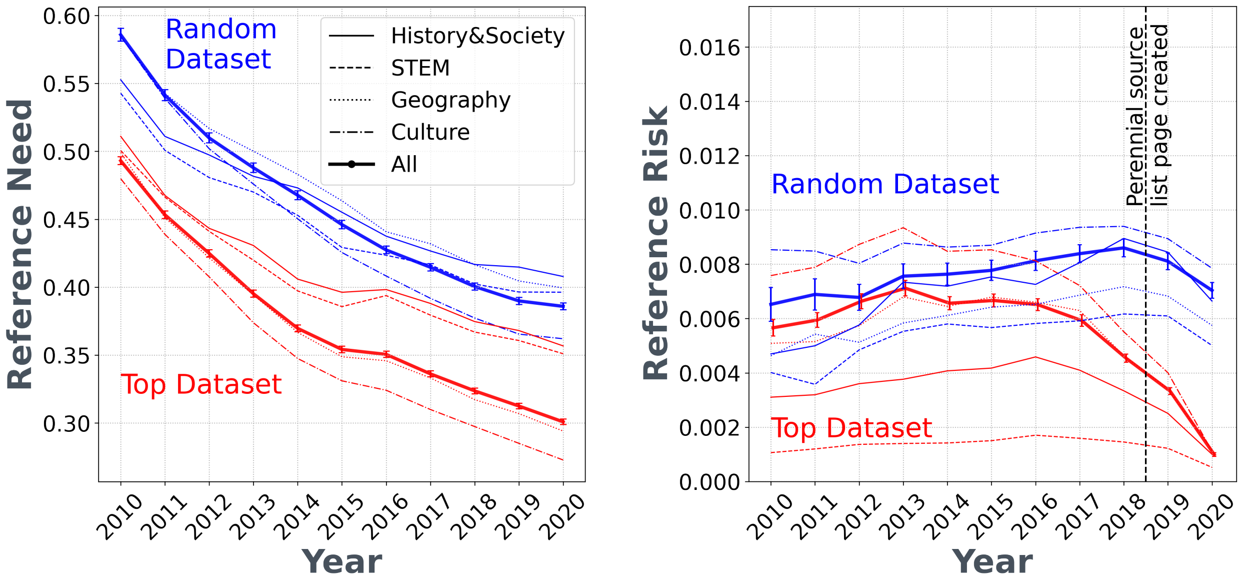

Figure 1: The evolution of reference quality in Wikipedia. Error bars represent the standard error. (Left) Reference need (RN) scores gradually decreased over the last decade, indicating an improved reference coverage of articles. The drop is nearly 20 percent point over the decade. (Right) Reference risk (RR) scores remain under 1% and show a decreasing trend in recent years, suggesting a reduction in the use of risky references after the introduction of the perennial sources list in 2018.

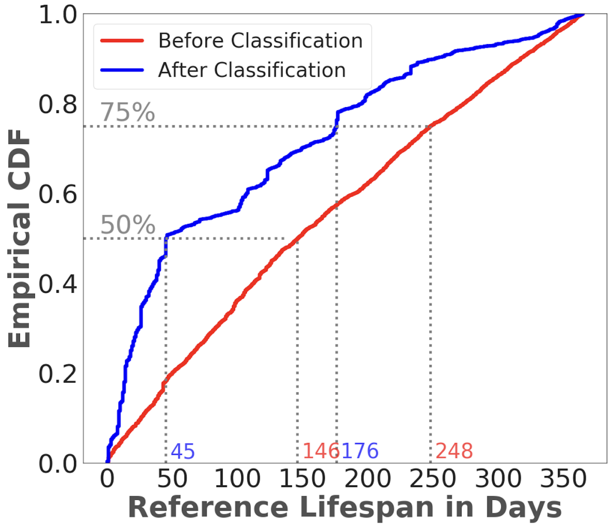

Figure 2: The lifespan of unreliable sources a year before and after being added to the perennial sources list. Sources have a short lifespan on Wikipedia once marked as unreliable.