nature machine intelligence

Article

https://doi.org/10.1038/s42256-023-00726-1

Improving Wikipedia verifiability with AI

In the format provided by the authors and unedited

Supplementary Information

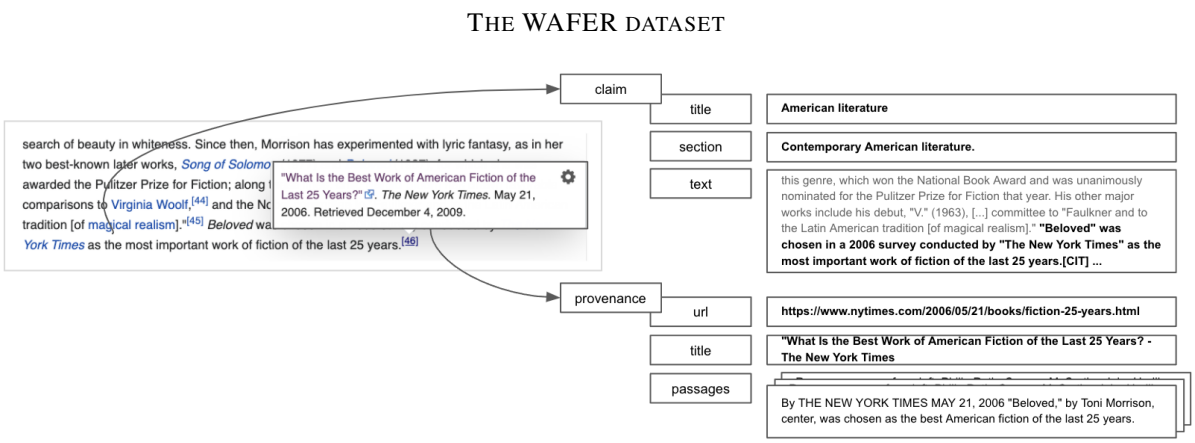

Figure 5. Example citation from the WAFER dataset, each datapoint contains the claim and source (document split in passages and other metadata)

	featured			random			micro avg.		
	P@1	SR@100	SR@200	P@1	SR@100	SR@200	P@1	SR@100	SR@200
1st stage - Sphere Retrieval									
1. dense - DPR multi-task pretrained	5.33	33.82	-	6.81	30.70	-	6.57	31.22	-
2. dense - DPR from scratch	11.25	66.94	73.71	18.25	72.09	78.04	17.12	71.26	77.34
3. sparse - BM25 no expansion	15.58	68.44	-	24.57	74.18	-	23.1	73.24	-
4. sparse - BM25 with expansion	15.72	73.17	77.78	26.14	80.16	84.36	24.45	79.02	83.30
2. dense + 4. sparse	-	-	**81.84**	-	-	**85.69**	-	-	**85.07**
2nd stage - Evidence Ranking									
verification engine (2. dense + 4. sparse)	**47.29**	**81.71**	-	**48.49**	**85.46**	-	**48.29**	**84.85**	-

Table 2. WAFER test results. Best numbers in bold.

Evaluation details

Experimental Data and Setting

We collect WAFER, a large scale dataset of English Wikipedia inline citations ($\approx 3.8M$ instances - see Table 4 for statistics) which are aligned to a snapshot of the web to obtain the full textual content of the cited sources. Each instance in WAFER contains metadata from the claim's article, the text around the citation within the article (with a marker indicating the citation position), and metadata of the cited source, including title and full textual content (see Figure 5 for an example). We create a cross-validation split at the article level—not at the claim level—to avoid potential test leakage into the training data. We find that only about 2% of the claims are in the development and evaluation set are repeated verbatim in the training set.

Both the Wikipedia snapshot we consider (*i.e.*, from KILT[30]) as well as the web snapshot (*i.e.*, a CCNet[31] dump from Sphere[14] which contains 134M web articles, split into 906.3M passages) are from August 2019. We use Sphere's web snapshot as the corpus for retrieval. Aligning the citations in the Wikipedia snapshot and Sphere's web snapshot leads to $\approx$ 250k retrievable citations. From those we sample $\approx$ 4.5k for testing and development each, making all the cited documents in our *test* and *dev* sets retrievable from the Sphere corpus. To increase the size of the training data, we match the remaining unaligned citations in the Wikipedia snapshot against several other Common Crawl snapshots from 2017 to 2019, collecting an additional $\approx$ 3.5M citations which are not retrievable from the Sphere corpus but which can be used for training models.

We distinguish two types of Wikipedia articles: *featured articles*[32] and *non-featured articles*. Featured articles are a tiny fraction (*i.e.*, 0.09%) of articles chosen by Wikipedia's editors as examples for their high quality. Therefore, we use the featured articles only for evaluation given their limited number ($\approx$ 16% of test and dev citations). The remaining instances of the evaluation data are sampled from non-featured articles which can vary in quality in terms of writing or verifiability. We do not

Wikipedia content	
Article	https://en.wikipedia.org/w/index.php?title=Jayda%20Fransen&oldid=907222168
Input	Jayda Fransen [SEP] Section::::Political career.:Leadership of Britain First. [SEP] she has often marched while holding a white cross, in "Christian patrols" through predominantly Muslim districts of British towns. In March 2018, Fransen was sentenced to 36 weeks imprisonment after being convicted of three counts of religiously aggravated harassment. Fransen had formerly been involved with the English Defence League, but left due to its association with violence. She was an unsuccessful candidate in the 2014 Rochester and Strood by-election, and the 2016 London Assembly election. Section::::Political career. Section::::Political career.:Leadership of Britain First. Britain First, formed in 2011, is a British fascist political party founded by Paul Golding and Jim Dowson. Golding became the leader following the resignation of Dowson, and during this time Fransen was the deputy leader of the party. **Golding handed over the leadership role to Fransen in November 2016 due to his being sentenced to 2 months in prison for breaching a court order, although Fransen stated that his leave was in order "to address some important, personal family issues".[CIT]** Fransen stepped down from her leadership role in January 2019. Section::::Political career.:Rochester and Strood by-election, 2014. Fransen stood as Britain First's first parliamentary candidate for the Rochester and Strood by-election on 20 November 2014, during which she expressed sympathy for the UK Independence Party (UKIP) and its candidate Mark Reckless (a Conservative MP who had switched allegiances to UKIP), who went on to win the seat. Britain First's campaign for the by-election drew attention when the party uploaded a photo of Fransen together with local activists from UKIP, who responded by saying
Wikipedia citation	
Source	http://www.searchlightmagazine.com/2016/12/more-questions-than-answers-a-searchlight-investigation
Title	More questions than answers: a Searchlight investigation
Passage	brought against Golding? It came as no shock that Golding suddenly stood down from the leadership of Britain First on the first day of Mair's trial, in favour of his deputy Jayda Fransen. When will Golding face a charge of incitement? When will somebody with responsibility and authority respond to these questions? Jayda Fransen with her Britain First
Score	2.6
SIDE citation	
Source	https://www.theguardian.com/uk-news/2016/nov/03/deputy-leader-britain-first-guilty-over-verbal-abuse-muslim-woman-jayda-fransen-hijab
Title	Deputy leader of Britain First guilty over verbal abuse of Muslim woman
Passage	Deputy leader of Britain First guilty over verbal abuse of Muslim woman | UK news | The Guardian Deputy leader of Britain First guilty over verbal abuse of Muslim woman Far-right group's Jayda Fransen convicted of religiously aggravated harassment for shouting at woman wearing hijab Thu 3 Nov 2016 13.19 EDT Last modified on Tue 28 Nov 2017 07.03 EST Jayda Fransen arriving at Luton magistrates court. Photograph: David Mirzoeff/PA The deputy leader of far-right group Britain First has been found guilty of religiously aggravated harassment after hurling abuse at a Muslim woman wearing a hijab in front of her four young children. Jayda Fransen, 30, was fined nearly £2,000 at Luton and South Bedfordshire magistrates court for'
Score	9.97

Table 3. In this example, both Wikipedia and SIDE citation get a relatively low score from the *verification engine*, suggesting the latter was unable to find enough evidence to verify the claim.

include in these datasets citations marked with a *failed verification* template[33], which indicates that the source does not support what is claimed in the Wikipedia article. We set these citations aside in specific dev and test sets (*i.e.*, *fail-dev* and *fail-test*) in order to evaluate the ability of models to detect citations that fail verification.

split	size	articles	featured
train	3805958	-	0
dev	4545	-	16% (727)
test	4568	-	16% (738)
fail-dev	725	-	0
fail-test	730	-	0

Table 4. Statistics for the WAFER dataset.

We use popular retrieval metrics to measure the performance to rank the gold-cited document as high as possible in the retrieved results. As our retrieval is passage-based, the highest ranked passage of a document determines its rank. We consider *precision-at-1* (P@1), that is the percentage of evaluation instances in which the originally cited document was ranked in the first position among all retrieved documents. Additionally, we use *success-rate-at-k* (SR@k)—sometimes also referred to as HITS@k—which is the percentage of cases in which the originally cited document was amongst the top-k documents. We also use the Precision-Recall curve which measures the performance in terms of Precision when Recall is fixed to a certain level.

CommonCrawl snapshots considered

2017-26, 2017-39, 2017-51, 2018-13, 2018-26, 2018-39, 2018-51, 2019-18, 2019-30, 2019-43, 2020-05, 2017-30, 2017-43, 2018-05, 2018-17 ,2018-30, 2018-43, 2019-09, 2019-22, 2019-35, 2019-47, 2020-10, 2017-34, 2017-47, 2018-09, 2018-22, 2018-34, 2018-47, 2019-13, 2019-26, 2019-39, 2019-51

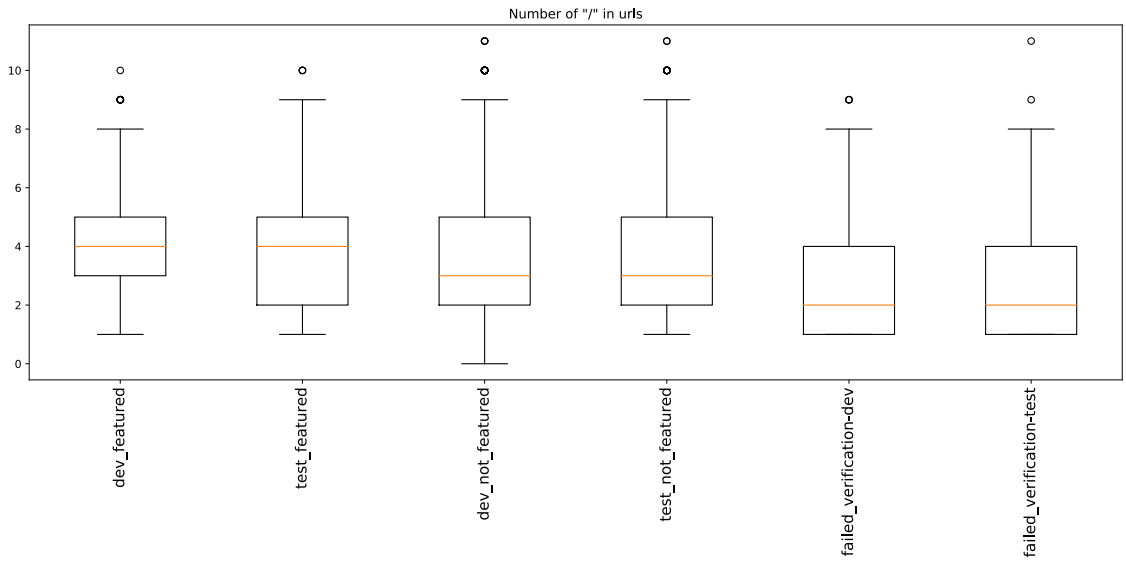

Figure 6. Url depth analysis. The Y-axes represents the number of "/" in URLs while on the X-axes URLs are grouped by the data split in which they are cited. URLs cited in featured Wikipedia articles tend to be more deep (i.e., specific, containing several "/") than URLs in citations marked as failed verification, that result more shallow (i.e., generic, containing few "/"). Statistics are as follows: dev_featured: Min=1, Q1=3.0, Median=4.0, Q3=5.0, Max=10. test_featured: Min=1, Q1=2.0, Median=4.0, Q3=5.0, Max=10. dev_not_featured: Min=0, Q1=2.0, Median=3.0, Q3=5.0, Max=11. test_not_featured: Min=1, Q1=2.0, Median=3.0, Q3=5.0, Max=11. failed_verification-dev: Min=1, Q1=1.0, Median=2.0, Q3=4.0, Max=9. failed_verification-test: Min=1, Q1=1.0, Median=2.0, Q3=4.0, Max=11.

Related Work

There is a large, passionate and engaged community who actively cares about, studies, and works to improve the verifiability of information in Wikipedia. The WikiProject Reliability[34], for instance, contains a set of tools, resources and reports which are aimed at improving the reliability of Wikipedia articles. One of these tools is Citation Hunt[35], which allows humans to check Wikipedia claims which have been flagged as not being backed by a reliable source and to propose a better citation. We

believe the technology presented in this paper can be integrated with similar tools to surface more unverified claims and suggest potential alternative citations to a human to validate.

Text-based classifiers able to detect claim needing citations[36,37] have received a lot of attention from both the scientific and the Wikimedia communities. We believe SIDE can be combined with such tools and recommend to Wikipedia editors a set of potential sources for claims needing a citation. Several studies have also been conducted on user interactions with citations[4–6,38,39] that are tangential to our work. There are a number of papers that approach citation recommendation for Wikipedia from different angles, such as by recommending citations from linked articles [40] to citation span detection[41] amongst other efforts. More broadly, citation retrieval and paper/source recommendation have also received attention in the scientific literature domain for many decades[37,42–44], albeit with less of a focus on verifiability of existing citations, with citations drawn from much smaller and less diverse sources than the open web, see[45] for a recent comprehensive review.

Several works have investigated the ability of AI to generate missing Wikipedia articles from scratch[46–49]. There exists AI tools, such as[50], that helps editors to bootstrap Wikipedia articles for underrepresented languages using these technologies. The SIDE engine can complement these systems and provide suggestions of supporting evidence from the web to back the article generation.

Finally, there exist a large body of research focused on fact-checking Wikipedia claims[12,13,51–56]. However, most of available resources are synthetically created to evaluate AI systems in a controlled environment. We believe that using real world supervision (*e.g.*, from Wikipedia citations) could be key to unlock a larger applicability of these systems.

Crowd Workers Annotation Guidelines

High level goal

In this HIT we give you a Wikipedia snippet with a highlighted claim and two passages from other web pages.
Your job is to:

1. **find evidence** supporting the Wikipedia claim in each of the Web passages;
2. **choose the passage** which you consider to be a better citation source for the claim.

How to identify the claim?

We highlight the sentence containing the claim in bold. Additionally, the **[CIT]** tag indicates the position either directly following the claim, or at the end of the sentence containing the claim. Please consider the *Wikipedia article title* as additional information for the claim.

Does the passage always contain evidence?

A passage may contain only partial evidence. Partial evidence is a sentence that only supports parts of the claim. In some cases a passage doesn't contain any evidence at all. Please use the passage url and title (both provided) as additional context for your decision but please don't retrieve and use the full text from the url.

Note: Some passages are control passages. We require good accuracy on our control passages to accept responses.

Detailed instructions for using the interface (Click to expand)

1. Read the Wikipedia snippet containing the claim and **identify the claim.**
2. For each of the two passages:
    - Read the passage.
    - Choose an option from the "Does it contain evidence?" drop down menu.
        - If the passage is not relevant, choose **"No"**. the passage contains evidence to fully support the claim, choose **"Yes, sufficient to support the claim"**.
        - If the passage contains partial evidence, choose **"Yes, partial"**.
    - If you selected yes, highlight the evidence in the passage using the mouse/cursor. The "Your Evidence Span" field will automatically populate with your highlighted evidence. If you made a mistake highlighting the evidence, simply do it again.
3. Once you have assessed both passages, you might need to answer some questions based on your responses. For instance, we ask you to provide an explanation if no evidence is found.

4. Finally, indicate which of the two provided passages you consider to be a better citation source for the **[CIT]** tag by choosing an option from the "Which passage do you prefer as citation" dropdown at the bottom of the page. Keep in mind that Wikipedia citations should be based on reliable, independent, published sources with a reputation for fact-checking and accuracy.
5. Click the submit button to end the HIT.