Improving Wikipedia Verifiability with AI

Fabio Petroni[1]*, Samuel Broscheit[2]†*, Aleksandra Piktus[1], Patrick Lewis[1], Gautier Izacard[1,4,5], Lucas Hosseini[1], Jane Dwivedi-Yu[1], Maria Lomeli[1], Timo Schick[1], Pierre-Emmanuel Mazaré[1], Armand Joulin[1], Edouard Grave[1], and Sebastian Riedel[1,3]

[1]Meta AI, [2]Amazon Alexa AI, [3]University College London,
[4]ENS, PSL University, [5]Inria

Abstract

Verifiability is a core content policy of Wikipedia: claims that are likely to be challenged need to be backed by citations. There are millions of articles available online and thousands of new articles are released each month. For this reason, finding relevant sources is a difficult task: many claims do not have any references that support them. Furthermore, even existing citations might not support a given claim or become obsolete once the original source is updated or deleted. Hence, maintaining and improving the quality of Wikipedia references is an important challenge and there is a pressing need for better tools to assist humans in this effort. Here, we show that the process of improving references can be tackled with the help of artificial intelligence (AI). We develop a neural network based system, called Side, to identify Wikipedia citations that are unlikely to support their claims, and subsequently recommend better ones from the web. We train this model on existing Wikipedia references, therefore learning from the contributions and combined wisdom of thousands of Wikipedia editors. Using crowd-sourcing, we observe that for the top 10% most likely citations to be tagged as unverifiable by our system, humans prefer our system's suggested alternatives compared to the originally cited reference 70% of the time. To validate the applicability of our system, we built a demo[1] to engage with the English-speaking Wikipedia community and find that Side's first citation recommendation collects over 60% more preferences than existing Wikipedia citations for the same top 10% most likely unverifiable claims according to Side. Our results indicate that an AI-based system could be used, in tandem with humans, to improve the verifiability of Wikipedia. More generally, we hope that our work can be used to assist fact checking efforts and increase the general trustworthiness of information online. All our code, data, indexes and models are publicly available at `https://github.com/facebookresearch/side`.

Introduction

Wikipedia is one of the most visited websites on the web (Ranking, 2022), and with half a trillion page views per year (Wikimedia, 2022), constitutes one of the most important knowledge sources today. As such, it is critical that any knowledge on Wikipedia is *verifiable*: Wikipedia users should be able to look up and confirm claims made on Wikipedia using reliable external sources (Verifiability, 2022). To facilitate this, articles provide inline citations that point to background material supporting the claim. Readers who challenge Wikipedia claims can follow these pointers and verify the information themselves (Piccardi et al., 2020; Lewoniewski et al., 2020; Kaffee & Elsahar, 2021). However, in practice this process can fail: a citation might either not entail the challenged claim, or its source might be questionable. Such claims may still be true, but a careful reader cannot easily verify them with the information at hand in the cited source. Under the assumption that a Wikipedia claim is true, its verification is hence a two stage process: 1) check the consistency of the existing source; 2) if that fails, search for new evidence, primarily online.

*Equal contribution.

†Work done during an internship with Meta AI.

[1] available at `https://verifier.sideeditor.com`

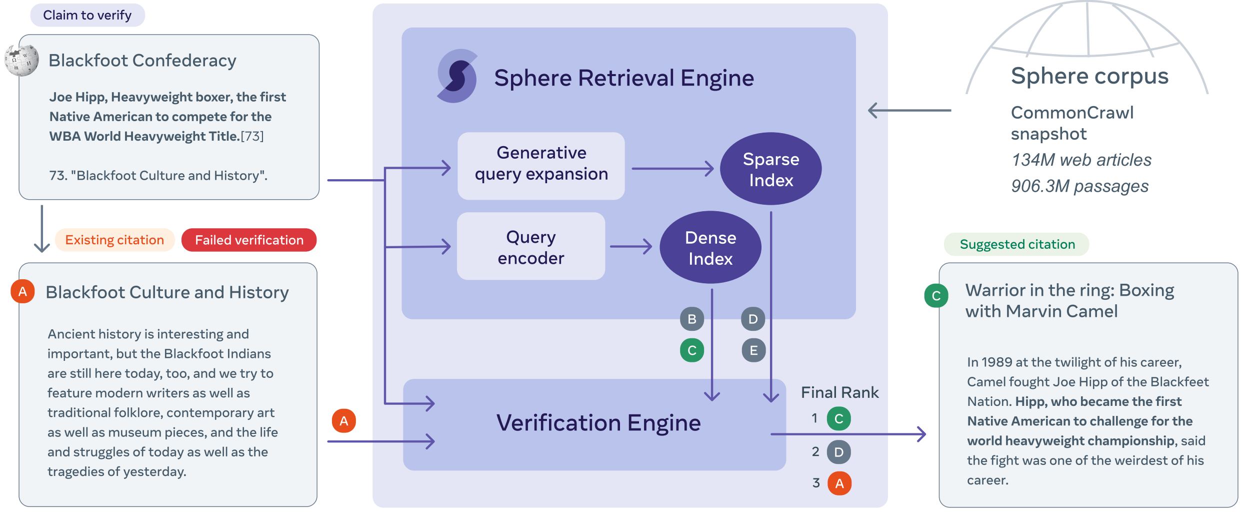

Claim to verify
Blackfoot Confederacy
Joe Hipp, Heavyweight boxer, the first Native American to compete for the WBA World Heavyweight Title.[73]
73. "Blackfoot Culture and History".
Existing citation
Failed verification
A
Blackfoot Culture and History
Ancient history is interesting and important, but the Blackfoot Indians are still here today, too, and we try to feature modern writers as well as traditional folklore, contemporary art as well as museum pieces, and the life and struggles of today as well as the tragedies of yesterday.
Sphere Retrieval Engine
Generative query expansion
Query encoder
Sparse Index
Dense Index
B
C
D
E
A
Verification Engine
Final Rank
1 C
2 D
3 A
Sphere corpus
CommonCrawl snapshot
134M web articles
906.3M passages
Suggested citation
C
Warrior in the ring: Boxing with Marvin Camel
In 1989 at the twilight of his career, Camel fought Joe Hipp of the Blackfeet Nation. Hipp, who became the first Native American to challenge for the world heavyweight championship, said the fight was one of the weirdest of his career.

Figure 1: The decision flow of SIDE from a claim on Wikipedia to a suggestion for a new citation is as follows: (1) the claim is sent to the *Sphere Retrieval Engine* which produces a list of potential candidate documents from the *Sphere corpus*; (2) the *verification engine* ranks the candidate documents and the original citation w.r.t. the claim; (3) if the original citation is not ranked above the candidate documents, then a new citation from the retrieved candidates is suggested. Note that the score of the *verification engine* can be indicative of a potential *failed verification*, as the one reported in the example.

Defined as above, verification of Wikipedia claims requires deep understanding of language and mastery of online search. To what extent can machines learn this behaviour? This question is important from the perspective of progress in fundamental AI. For example, verification requires the ability to detect logical entailment in natural language and to convert claims and their context to the best search term for finding evidence—two long-standing problems that have been primarily investigated in somewhat synthetic settings (Bowman et al., 2015; Wang et al., 2018; Camburu et al., 2018; Nie et al., 2019; Pérez-Rosas et al., 2017; Thorne et al., 2018; Thorne & Vlachos, 2018). It is equally important from a practical perspective. A machine verifier can assist Wikipedia editors by both flagging what citations might trigger failed verifications and suggesting what to replace citations with in case they currently do not support their respective claim. This can be significant: searching potential evidence and carefully reading the search results requires time and high cognitive effort. Integrating an AI assistant into this process could help to reduce both.

In this work we develop SIDE, an AI-based Wikipedia citation verifier. SIDE finds claims on Wikipedia that likely cannot be verified given the current citation, and for such, scans a web snapshot for an alternative. Its behaviour is learnt using Wikipedia itself: using a carefully curated corpus of Wikipedia claims and their current citations, we train a) a retriever component that converts claims and contexts into symbolic and neural search queries optimised to find candidate citations in a web-scale corpus; and b) a verification model that ranks existing and retrieved citations according to how likely they might verify a given claim.

We evaluate our model using both automatic metrics and human annotations. To measure the accuracy of our system automatically, we check how well SIDE recovers existing Wikipedia citations *in high quality articles* as defined by the Wikipedia featured article class. We find that in nearly 50% of the cases, SIDE returns exactly the source that is used in Wikipedia as its top solution. Notably, this does not mean the other 50% are wrong but they are not what Wikipedia is currently using as a source.

We also test SIDE's ability to be a citation assistant. In a user study we present existing Wikipedia citations next to the ones that SIDE produces. These users then assess to what extent the presented citations support the claim, and which citation—from SIDE or Wikipedia—would be better for verification. Overall, more than 60% of the time users prefer SIDE citations over Wikipedia's ones, and this percentage grows above 80% for cases in which SIDE associates a very low verification score to the Wikipedia citation.

System Architecture

In Figure 1, we provide a high level overview of SIDE that shows an example of the decision flow given a Wikipedia claim. In the following, we briefly describe all major components of the system and how they interact with one another. Note that we use the term *claim* to refer to the sentence (or clause) preceding a Wikipedia citation, but any given sentence can contain a multitude of logical claims, and the claim's meaning might depend on its context. The cited documents are represented as a list of passages, i.e., chunks of text with a fixed number of words.

The Retrieval Engine

Given a claim tagged as *failed verification* by a human editor, or flagged by our *verification engine*, SIDE needs to retrieve a list of documents that support the claim. A human verifier would do so by 1) synthesizing a search query based on the claim's context; and 2) executing this query against a search engine. Fundamentally, SIDE *learns* to do the same, using both sparse and dense retrieval sub-systems that we explain in more detail below. The claim's context is represented using the sentences preceding the citation, as well as the section title and the title of the enclosing Wikipedia article. We use *Sphere* (Piktus et al., 2021), a web-scale corpus and search infrastructure for web-scale data, as a source of candidate web pages. Classic sparse and neural dense approaches are known to have complementary strengths (Mao et al., 2020) and hence we merge their results to produce the final list of recommended evidence.

The *sparse retrieval* sub-system uses a seq2seq model (Lewis et al., 2019; Mao et al., 2020) to translate the citation context into query *text*, and then matches the resulting query—a sparse bag-of-words vector—on a BM25 index (Robertson et al., 1995; Baeza-Yates et al., 1999; Manning et al., 2008; Robertson & Zaragoza, 2009; Lin et al., 2021) of Sphere. We train the seq2seq model using data from Wikipedia itself: the target queries are set to be web page titles of existing Wikipedia citations. In practice, we enrich the generated queries with the sentence preceding the citation and the Wikipedia title. The *dense retrieval* sub-system is a neural network which learns from Wikipedia data to encode the citation context into a dense query vector (Wu et al., 2019; Karpukhin et al., 2020; Maillard et al., 2021; Oğuz et al., 2021; Luan et al., 2021). This vector is then matched against the vector encodings of all passages in *Sphere* and the closest ones are returned. The context and passage encoders are trained such that the context and passage vectors of existing Wikipedia citation and evidence pairs are maximally similar (Karpukhin et al., 2020).

The Verification Engine

Given a claim and possible evidence document, either existing on Wikipedia or proposed by the retrieval engine, a human would carefully evaluate to what extent the claim is supported by the provided evidence. This is the role played by our *verification engine*, a neural network taking the claim and a document as input, and predicting how well it supports the claim. Due to efficiency reasons, it operates on a per passage level and calculates the verification score of a document as the maximum over its per-passage scores. The verification scores are calculated by a fine-tuned BERT (Devlin et al., 2019) transformer that uses the concatenated claim and passage as input. This architecture is akin to prior work for textual entailment in natural language inference (MacCartney & Manning, 2008), i.e., testing whether a particular premise supports or contradicts a hypothesis.

The *verification engine* is optimised to rank claim-document pairs in order of verifiability rather than making verified versus failed-verification decisions. This is motivated by the way we envision SIDE's usage in practical setting: we want to prioritise *existing claims* for humans to check by starting with those that are *less likely* supported by their current evidence, and to highlight *recommended evidence* for a given claim by starting with documents that are *more likely* to support the claim. To train the *verification engine* model, we use a training objective that rewards models when they rank existing Wikipedia evidence higher than evidence returned by our retrieval engine. Assuming that some existing Wikipedia evidence is of poorer quality—a core motivation behind this work— even though this training signal could be noisy, we found that, on average, it still provides a meaningful signal. We test this empirically further in the next section.

Evaluation and results

Evaluating the performance of our system is challenging because we cannot be certain that existing citations are always accurate and because of the lack of annotations for citations that fail verification. Therefore, we first evaluate the components of our system in isolation by addressing the following two questions: 1) given a Wikipedia claim, can our retrieval solutions surface the existing citation source from more than 100M web articles? and 2) Is our *verification engine* able to assign low scores to citations marked as failing verification in Wikipedia? After investigating these two questions, we conduct a large scale human annotation campaign to evaluate the overall system.

Experimental Data and Setting

We collect WAFER, a large scale dataset of English Wikipedia inline citations ($\approx 3.8M$ instances - see table 3 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 on the article level—not on the citation level—to avoid potential test leakage into the training data.

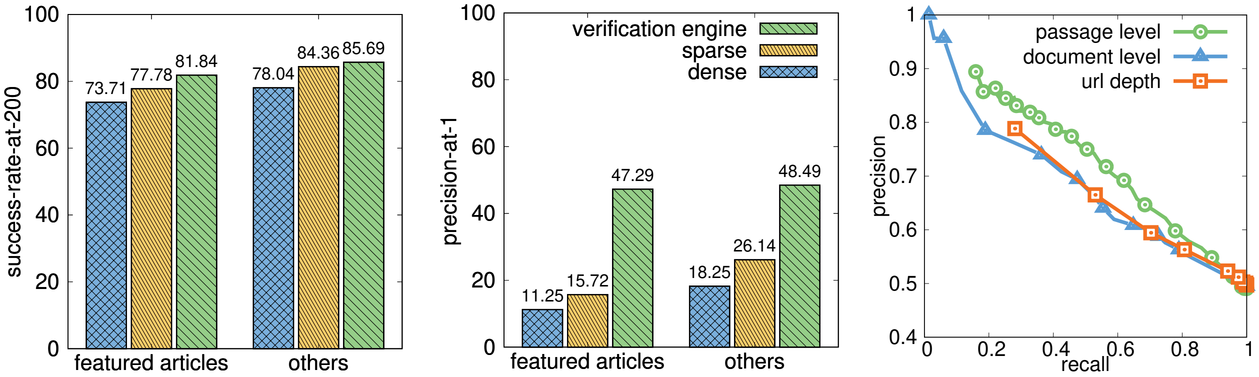
Both the Wikipedia snapshot we consider (*i.e.*, from KILT (Petroni et al., 2021)) as well as the web snapshot (*i.e.*, a CCNet (Wenzek et al., 2019) dump from Sphere (Piktus et al., 2021) which contains 134M web articles, split into 906.3M passages) are from August 2019. We use Sphere's web snapshot as the corpus for retrieval. Aligning the citations in the Wikipedia snapshot and Sphere's web snapshot leads to ≈ 250k retrievable citations. From those we sample ≈ 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 ≈ 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* (articles, 2022) 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 (≈ 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 include in these datasets citations marked with a *failed verification* template (verification, 2022), 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.

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.

Retrieval evaluation

We report our results in Figure 2. We note that the sparse retrieval solution outperforms the dense approach for retrieval from the web, which is consistent with previous observations (Piktus et al., 2021). However, we obtain our best overall SR@200 by combining 100 results from each given they are are highly complementary (Mao et al., 2020) (see Figure 2a) — this ensemble is what we use to retrieve passages to feed into the *verification engine* component. Notably, the *verification engine* component surfaces the original citation document in the highest-ranked position nearly 50% of the time (see Figure 2b). However, these numbers



(a) Percentage of times our retrievers can surface the gold source among the top-200 results, for citations in featured and other Wikipedia articles. The *verification engine* bar (*i.e.*, green) combines sparse and dense retrievers, 100 passages each.

(b) Accuracy in surfacing the gold source in first position, for citations in featured and other articles. The *verification engine* (*i.e.*, green bar) takes as input a combination of 100 passages from the sparse and 100 from the dense retriever and reranks those.

(c) Precision versus recall in detecting citations marked as *failed verification* against citations in *featured* articles. We compare a passage versus a document-level approach for the *verification engine* and a baseline using the depth of the cited url.

Figure 2: Automatic evaluation of Side components on the WAFER test set.

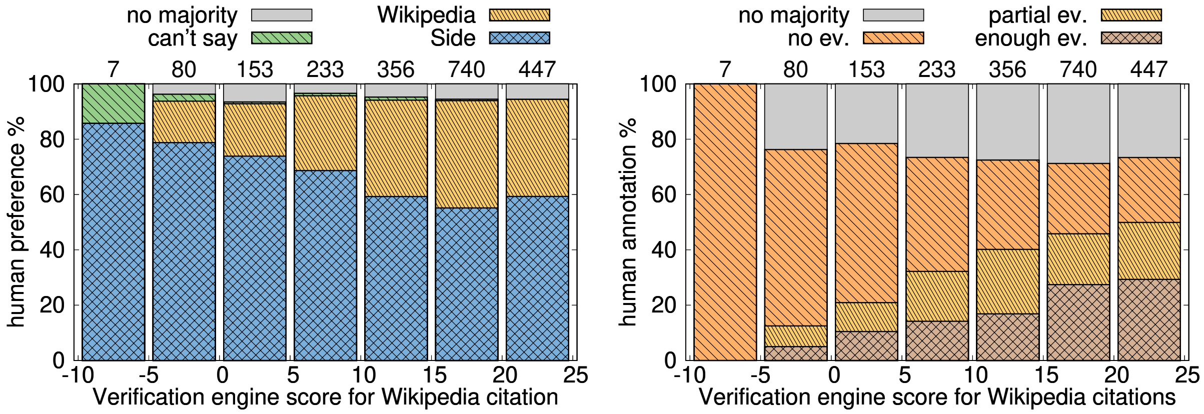
have to be interpreted in the context of our background corpus, i.e., despite containing ≈ 900M passages from more than 100M documents, it can only approximate a real-world scenario where evidences are to be sought on the open web.

In general, retrieving evidence for claims in featured articles is more challenging than for other claims in Wikipedia, e.g., we observe a large difference of -7.0%/-10.4% P@1 (for dense/sparse) between featured and non-featured articles. One hypothesis for this is that there exists an intrinsic popularity bias associated with featured content. Featured content might often correlate with popular topics, which in turn means that more sources on the web contain relevant information. In contrast, claims in more niche articles have much less coverage on the web and therefore are easier to find. Another factor is that featured articles are typically edited a lot more frequently, which is how they achieved their high quality, which in turn also could lead greater deviation from the original phrasing of the cited source. Assuming that dense retrievers are better at recognising paraphrases, we would expect a smaller increase in performance between featured and non-featured for dense vs. sparse, which is indeed the case.

The *verification engine* model considerably boosts the accuracy of the retrieval component and almost levels the gap for featured articles, suggesting greater ability to identify evidence, even among a large set of relevant documents. This performance can be explained by its ability to leverage fine-grained language comprehension, when the model can directly compare the contents of the two texts using a cross-attention mechanism to overcome the representational decomposability gap suffered by the retrievers (Seo et al., 2019). Another relevant factor is that simple, helpful indicators like quoted phrases from the cited source seem to be easier to detect in token-level comparison.

Detecting Failed Verification

Our goal in this analysis is to measure to which degree the score of the *verification engine* can be used to detect whether a citation fails verification. To this aim, we rank the union of test citations in featured articles and test *failed verification* citations. An ideal system would place all failed verification at the bottom end of the ranked list and featured citations at the top. To compute the rank, we consider two different instantiations of the *verification engine*, that operate either at a passage or document level. As many failed citations include a link to an over-generic URL (*e.g.*, a generic newspaper website instead of a specific page covering the claim), we include a simple baseline based on the depth of a source URL (*i.e.*, the number of



(a) Crowd annotators preference for citations suggested by Side versus those on Wikipedia for a given claim, without knowing their identity. Fleiss' kappa Inter-Annotator Agreement = 0.2.

(b) Evidence annotations for Wikipedia citations: (1) *enough* to verity the claim; (2) the claim is only *partially* verified; (3) *no evidence*. Fleiss' kappa Inter-Annotator Agreement = 0.11.

Figure 3: Crowd annotator evaluation for 2016 claims in the WAFER test set for which Side produces a citation with higher evidence score than the existing Wikipedia citation. We collect 5 annotations per claim and report majority voting results, bucketed according to the *verification engine* score associated with the existing Wikipedia citation (bucket size reported on top).

elements in an URL path). In the passage-level solution, we independently compute a score for each passage in a document with the *verification engine* and rank citations according to the maximum score. For the document-level approach, we feed as much text as possible (*i.e.*, on average the first 2 or 3 passages) for the source document as input to a seq2seq model (Lewis et al., 2019) and use the prediction score for the ranking.

The resulting precision-recall curve is in Figure 2c. Overall, the passage-level *verification engine* performs very well; if we only consider a conservative Recall of 15%, for instance, ≈ 90% are *failed verification* citations. Notably, these results are achieved without any explicit supervision on failed verification instances, given that the *verification engine* is trained only on positive examples. A document-level approach leads to worse results (*i.e.*, ≈ 80% precision at 15% recall), mainly due to the impossibility of considering the whole document (given model architectural constraints on maximum input size). Considering url depth turns out to be a remarkably solid baseline. To further investigate this aspect, we study the distribution of depths for urls in our data (see Figure 7) and find that citations in featured articles tend to be deep (*i.e.*, very specific urls) while citations marked as failed verification are usually shallow (*i.e.*, very generic urls).

Evaluation of the final system

To test the performance of our final system, we perform a two-stage human assessment: (1) a large scale crowd annotation campaign followed by (2) a smaller scale fine-grained evaluation. First, we select claims in the *test set* for which Side outputs a citation source with a higher score than what is currently on Wikipedia. We then ask crowd annotators to express their preference on which one of the two (i.e., Side's suggested citation or Wikipedia's one) better supports a given claim. Additionally, we ask them to assess if a source contains *enough evidence* to support the claim, *partial evidence* (meaning that only parts of the claim are supported by the source), or *no evidence* whatsoever. To keep the annotation load tractable, we use our *verification engine* component to select a single passage from each source, making sure to consider overlapping passages for Wikipedia sources so as to avoid cutting evidentiary sentences.

Results are reported in Figure 3. We note that both preferences for Side's suggested source (*i.e.*, Figure 3a) and Wikipedia evidence annotations (*i.e.*, Figure 3b) are proportional to the ranker score associated

No evidence	41.3%
Partial evidence	18.2%
Full evidence in one passage	16.7%
Full evidence in multiple passages	13.5%
Evidence not in crawled text (e.g., multimedia)	7.1%
Pay wall access	3.2%

Table 1: Fine-grained human annotations for Wikipedia citations for which crowd annotators indicate no evidence for a total of 136 instances.

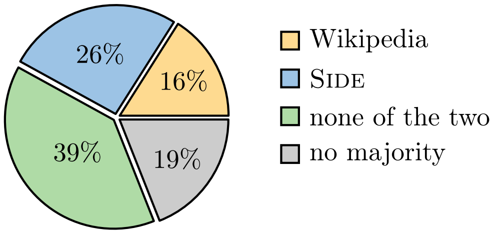


Figure 4: Annotations of Wikipedia authenticated users via our demo. P value = 0.0178.

to the existing Wikipedia citation—the lower the score the more preferences for Side and the less evidence found within Wikipedia. These results suggest that the ranker score might be a valid proxy for the presence (or absence) of evidence in a citation, and might help in surfacing cases that require attention from Wikipedia editors. To verify the noise introduced by automatically selecting a single passage for each source, we conduct a control study on more than 500 sources where we ask annotators if they prefer the selected passage (*i.e.*, the top scored) with respect to a random one within the source. We find that for over 80% of the cases annotators prefer the selected passage, with an Inter-Annotator Agreement of 0.27 (Fleiss' κ). Finally, to validate crowd annotators accuracy, we annotate more than 100 cases where evidence was not found in the Wikipedia citations. We find (see Table 1 for the complete picture) that sometimes the evidence is in the source but not within the crawled text (*e.g.*, multimedia content); other times, it is spread across multiple passages (which the current system can't detect, but that we plan to tackle in future work). Overall, more than 40% of the time no evidence can be found in the reference to verify a Wikipedia claim.

Finally, we build a demo of Side and engage with the English-speaking Wikipedia community, asking users if they would use the citation already present on Wikipedia, the top-1 citation suggested by Side or none of the two to verify a given claim. We do not reveal the source of a citation in the user interface (i.e., Wikipedia or Side), select claim-citation pairs on Wikipedia that are likely to fail verification (i.e., with a verifier score below 0) and allow access to the full text for each citation (instead of a single passage). Results (see Figure 4) reveal that Side can indeed select claim-citation pairs that fail verification — users selected the Wikipedia citation in only 16% of cases, compared to the 65% of citations where either Side's recommendation or neither of the two were preferred. Moreover 26% of the times Side can provide a top-1 recommendation that is judged appropriate by the community. We additionally conduct a sign test between Side and Wikipedia preferences resulting in a P value of 0.0178 and two-tail P value of 0.0357. So far 43 authenticated Wikipedia users[2] participated to our study, for a total of 106 annotations, with an average of 1.8 annotations per claim. We plan to keep collecting annotations through our demo and update these figures in future iterations of the paper.

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 (Reliability, 2022), 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 (Gonçalves, 2022), 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 (Redi et al., 2019; Chou et al., 2020) have received a lot of attention from both the scientific and the Wikimedia communities. We believe Side can be

[2] We exclude annotations performed by the authors of this paper.

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 (Piccardi et al., 2020; Lewoniewski et al., 2020; Piccardi et al., 2021; Kaffee & Elsahar, 2021; Zagorova et al., 2022) 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 (Jana et al., 2018) to citation span detection (Fetahu et al., 2017) amongst other efforts. More broadly, citation retrieval and paper/source recommendation have also received attention in the scientific literature domain for many decades (McNee et al., 2002; Ren et al., 2014; Bhagavatula et al., 2018; Chou et al., 2020), 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 Färber & Jatowt (2020) for a recent comprehensive review.

Several works have investigated the ability of AI to generate missing Wikipedia articles from scratch (Liu et al., 2018; Prabhumoye et al., 2019; Fan & Gardent, 2022; Kaffee et al., 2022). There exists AI tools, such as Scribe (2022), 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 (Thorne & Vlachos, 2018; Thorne et al., 2018, 2019; Schuster et al., 2021; Trokhymovych & Saez-Trumper, 2022). 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.

Discussion

We introduce Side, an AI-based system for improving the quality and verifiability of Wikipedia citations. Building on recent advances in natural language processing, we demonstrate that machines can help humans finding better citations, a task requiring understanding of language, and mastery of online search. While previous works (Bowman et al., 2015; Wang et al., 2018; Camburu et al., 2018; Nie et al., 2019; Pérez-Rosas et al., 2017; Thorne et al., 2018; Thorne & Vlachos, 2018) have shown the ability of large neural networks to perform well on natural language understanding tasks, these results were mostly obtained for well specified tasks, on synthetic datasets specifically created for evaluating AI systems. Here we show similar results in a real world scenario, implying noisier data and a more loosely defined task.

While our results are promising, and we believe our system could already be used to improve Wikipedia, there exist a variety of future research directions that can be pursued. For instance, we only considered references corresponding to web pages, but Wikipedia also cites books, scientific articles and other kind of documents. These include other modalities than just text, such as images and videos. To fully assess the quality of Wikipedia references, Side needs to become multi-modal. Second, our system currently only supports the English language, while Wikipedia exists for more than two hundreds languages. Making Side multi-lingual raises interesting research questions, such as the capabilities of performing *cross-lingual* citation improvements: given a claim in one language, if the system cannot find good evidence in that particular language, can it find references in other languages?

Finally, our work currently assumes that Wikipedia claims are verifiable, and only improves the quality of the references for existing claims. A natural extension of our work would be to detect claims that are not verifiable, and flag them for review by human editors. This comes with challenges, as a way to show that a claim is unverifiable is to find a contradicting evidence. Unfortunately, Wikipedia currently does not contain such information, and thus training AI-based systems to perform this task is not straightforward. However, we believe that Side could be a first step towards surfacing unverifiable claims: if Side cannot find good evidence for a claim, it might be impossible to verify. We report one example of such claims in the Appendix (Table 2), showing that a lack of good evidence from Side could be an indication of unverifiability.

We release all data, code and models described in this paper. We hope that this work could be used in a broader context than just Wikipedia, for example helping humans to perform fact-checking. More generally, we believe that this work could lead to more trustworthy information online.

Acknowledgement

The authors would like to greatly thank Miriam Redi and Diego Saez-Trumper from the Wikimedia Foundation for their invaluable help and support throughout the project; Ross Nkama, for helping set up our annotation interface; and all Wikipedia users who helped us evaluate Side.

References

Featured articles. Wikipedia. `https://en.wikipedia.org/wiki/Wikipedia:Featured_articles`, 2022.

Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.

Chandra Bhagavatula, Sergey Feldman, Russell Power, and Waleed Ammar. Content-based citation recommendation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 238–251, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1022. URL `https://aclanthology.org/N18-1022`.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-snli: Natural language inference with natural language explanations. *arXiv preprint arXiv:1812.01193*, 2018.

Ai-Jou Chou, Guilherme Gonçalves, Sam Walton, and Miriam Redi. Citation detective: a public dataset to improve and quantify wikipedia citation quality at scale. Wiki-Workshop, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://www.aclweb.org/anthology/N19-1423`.

Angela Fan and Claire Gardent. Generating full length wikipedia biographies: The impact of gender bias on the retrieval-based generation of women biographies. *ACL*, 2022.

Michael Färber and Adam Jatowt. Citation recommendation: approaches and datasets. *International Journal on Digital Libraries*, 21(4):375–405, August 2020. doi: 10.1007/s00799-020-00288-2. URL `https://doi.org/10.1007/s00799-020-00288-2`.

Besnik Fetahu, Katja Markert, and Avishek Anand. Fine grained citation span for references in Wikipedia. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1990–1999, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653 /v1/D17-1212. URL `https://aclanthology.org/D17-1212`.

Guilherme Gonçalves. Citation Hunt. `https://citationhunt.toolforge.org`, 2022.

Abhik Jana, Pranjal Kanojiya, Pawan Goyal, and Animesh Mukherjee. WikiRef: Wikilinks as a route to recommending appropriate references for scientific Wikipedia pages. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 379–389, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL `https://aclanthology.org/C18-1032`.

Lucie-Aimée Kaffee and Hady Elsahar. References in wikipedia: The editors' perspective. In *Companion Proceedings of the Web Conference 2021*, pp. 535–538, 2021.

Lucie-Aimée Kaffee, Pavlos Vougiouklis, and Elena Simperl. Using natural language generation to bootstrap missing wikipedia articles: A human-centric perspective. *Semantic Web*, 13(2):163–194, 2022.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*, 2020.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *ArXiv*, abs/1910.13461, 2019.

Włodzimierz Lewoniewski, Krzysztof Węcel, and Witold Abramowicz. Modeling popularity and reliability of sources in multilingual wikipedia. *Information*, 11(5):263, 2020.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. Pyserini: An easy-to-use python toolkit to support replicable ir research with sparse and dense representations, 2021.

Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*, 2018.

Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. Sparse, Dense, and Attentional Representations for Text Retrieval. *Transactions of the Association for Computational Linguistics*, 9:329–345, 04 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00369. URL https://doi.org/10.1162/tacl_a_00369.

Bill MacCartney and Christopher D. Manning. Modeling semantic containment and exclusion in natural language inference. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pp. 521–528, Manchester, UK, August 2008. Coling 2008 Organizing Committee. URL https://aclanthology.org/C08-1066.

Jean Maillard, Vladimir Karpukhin, Fabio Petroni, Wen-tau Yih, Barlas Oguz, Veselin Stoyanov, and Gargi Ghosh. Multi-task retrieval for knowledge-intensive tasks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1098–1111, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.89. URL https://aclanthology.org/2021.acl-long.89.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge, 2008.

Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. Generation-augmented retrieval for open-domain question answering. *arXiv preprint arXiv:2009.08553*, 2020.

Sean M. McNee, Istvan Albert, Dan Cosley, Prateep Gopalkrishnan, Shyong K. Lam, Al Mamunur Rashid, Joseph A. Konstan, and John Riedl. On the recommending of citations for research papers. In *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work*, CSCW '02, pp. 116–125, New York, NY, USA, 2002. Association for Computing Machinery. ISBN 1581135602. doi: 10.1145/587078.587096. URL https://doi.org/10.1145/587078.587096.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*, 2019.

Barlas Oğuz, Kushal Lakhotia, Anchit Gupta, Patrick Lewis, Vladimir Karpukhin, Aleksandra Piktus, Xilun Chen, Sebastian Riedel, Wen tau Yih, Sonal Gupta, and Yashar Mehdad. Domain-matched pre-training tasks for dense retrieval, 2021.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic detection of fake news. *arXiv preprint arXiv:1708.07104*, 2017.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2523–2544, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.200. URL https://aclanthology.org/2021.naacl-main.200.

Tiziano Piccardi, Miriam Redi, Giovanni Colavizza, and Robert West. Quantifying engagement with citations on wikipedia. In *Proceedings of The Web Conference 2020*, pp. 2365–2376, 2020.

Tiziano Piccardi, Miriam Redi, Giovanni Colavizza, and Robert West. On the value of wikipedia as a gateway to the web. In *Proceedings of the Web Conference 2021*, pp. 249–260, 2021.

Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Dmytro Okhonko, Samuel Broscheit, Gautier Izacard, Patrick Lewis, Barlas Oguz, Edouard Grave, Wen-tau Yih, and Sebastian Riedel. The web is your oyster - knowledge-intensive NLP against a very large web corpus. *CoRR*, abs/2112.09924, 2021. URL https://arxiv.org/abs/2112.09924.

Shrimai Prabhumoye, Chris Quirk, and Michel Galley. Towards content transfer through grounded text generation. *arXiv preprint arXiv:1905.05293*, 2019.

Top Websites Ranking. similarweb. https://www.similarweb.com/top-websites/, 2022.

Miriam Redi, Besnik Fetahu, Jonathan Morgan, and Dario Taraborelli. Citation needed: A taxonomy and algorithmic assessment of wikipedia's verifiability. In *The World Wide Web Conference*, pp. 1567–1578, 2019.

WikiProject Reliability. Wikipedia. https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Reliability, 2022.

Xiang Ren, Jialu Liu, Xiao Yu, Urvashi Khandelwal, Quanquan Gu, Lidan Wang, and Jiawei Han. Cluscite: Effective citation recommendation by information network-based clustering. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pp. 821–830, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450329569. doi: 10.1145/2623330.2623630. URL https://doi.org/10.1145/2623330.2623630.

Stephen Robertson and Hugo Zaragoza. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc, 2009.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. Okapi at trec-3. *Nist Special Publication Sp*, 109:109, 1995.

Tal Schuster, Adam Fisch, and Regina Barzilay. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 624–643, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.52. URL https://aclanthology.org/2021.naacl-main.52.

Scribe. Wikimedia. https://meta.wikimedia.org/wiki/Scribe, 2022.

Minjoon Seo, Jinhyuk Lee, Tom Kwiatkowski, Ankur Parikh, Ali Farhadi, and Hannaneh Hajishirzi. Real-time open-domain question answering with dense-sparse phrase index. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4430–4441, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1436. URL https://aclanthology.org/P19-1436.

James Thorne and Andreas Vlachos. Automated fact checking: Task formulations, methods and future directions. *arXiv preprint arXiv:1806.07687*, 2018.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*, 2018.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. The fever2. 0 shared task. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pp. 1–6, 2019.

Mykola Trokhymovych and Diego Saez-Trumper. Wikifactfind: Semi-automated fact-checking based on wikipedia. 2022.

Verifiability. Wikipedia. `https://en.wikipedia.org/wiki/Wikipedia:Verifiability`, 2022.

Failed verification. Wikipedia. `https://en.wikipedia.org/wiki/Template:Failed_verification`, 2022.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. Ccnet: Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv:1911.00359*, 2019.

Wikimedia. Statistics. `https://stats.wikimedia.org/#/all-projects/reading/total-page-views/normal|bar|2-year|~total|monthly`, 2022.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. Zero-shot entity linking with dense entity retrieval. *arXiv preprint arXiv:1911.03814*, 2019.

Olga Zagorova, Roberto Ulloa, Katrin Weller, and Fabian Flöck. "i updated the": The evolution of references in the english wikipedia and the implications for altmetrics. *Quantitative Science Studies*, 3(1):147–173, 2022.

Additional information

Sphere retrieval

Sparse retriever with generative query expansion Sparse retrieval methods rank documents by weighted lexical overlap and represent queries and documents as high-dimensional *sparse* vectors with dimensions corresponding to vocabulary terms. BM25 is by nature very successful in retrieving passages that require high lexical overlap, also for long tail names and words. The disadvantage for BM25 in this setting is that we do not know where the claim in the text in front of the citation is located, it could be just a short span of text, or the claim could be fragmented over multiple sentences and require references to the context of the Wikipedia article. Indeed, in a manual evaluation of a small sample (30 instances) we found that roughly 1/3 of the sentences had some kind of co-reference which was crucial for understanding the claim. Empirically we found that only using the first sentence in front of the claim and also adding the Wikipedia article's title to the query did yield the best BM25 results.

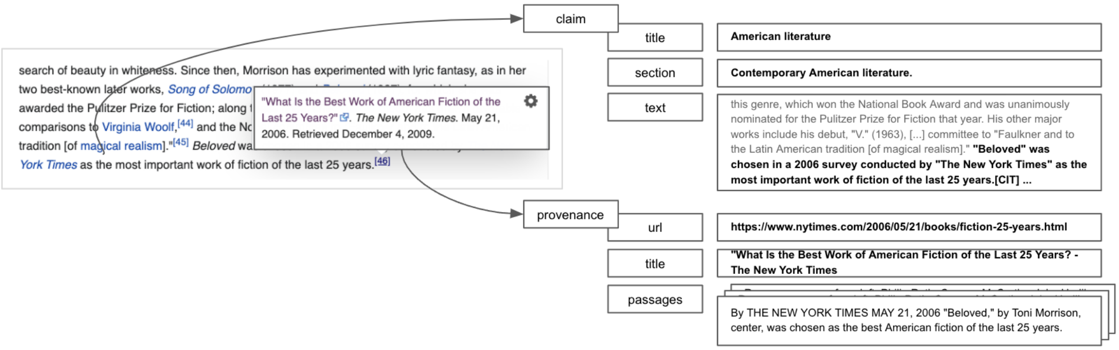
Dense retriever DPR is a method that learns to embed queries and documents as low-dimensional *dense* vectors. The basic building block of DPR is a BERT-like neural encoder, that consumes a sequence of tokens and predicts one dense vector. DPR consists of two such neural encoders, one for the query and one for a document's passage. DPR is then trained on a dataset with instances consisting of (query, correct document) tuples. The training objective is to maximize the inner product between the query vector and the passage vectors of a correct document, and to minimize the inner product for incorrect documents. In contrast to BM25, DPR can learn which parts of the text are likely the important elements. Another advantage is that DPR is typically stronger in retrieving passages with rephrased versions of the claim.

Training Many components of our system, such as the dense retriever and the *verification engine*, are based on neural networks requiring examples to be trained. We propose to leverage the scale of Wikipedia, and its millions of existing citations, to build a training set for our models. It should be noted that the obtained data is noisy, as existing citations might be failing verification, and determining if it could be used to train our system is an interesting research question. Moreover, our system processes references at the passage level, while our training data corresponds to pairs of claims and documents. Thus, we train the retriever and the *verification engine* using an expectation-maximization algorithm, modeling the passsage containing the evidence as a latent variable. Finally, our data only contains *positive* examples of claims and references. A standard solution for training retrievers is to mine *negative* examples, and we follow this approach here. While this work well for training retrievers, it is unclear how well this supervision would work for training the *verification engine*, and in particular, to determine if an *existing* reference is failing verification for a particular claims. Indeed, the problem of ranking a set of candidate documents for a particular claim is different from ranking existing pairs of documents and claims.

Evaluation details

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

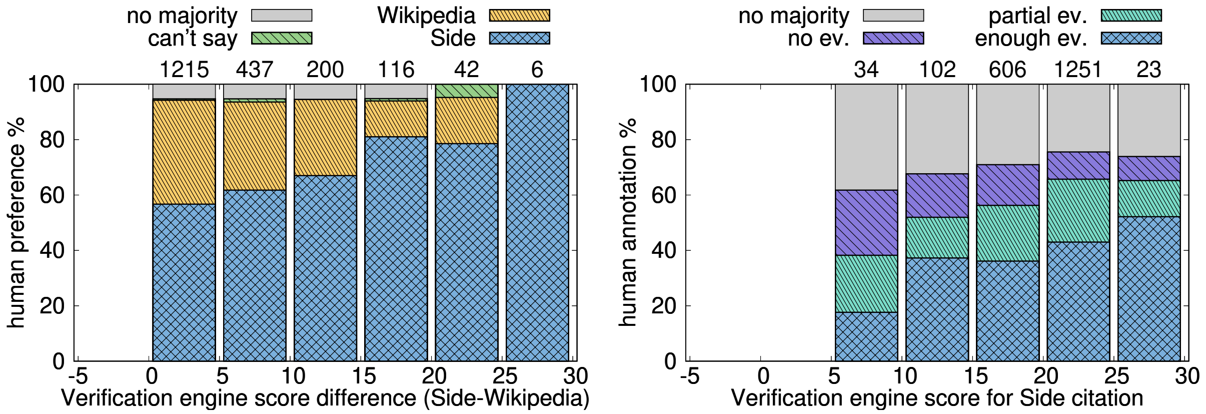


search of beauty in whiteness. Since then, Morrison has experimented with lyric fantasy, as in her two best-known later works, *Song of Solomo* awarded the Pulitzer Prize for Fiction; along t comparisons to Virginia Woolf,[44] and the Nc tradition [of magical realism]."[45] *Beloved* wa *York Times* as the most important work of fiction of the last 25 years.[46]

"What Is the Best Work of American Fiction of the Last 25 Years?". *The New York Times*. May 21, 2006. Retrieved December 4, 2009.

claim	
title	**American literature**
section	**Contemporary American literature.**
text	this genre, which won the National Book Award and was unanimously nominated for the Pulitzer Prize for Fiction that year. His other major works include his debut, "V." (1963), [...] committee to "Faulkner and to the Latin American tradition [of magical realism]." **"Beloved" was chosen in a 2006 survey conducted by "The New York Times" as the most important work of fiction of the last 25 years.[CIT] ...**
provenance	
url	**https://www.nytimes.com/2006/05/21/books/fiction-25-years.html**
title	**"What Is the Best Work of American Fiction of the Last 25 Years? - The New York Times**
passages	By THE NEW YORK TIMES MAY 21, 2006 "Beloved," by Toni Morrison, center, was chosen as the best American fiction of the last 25 years.

Figure 5: Example citation from the WAFER dataset.



(a) Crowd annotators preference for citations suggested by our System versus those present in Wikipedia for a given claim. Fleiss' kappa Inter-Annotator Agreement = 0.2.

(b) Evidence annotations for Side citations: (1) *enough* to verity the claim; (2) the claim is only *partially* verified; (3) *no evidence*. Fleiss' kappa Inter-Annotator Agreement = 0.09.

Figure 6: Crowd annotators evaluation for 2016 claims in the WAFER test set for which Side produces a citation with higher evidence score that the existing Wikipedia citation. We collects 5 annotations per claim and report majority voting results, bucketed according to the evidence ranker score (bucket size reported on top).

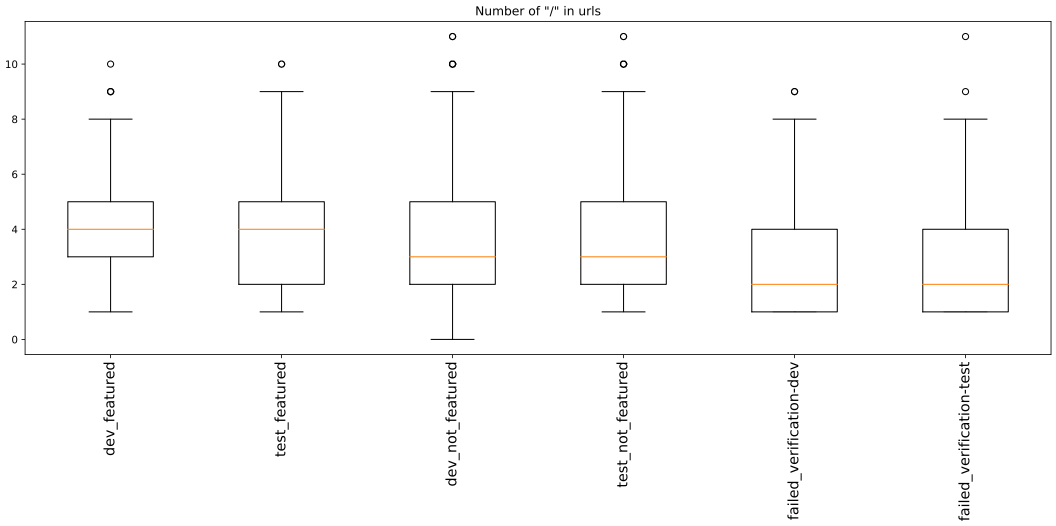


Figure 7: Url depth analysis.

	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 2: 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.

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

Table 3: Statistics for the WAFER dataset.

	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 4: WAFER test results.