Generating Full Length Wikipedia Biographies The Impact of Gender Bias on the Retrieval-Based Generation of Women Biographies

Angela Fan
FAIR / LORIA
Université de Lorraine
angelafan@fb.com

Claire Gardent
CNRS/LORIA
Nancy, France
claire.gardent@loria.fr

Abstract

Generating factual, long-form text such as Wikipedia articles raises three key challenges: how to gather relevant evidence, how to structure information into well-formed text, and how to ensure that the generated text is factually correct. We address these by developing a model for English text that uses a retrieval mechanism to identify relevant supporting information on the web and a cache-based pre-trained encoder-decoder to generate long-form biographies section by section, including citation information. To assess the impact of available web evidence on the output text, we compare the performance of our approach when generating biographies about women (for which less information is available on the web) vs. biographies generally. To this end, we curate a dataset of 1,500 biographies about women. We analyze our generated text to understand how differences in available web evidence data affect generation. We evaluate the factuality, fluency, and quality of the generated texts using automatic metrics and human evaluation. We hope that these techniques can be used as a starting point for human writers, to aid in reducing the complexity inherent in the creation of long-form, factual text.

1 Introduction

Wikipedia has become one of the major sources of dissemination of knowledge across the globe. However, the knowledge contained in Wikipedia is not neutral — it is biased in various ways (Hinnosaar, 2019; Schmahl et al., 2020). Many studies, including those from the Wikimedia Foundation itself, have emphasized that biographies in particular are overwhelmingly written about men. This leads to many subtle yet far-reaching effects, from students not writing their first book reports on a woman to bias in models trained on Wikipedia, as Wikipedia has long been used as a source of data. Many existing efforts, such as the Wikipedia Women in Red project, focus on encouraging article creation to mitigate this gender gap. However, Wikipedia articles remain painstakingly written and edited primarily by a network of human contributors. Despite advances in text generation and modeling architectures that retrieve information, the automatic creation of Wikipedia articles is incredibly challenging (Liu et al., 2018). Even the functionality of tools that aid human editors are limited.

In this work, we strive to create a system that could write an entire Wikipedia article in English, focusing on the biography domain. We confront several major challenges. First, this is fundamentally a long-form generation task. Improvements driven by pretraining (Radford et al., 2019; Lewis et al., 2019) have improved generation fluency at the level of multiple sentences. However, Wikipedia biographies contain multiple paragraphs in a structured form with headings, as well as citations to indicate where the information originated from. Second, the task confronts obstacles around the factuality (Elazar et al., 2021) of generated content, as articles must be factually accurate. Third, Wikipedia articles are written using reference material, often found on the web (Piktus et al., 2021). Thus, models need to find and ingest web searches as a pre-requisite to writing accurate biographies.

We develop a method for English Wikipedia that starts with the subject and occupation of the biography, then leverages web search to find relevant evidence. Given search results, we employ a retrieval-augmented generation architecture (Lewis et al., 2020; Guu et al., 2020) based on large-scale pre-training to identify relevant information and write the biography. We generate section by section, using a caching mechanism similar to Transformer-XL (Dai et al., 2019) to reference previous sections and achieve greater document-level context. Finally, after each section, we append a citation based on which web searches were retrieved.

We quantify the quality of generation using several automatic metrics such as ROUGE-L (Lin,

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 8561 - 8576

2004), entailment, and named entity coverage. Further, we study the strong dependency of our method on accurate retrieval, and design a specific evaluation dataset that highlights this challenge. The dataset consists of 1,527 Wikipedia biographies about women, where information on the internet is not as easily retrieved. We use this dataset to analyze the gap between model quality when retrieval is challenging (our novel evaluation dataset with biographies about women) and model quality when retrieval is more accurate (a random set of evaluation biographies). Finally, we conduct a large-scale human evaluation to measure the factuality and coverage of our generated biographies. We hope that our techniques can eventually be used as a starting point for human Wikipedia writers, for biographies and beyond.

2 Related Work

2.1 Generation of Wikipedia Articles

A large body of work in generation utilizes Wikipedia, often for data-to-text tasks that use Wikidata or DBpedia RDF triples (Gardent et al., 2017; Castro Ferreira et al., 2020; Kaffee et al., 2018b; Vougiouklis et al., 2018; Sha et al., 2018; Puduppully et al., 2019; Chen et al., 2020b; Wang et al., 2020; Agarwal et al., 2020; Parikh et al., 2020), as well as graphs (Jin et al., 2020) as input. Some have focused on long text, such as writing summaries (Chen et al., 2020a) or sections of articles (Kaffee et al., 2020), expanding stubs (Banerjee and Mitra, 2015), and writing full articles (Liu et al., 2018). Some of these works utilize structure to learn templates (Sauper and Barzilay, 2009), Markov logic networks (Liu et al., 2010), or word graphs (Banerjee and Mitra, 2015), but we anticipate that pretraining and large neural network based techniques will vastly improve upon this quality.

Closest to our work, Liu et al. (2018) use web evidence to write full length articles, but do not focus on biographies and use extractive summarisation techniques rather than a retrieval mechanism to identify relevant information. Further, their work generates the entire Wikipedia article at once, whereas we demonstrate that breaking down the article to generate section by section is more effective. We also include a mechanism for the model to generate citations, which was not included in existing work. Thus, our model can produce a full-form Wikipedia article that would look like what a human editor wrote. Finally, our work (i) leverages recent advances in large-scale pretraining, which improves generation fluency and (ii) investigates the impact of available web evidence on the generated texts.

Other work has focused on automatic creation of biographies, such as generation from infoboxes (Lebret et al., 2016) or Wikidata (Chisholm et al., 2017), as well as extracting biographical sentences (Biadsy et al., 2008). The majority of existing research focused on short biographies.

2.2 Retrieval in Generative Models

Retrieval mechanisms have been used to support a variety of tasks, including dialogue (Moghe et al., 2018; Dinan et al., 2018; Shuster et al., 2021), fact verification (Thorne et al., 2018), and sentence generation (Guu et al., 2018). Most notably, retrieval has been heavily used in question answering (Chen et al., 2017; Kwiatkowski et al., 2019; Seo et al., 2019; Karpukhin et al., 2020). Recent innovations in incorporating retrieval mechanisms have increased the quality and scale of retrieval-augmented generative methods (Guu et al., 2020; Lewis et al., 2020; Izacard and Grave, 2020).

2.3 Bias in Wikipedia Biographies

Gender bias on Wikipedia is a well-known problem (Hinnosaar, 2019; Dinan et al., 2020; Schmahl et al., 2020), particularly in the case of biographies (Graells-Garrido et al., 2015; Stratigakos, 2016; Luo et al., 2018; Schmahl et al., 2020). This bias is compounded by geographical location, as information about certain areas of the world is far more prevalent (Kaffee et al., 2018a; Beytía, 2020). This bias exists not only in what articles are written, but also in articles targeted for deletion — articles about certain marginalized groups are removed at higher rates (Worku et al., 2020). Wikipedia reflects biases present in society (De-Arteaga et al., 2019; Young et al., 2020; Schmahl et al., 2020), though numerous initiatives exist to de-bias Wikipedia. These range from training programs (Iglesias, 2020) to projects such as Women in Red[1] and WikiProject Women[2]. The success of these initiatives has been studied (Langrock and González-Bailón, 2020) and found to be effective, but not at addressing the systemic challenges that create bias in the first place.

[1] https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Women_in_Red

[2] https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Women

In the natural language processing community, work has focused on combating gender bias in co-reference resolution (Zhao et al., 2018), dialogue (Dinan et al., 2019; Lee et al., 2019; Liu et al., 2020), detection of abusive language (Park et al., 2018), machine translation (Stanovsky et al., 2019), and word embeddings (Gonen and Goldberg, 2019). These works present a variety of strategies, including data augmentation, additional data collection efforts, modified generation, and fair evaluation (Yeo and Chen, 2020). A comprehensive survey can be found in Blodgett et al. (2020). However, most of these efforts are focused on specific tasks or models — our work uniquely targets generation of full Wikipedia biographies to combat gender bias present on Wikipedia.

3 Task

Given a person's name, one or more occupation(s), and CommonCrawl as a source of evidence, the task is to generate a Wikipedia biography and to associate each generated section with adequate bibliographic references. We model this task by generating a biography section by section using section headers as additional information. A special section header called *toplevel* is used as the start of the article. The subsequent headers are automatically generated at the end of each section as input for the next. Thus for each section, the input includes a name, one or more occupations, a section header, and CommonCrawl as a retrieval corpus.

4 Method

Wikipedia biographies begin with an introductory paragraph followed by various subsections[3]. To account for this structure and generate long-form text based on retrieved web evidence, our system, illustrated in Figure 1, generates a biography section by section. Based on the subject, their occupation(s), and the section heading, the model first identifies a subset of relevant evidence from a set of web search results found using that triplet (*retrieval module*). It then conditions upon that evidence to generate the section, using a Sequence-to-Sequence model (*generation module*) which can access previous sections using a caching mechanism. Finally, the model indicates which evidence documents it used and outputs those as citations, mimicking a standard Wikipedia article (*citation module*). We focus on generation in English.

[3]Many biographies contain infoboxes, which we do not generate.

4.1 Retrieval Module

Given a query Q and a set of web documents D retrieved from the web based on this query, the task of the retrieval module is to retrieve the subset of D that is most relevant given Q. The challenge is sifting through the large quantity of potentially useful information.

Query. The query Q consists of three parts: **(1)** the name of the person for which the biography is generated, **(2)** their , possibly multiple, occupation(s), and **(3)** a section heading. Including the occupation narrows the realm of potential relevant content, especially as proper names are often ambiguous (e.g. *Jane Wang*). Similarly, the section header allows the model to retrieve different information for each section (e.g. *Personal Life* compared to *Career*).

Documents. The query Q is put through a search engine to retrieve web hits, which form the set of documents D that are candidates for retrieval. The web results are represented only as text, and all non-text information is discarded.

Retrieval. To retrieve the relevant subset of D, each sentence in D is encoded with RoBERTa base trained with LayerDrop (Fan et al., 2019b; Liu et al., 2019; Devlin et al., 2018). The concatenation of the subject's name, occupation(s), and section header is also encoded. We then calculate the dot product to identify which encoded document sentences are most relevant given the currently encoded query Q, following the strategy used in other retrieval works (Karpukhin et al., 2020). The representation of the top k most relevant sentences are then passed onwards through the model. Note that compared to some other retrieval-augmented generation (Lewis et al., 2020), the RoBERTa encoder is not fixed, so the retrieval module learns based on the performance of the generation module. This is possible because our retrieval is far smaller scale, we limit the search to approximately 40 sentences (1,000 words) that could be used to generate each section.

4.2 Generation Module

To generate the sections we use a Transformer-based Sequence-to-Sequence model initialized with BART-Large (Lewis et al., 2019). The input to BART is the concatenation of the subject's name,

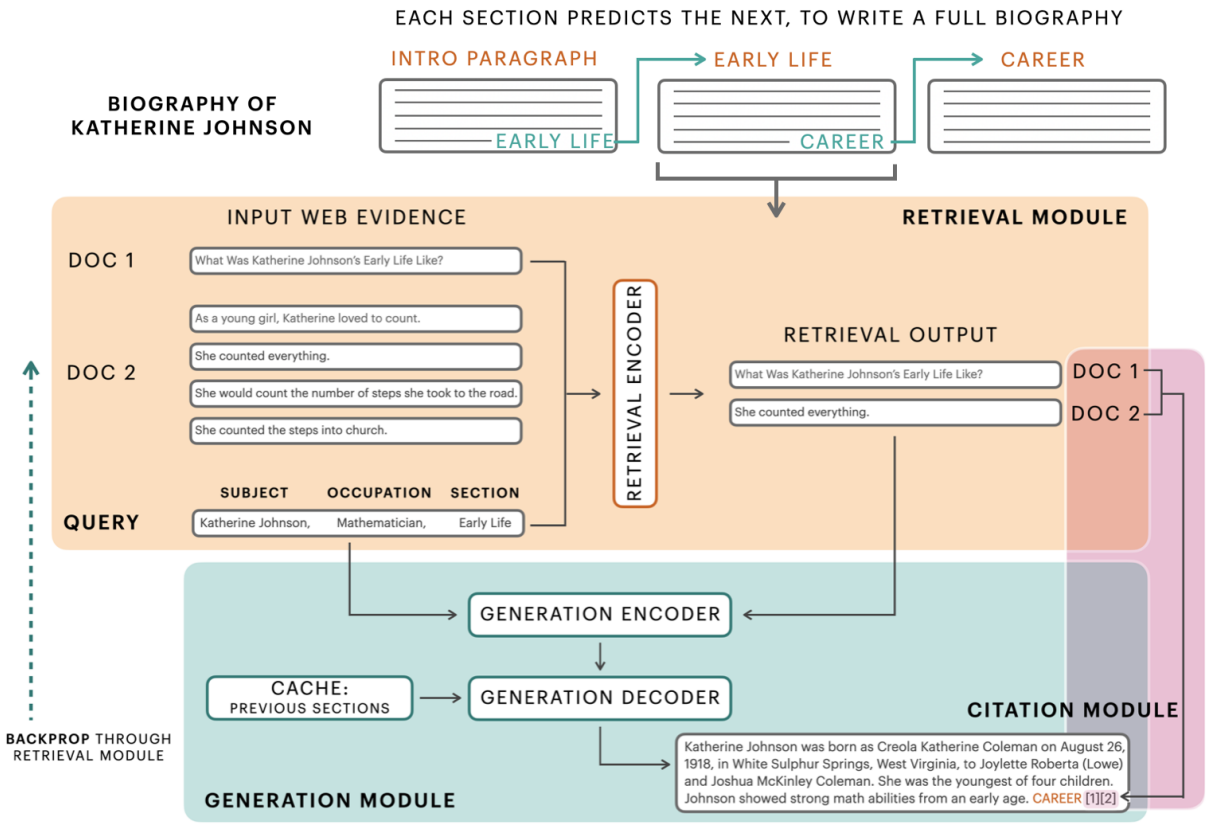

Figure 1: **Model Architecture.** Our method writes a Wikipedia article section by section, with each section predicting the next in sequence. To write one section, the model starts with a *retrieval module* that uses a query consisting of the subject name, occupation, and section heading to identify the most relevant information from the web. The query and retrieval output passes to the *generation module*, which generates the desired section while using a cache to reference previously written sections. Finally, to complete the full Wikipedia article, the *citation module* appends citations based on the retrieved content. The entire system is learned end-to-end, with backpropagation from the generation module through the retrieval module.

occupation(s), the section header and the retrieved evidence. Note that the maximum number of input tokens for BART is 1024 words, which is why we cap the retrieval at 1000 words, as described in the previous section. The decoder conditions on the input information to generate the section.

One challenge with this is that the sections would be generated completely independently, which might result in redundancy between generated sections. Thus, we equip the Sequence-to-Sequence model with a mechanism to refer to previous sections using the cache mechanism from Transformer-XL (Dai et al., 2019). This mechanism caches the previous section's hidden states at every layer, using it as memory to generate the current section.

4.3 Citation Module

Recent work has focused on models that not only perform a task, but also produce an explanation (DeYoung et al., 2019). Much of this work has focused on question answering (Latcinnik and Berant, 2020; Lamm et al., 2020; Lakhotia et al., 2020; Gonzalez et al., 2020) and generating explanations in natural language (Camburu et al., 2019; Narang et al., 2020; Kumar and Talukdar, 2020; Hase et al., 2020). A similar requirement exists on Wikipedia — not only to collate the information into an article, but to provide the *original references* for users to verify. Thus, to complete the generation of a full Wikipedia biography, we cite the information used, as in any real article. On Wikipedia itself, each sentence could contain citations. We simplify this, citing at the end of each section. To do this, we track the original document the retrieved evidence originates from, and reference that document at the end of the generated section.

4.4 Bringing it All Together

To write a full biography, models must generate the introductory paragraph followed by each section.

For a new article, the introductory paragraph is given as a section heading called *toplevel*. For each subsequent section, we follow the process outlined above to retrieve evidence, then write a section, then add citations. At the end of each section, the model generates the *section heading* of the next section. This allows the model to generate an entire article section by section.

5 Creating an Evaluation Dataset

A possible failure point for our method is the retrieval step as good biography generation requires access to sufficient relevant information. To study the impact of accurate retrieval on generation quality, we design a specific evaluation dataset that pushes this problem to the forefront. Specifically, we create a novel evaluation dataset which consists exclusively of biographies about women.

Ongoing efforts to write biographies about women in the Wikipedia editor community, such as the Women in Red project, have identified *insufficient online evidence* as a major challenge for writing Wikipedia biographies about women. To study the importance of retrieval on model quality, we therefore create an evaluation dataset where the target Wikipedia articles are women bios. We collate candidate biographies, retrieve information about their occupation, and gather web sources using web search. The resulting dataset, summarized in Table 2, consists of 1,527 biographies, each linked to a set of retrieved web articles.

Identifying Biographical Subjects. We first source various notable women on Wikipedia using internet lists (e.g. *Famous Women you should know*) and existing efforts by collective groups of Wikipedia editors, such as the Women in Red project. Several recent efforts focus on Women in Science [4], and so we specifically include scientists as a category. Overall, we collate almost two thousand candidate Wikipedia women biographies. We then narrow down by selecting articles that have previously *Featured Article* or *Good* quality. The final evaluation dataset contains 1,527 biographies in four groups: Women, Women in Science, Women in Asia, and Women in Africa (see Table 2).

Biography Text and Occupation. After finalizing candidate Wikipedia biographies, we use the MediaWiki API[5] to query the text of the article. We use the Wikidata API[6] to retrieved the individuals, possibly multiple, occupations (e.g. *Rachel Carson* is an author and an environmental activist). As seen in Table 2, on average, articles have around 6 sections with 130 words each. The most common occupations include writers, teachers, and doctors (see Table 1), though the entire dataset contains almost 500 different occupations, with people having on average 2 occupations (see Table 2).

Retrieving Web Evidence. Next, we identify web sources with reference evidence for each biography. We follow the construction of similar datasets, such as WikiSum (Liu et al., 2018) and ELI5 (Fan et al., 2019c), which searches through CommonCrawl. We query CommonCrawl based on the subject's name and occupation(s) and return the top 20 search results. We reject all CommonCrawl links from Wikipedia, to prevent querying the Wikipedia articles in our dataset. Statistics are presented in Table 2. Out of a maximum of 20 possible hits, on average each biography returns around 18.

6 Experimental Details

We describe our training data, baselines, and automatic and human evaluation metrics.

Training Data. We utilize the WikiSum (Liu et al., 2018) dataset of Wikipedia articles paired with web references. We filter to biographies using a combination of querying for occupations in Wikidata and using Named Entity Recognition[7] to recognize names. We query each article title in the WikiSum dataset to attempt to find an occupation and see the title is recognized as a named entity, to identify the bibliographical subset of WikiSum. This produces 677,085 biographies, each associated with a set of web articles.

Evaluation Data. We utilize the WikiSum (Liu et al., 2018) dataset, filtered to biographies, for evaluation. Similar to the training dataset, we query to identify occupational information. To study the impact of retrieval and available evidence on model

[4] https://towardsdatascience.com/who-is-wikipedia-famous-within-natural-language-processing-fa0c8e91bdf6?gi=b910dd838c47,https://www.newscientist.com/article/mg24532680-800-jess-wades-one-woman-mission-to-diversify-wikipedias-science-stories/

[5] https://www.mediawiki.org/wiki/API

[6] https://query.wikidata.org/

[7] https://spacy.io/usage/linguistic-features/

Most Common Section Headings	Career, Personal Life, Early Life, Biography, History
Most Common Occupations	Writer, Politician, University Teacher, Physician, Researcher

Table 1: **Example Section Headings and Occupations in Wikipedia Biographies.**

WikiSum Evaluation Dataset	
Average Number of Sections	7.2
Average Length of a Section	151.0
Average Length of Total Article	892.3
Avg overlap of Web Hits and Biography	39.8%
Our Evaluation Dataset	
Average Number of Sections	5.8
Average Length of a Section	132.3
Average Length of Total Article	765.9
Avg Number of Web Hits (max 20)	18.1
Avg overlap of Web Hits and Biography	24.9%
Biographies about Women	419
Biographies about Women in Science	808
Biographies about Women in Asia	164
Biographies about Women in Africa	136
Total Biographies	1,527

Table 2: **Breakdown and Statistics of Biographies** of a random sample of Wikipedia biographies compared to our created evaluation dataset.

quality, we also evaluate on our constructed evaluation dataset about women (which has substantially less web-based evidence). As shown in Table 2, these two datasets differ in the length and quality of both the Wikipedia articles and the web-based evidence.

Baseline. We compare our method described in Section 4 to a pretraining and finetuning generation baseline. We use the BART model (Lewis et al., 2019) and finetune on the Biography subset of the WikiSum data. Note that BART has a token limit of 1024, thus the entirety of the web retrieval is not available to this model. We take the web search hits ordered by the search engine, and provide the first 1000 available tokens. To compare this baseline with our method equitably, the baseline is also trained to generate section by section. However, it does not use the retrieval module (all evidence is given), the caching mechanism, or the citation module (as described in Section 4), meaning citations are not added to the generated text. Additional training details are in the Appendix.

Generation. We generate from all models with beam search, setting the beam size to 5. We allow the model to generate an output of any length, with no restrictions. For human evaluations, we set the minimum and maximum length such that it matches the length of the gold target to minimize the effect of length on human interpretations.

Automatic Evaluation. We evaluate the quality of generated biographies with three automatic metrics. First, we measure the **ROUGE-L** between the generated text and the Wikipedia reference text to assess the similarity. ROUGE-L is commonly used in multi-sentence summarization and is a measure of longest common substring overlap.

Next, we use **Natural Language Entailment** as a high level proxy for quantifying a form of factuality: if two sentences entail each other in both directions, then they are semantically equivalent. We use a model pretrained and finetuned on MNLI, open sourced by Liu et al. (2019). To evaluate entailment, we split the generated biography and reference biography into sentences, then for each sentence in the generated biography we calculate if it is semantically equivalent to a sentence in the reference. We then compute the percentage of generated sentences that are semantically equivalent to at least one sentence in the reference biography, where entailment is evaluated bidirectionally.

Finally, we assess the **Coverage** of information in the generated biography, constraining this to analyzing mentions of named entities. We report the percentage of named entities detected in the reference which are also detected in the generated text. We extract entities with BLINK, a BERT-based entity linking system (Wu et al., 2019).

Human Evaluation Long-form text generation is very difficult to assess automatically (Thomson and Reiter, 2020; Howcroft et al., 2020), particularly for factuality (Goodrich et al., 2019; Maynez et al., 2020; Peshterliev et al., 2021) and hallucination (Zhou et al., 2020; Dušek and Kasner, 2020). We conduct a detailed, large-scale human evaluation with the goal to assess **Coverage** (How much of the information in the reference section is in the generated section?) and **Factuality** (How much of the generated section is in the reference and, for the information added in the generated text, how much of that information is verifiable based on the web evidence?).

To reduce the challenge of evaluation, the text is compared section by section, and the generated text is the same length as the reference by constraining the max length of beam search (to remove length as an evaluation artifact). First, each sentence of the generated section is shown next to the full reference section and the entire document cited in the generated section (recall our generated biographies cite the retrieved evidence). Evaluators are asked to decide **(1)** if the information in the generated sentence is present in the reference section (ground truth) and **(2)** if the information in the generated sentence is present in the cited document (web evidence). This question assesses if the information from the generated section is factual with respect to either the reference Wikipedia text or the retrieved web documents. Then, the evaluation is flipped to assess coverage with respect to the Wikipedia reference. Each sentence of the reference is shown next to the generated section, and evaluators are asked to decide **(3)** if the information in the reference sentence is present in the generated section. In total, human annotators evaluated 100 sections with length between 200 to 500 words. Each section is reviewed by one annotator. Additional details are in the Appendix.

7 Results and Discussion

We describe our main results and analyze the importance of retrieval on model quality. An example generation is shown in Figure 2.

7.1 Quality of Generated Biographies

Automatic Evaluation. We examine the model's overall performance. Results are summarized in Table 3. Compared to the pretraining+finetuning baseline, adding the retrieval module statistically significantly[8] increases results by 1.4 ROUGE-L. Adding a caching mechanism improves further by 0.5 ROUGE-L. This trend is reflected across the entailment and entity coverage metrics, indicating that retrieving the most relevant information to write a biography is critical.

Next, we examine the impact of our modeling choices using ablation (Table 4). Compared to previous work on WikiSum (Liu et al., 2018; Fan et al., 2019a), we add an end-to-end retrieval mechanism based on RAG that substantially improves results. Further, instead of retrieving solely based on the subject name, as was previously done (Liu et al., 2018), we retrieve on a detailed query (the name, occupation, and section heading). Table 4 indicates that this enriched query improves the retrieval quality by almost 2 ROUGE-L. We conjecture it helps improve disambiguation and retrieve evidence that is relevant to the desired entity rather than to one of its homonyms.

We also generate the biographical articles section by section, rather than an entire article at once. This allows the retrieval mechanism to be focused on the section information. As shown in Table 4, this also has a positive effect of +1.5 ROUGE-L.

Human Evaluation. Next, we examine quality with human evaluation, as shown in Figure 3. Models generating nonfactual or hallucinated content is an ongoing area of study (Tian et al., 2019; Nie et al., 2019; Liu et al., 2021). Our goal is to understand how much information in the generated text is present in the reference text or the web evidence, as a proxy for factuality and coverage. Overall, 68% of the information in generated sections is not present in the reference text. Conversely, 71% of information in the reference text is not in the generated text. This indicates that the generated text has far from perfect coverage. However, we found that 17% of the added information can be validated by examining the web evidence, which shows that some information added by the generative model is valid biographical information.

We examine why there is low information overlap between the generated and reference text. First, information in the reference biography may not be available on the web[9] or may not be retrieved. In a manually examined subset of 250 sentences taken from reference biographies, we found that about 50% of the information was not contained in the web evidence. The other 50% was partially present in the web evidence but were not retrieved by the model. Second, annotators must compare sentences, but sentences contain partial information. For example, if *Person is was born in Chicago in 1968* was in the generated text and *Person was born in Chicago* was in the reference text, this would count as the generation having information not in the reference. Annotators were very precise in sticking to the requested standard that the *entire sentence* should be factual to count as fully factual, which is reflected by annotators marking *partial*

[8] We use the confidence interval reported in the ROUGE package.

[9] Note that search hits from the Wikipedia domain are removed from web search results.

Model	ROUGE-L	Entailment	Named Entity Coverage
BART Pretraining + Finetuning	17.4	15.8	21.9
+ Retrieval Module	18.8	17.2	23.1
+ Caching Mechanism	19.3	17.9	23.4

Table 3: **Full Results on Biography Generation.** We compare the BART baseline with our method across different automatic metrics to assess fluency, factuality, and coverage. Results are shown on the test set.

hyman is best known for her work on the classification of invertebrates. she was the author of a six-volume set of reference books titled the invertebrate treatise, which was published by mcgraw-hill in the united states and in germany. she also wrote a series of laboratory manuals for the teaching of zoology classes nationwide. hyman's work has had a lasting influence on scientific thinking about a number of animal groups, and the only works that can be compared with hers are of composite authorship.

Figure 2: **Example Generation** of the *Work* section for a biography about Libbie Hyman, a zoologist. Green indicates text in the reference article, Pink indicates text in the web evidence, and Orange (underlined) indicates hallucination. See the biography on Wikipedia: https://en.wikipedia.org/wiki/Libbie_Hyman.

Model	ROUGE-L
Retrieval with Different Queries	
with Subject Name Only	19.6
with Name and Occupation	19.8
with Name, Occupation, Section Heading	21.4
Writing Articles in Sections	
Entire Article	14.4
Section by Section	15.9

Table 4: **Ablations** of types of Queries for the Retrieval Module and generation section by section. Results are shown on the dev set.

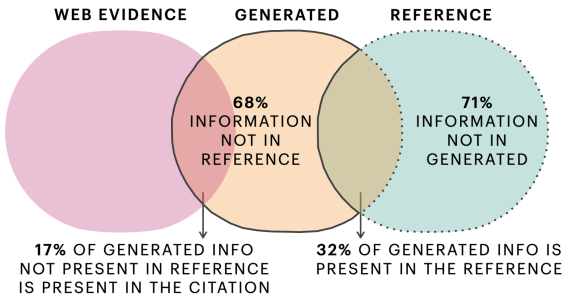

Figure 3: **Human Evaluation.** We compare the coverage of content between generated and reference biographies, as well as the factuality of generated content.

factuality as not factual. Our stringent standard for factuality produces a clearer understanding of hallucinations at the sentence-level.

In summary, our investigation suggests two explanations for the low coverage reported by human annotators: lack of information in the web evidence and difficulty assessing whether two sentences contain the same core knowledge.

7.2 Performance with Unreliable Retrieval

One major challenge of accurate Wikipedia article generation is when information is not available on the web or not easily retrieved. For example, information could simply not exist on the internet. Writing a Wikipedia biography about any randomly chosen person on the street would likely manifest this scenario. Other situations could include having a large number of search results returned but difficulty identifying which are relevant, having too few search results to write a good biographic article, or even having only noise returned in the search results. We discuss these challenges and possible mitigations in this section.

The Evidence Gap. We compare the results on our evaluation set about women with those on the WikiSum test set. Compared to WikiSum, the unigram overlap of the web hits with the biographical article is substantially lower for our evaluation dataset (see Table 2). As shown in Table 5, across the board, the quality of generated biographies is higher for the WikiSum Test set. This is especially prominent for Women in Asia and Africa, which are more than 2.5 ROUGE-L worse on average.

Reducing the Dependency on Retrieval. One challenge is that there is a disconnect between the training dataset, where retrieval information is readily available, and the women-focused evaluation dataset, where retrieval information is noisy or missing. We investigate the potential of a straightforward strategy to mitigate differences in training data: that of training on biographical articles with less reliable web evidence. We mimic this by finetuning our model on a subset of our evalu-

Model	WikiSum Test	Women	Scientists	Women in Asia	Women in Africa
BART Pretraining	19.0	17.4	18.2	16.7	16.4
+ Retrieval	21.4	18.8	19.3	17.9	17.1
+ Caching	21.8	19.3	19.7	18.4	17.3

Table 5: **ROUGE-L Performance broken down by sub-categories.** We compare the BART baseline with our method across different subsets of women, as well as the biography subset of WikiSum Test.

Model	WikiSum Test	Women Asia	Women Africa
Our Method	19.0	16.7	16.4
+ finetune on Women	18.9	17.3	16.8

Table 6: **Improved Performance when Finetuning** on biographical articles with less web evidence. We finetune on biographies about women that do not include this subset of women in Asia and Africa.

ation dataset, and then testing on Women in Asia and Africa, the two categories that perform most poorly. As shown in Table 6, finetuning statistically significantly improves performance, though the improvement is not large (+0.5 ROUGE-L). Another phenomenon that arises with noisy web evidence is that retrieving more is not necessarily better. Perhaps only one website has really relevant information. In the retrieval module, all available web documents are encoded at the sentence level, and the model can select sentences across all documents. We next explore an approach where the model first scores documents, then selects sentences from the most relevant document. We found this had very similar performance, and thus conclude that the challenge of identifying relevant documents and then sentences is probably similar in difficulty to identifying relevant sentences directly.

8 Conclusion

We developed a novel retrieval and cache-augmented generative model to generate long-form biographies based on evidence from the web. Experimental evidence reveals that an enriched query including occupations, caching, and backpropagation through the retrieval module contributes to improved performance. We investigate the dependency on high-quality web evidence, which manifests strongly in our constructed evaluation dataset of biographies about women. We discuss this challenge and possible mitigations.

9 Acknowledgments

We thank the anonymous reviewers for their feedback. We gratefully acknowledge the support of the French National Research Agency and of Facebook AI Research Paris (for Claire Gardent; award ANR-20-CHIA-0003, XNLG "Multilingual, Multi-Source Text Generation").

We thank Adina Williams, Emily Dinan, Ledell Wu, and Aleksandra Piktus for thoughtful discussions and feedback on this entire effort, as well as previous collaborations that influenced this work. We thank Sebastian Riedel, Douwe Kiela, Mona Diab, and Michael White for their suggestions to improve this work. We thank Mojtaba Komeili for developing the web query service we used to create the evaluation dataset.

Finally, we thank all of the editors of Wikipiedia, particularly those in the Women in Red Project, for their hard work and dedication to creating, moderating, editing, and all that is necessary to keep Wikipedia running. We encourage readers to donate to Wikipedia to support this public project.

10 Ethical Considerations

In this section, we discuss several known limitations and ethical considerations of our work. We do not recommend any kind of text generation technology to be deployed on Wikipedia given this is an active area of research.

10.1 Dependency on Evidence from the Web reflects Bias on the Internet

Biographies, whether written as books or available online, reflect societal bias. While many Wikipedia editors rely on web-based references to create their articles, and we follow the same strategy in this work, relying on the web is flawed. The prominent reason is that the internet is full of bias in it of itself. For example, Donna Strickland, who received a Nobel Prize, did not have a Wikipedia article[10]

[10] https://wikimediafoundation.org/news/2018/10/04/donna-strickland-wikipedia/#:~:text=Donna%20Strickland%20is%20an%

as there was not sufficient content about her on the web as a basis for her article. Thus, it is important to recognize that the availability of references is problematic, affecting the downstream ability to write accurate, comprehensive biographies. Further, information on the web can be contradictory, information can be affected by the passage of time, and not information on the web is necessarily factually correct. Our proposed modeling mechanism does not have a way to explicitly recognize or correct for these challenges, which also plagues text generation generally.

10.2 Focus on English Limits Inclusivity from Other Languages

Our work focuses on text generation in English only, which limits inclusivity purely on the basis of language. This is challenging as the content of the internet and Wikipedia itself is different in various languages. For example, articles about people from Germany may be more likely to be located on the German version of Wikipedia. Another factor is that the content of the references may be written in another language, and then used by a bilingual individual to write an article in English about that subject. This is often the case for many biographical subjects who may be more well known in a non-English speaking area.

10.3 Evaluation focuses on Women Only, Not Other Groups

There are a very large number of marginalized groups in the world and numerous important intersectional aspects to consider. When discussing identity, a wide variety of factors and personal views influence individuals when thinking about how they describe themselves. Our evaluation dataset focuses on women alone, which leaves out many groups, including non-binary people. Further, Wikipedia may not reflect the up-to-date information — names and gender are both mutable, for example — and Wikipedia articles do not ask each subject to self-report their gender. Finally, we note that by grouping people into hard categories, there can potentially be harm — such as limiting people from opportunities because of their gender or race. However, we strongly believe that it is important to recognize bias in its various forms as it exists, particularly in popular, default online sources of information such as Wikipedia.

20optical,of%20a%20Sloan%20Research%20Fellowship.

10.4 Bias in Style, Word Choice, and Tone

In this work, we focus on bias manifesting as unequal prevalence and length of biographical content on Wikipedia, focusing specifically on different intersectional groups of women. However, bias manifests in a number of other ways. Studies have indicated that the words used in biographies about women compared to biographies about men (Dinan et al., 2019) also differs, and is reflective of gendered terminology. For example, many articles about women are actually written with a lot of information about men, such as their husband's careers, and articles about actresses describe more often their physical appearance. This is also a manifestation of bias, and we do not present any focused modeling techniques to address this type of bias explicitly.

10.5 Biographies as Records

In the modern internet, a large number of events are recorded for the public record. These include events that people may personally prefer to forget, often termed *right to be forgotten*[11]. Automatically generating biographies about individuals may collate such information in an easily accessible public place, which can conflict with this personal right. This has a complex but important interaction with marginalized groups. For example, many celebrities who are women, transgender, or a part of another marginalized group are far more likely to have news articles written about intimate personal details such as plastic surgeries. Thus, it is important to consider the interaction of biographical data with individual privacy. This is a larger challenge of biographical information generally.

References

Oshin Agarwal, Heming Ge, Siamak Shakeri, and Rami Al-Rfou. 2020. Large scale knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training. *arXiv preprint arXiv:2010.12688*.

Siddhartha Banerjee and Prasenjit Mitra. 2015. Wikikreator: Improving wikipedia stubs automatically. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 867–877.

[11] https://en.wikipedia.org/wiki/Right_to_be_forgotten

Pablo Beytía. 2020. The positioning matters: Estimating geographical bias in the multilingual record of biographies on wikipedia. In *Companion Proceedings of the Web Conference 2020*, pages 806–810.

Fadi Biadsy, Julia Hirschberg, and Elena Filatova. 2008. An unsupervised approach to biography production using wikipedia. In *Proceedings of ACL-08: HLT*, pages 807–815.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

Oana-Maria Camburu, Brendan Shillingford, Pasquale Minervini, Thomas Lukasiewicz, and Phil Blunsom. 2019. Make up your mind! adversarial generation of inconsistent natural language explanations. *arXiv preprint arXiv:1910.03065*.

Thiago Castro Ferreira, Claire Gardent, Nikolai Ilinykh, Chris van der Lee, Simon Mille, Diego Moussallem, and Anastasia Shimorina. 2020. The 2020 bilingual, bi-directional WebNLG+ shared task: Overview and evaluation results (WebNLG+ 2020). In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 55–76, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*.

Mingda Chen, Sam Wiseman, and Kevin Gimpel. 2020a. Generating wikipedia article sections from diverse data sources. *arXiv preprint arXiv:2012.14919*.

Wenhu Chen, Yu Su, Xifeng Yan, and William Yang Wang. 2020b. Kgpt: Knowledge-grounded pre-training for data-to-text generation. *arXiv preprint arXiv:2010.02307*.

Andrew Chisholm, Will Radford, and Ben Hachey. 2017. Learning to generate one-sentence biographies from wikidata. *arXiv preprint arXiv:1702.06235*.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2019. Eraser: A benchmark to evaluate rationalized nlp models. *arXiv preprint arXiv:1911.03429*.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2019. Queens are powerful too: Mitigating gender bias in dialogue generation. *arXiv preprint arXiv:1911.03842*.

Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. 2020. Multi-dimensional gender bias classification. *arXiv preprint arXiv:2005.00614*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

Ondřej Dušek and Zdeněk Kasner. 2020. Evaluating semantic accuracy of data-to-text generation with natural language inference. *arXiv preprint arXiv:2011.10819*.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Angela Fan, Claire Gardent, Chloé Braud, and Antoine Bordes. 2019a. Using local knowledge graph construction to scale seq2seq models to multi-document inputs. In *2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing*.

Angela Fan, Edouard Grave, and Armand Joulin. 2019b. Reducing transformer depth on demand with structured dropout. *arXiv preprint arXiv:1909.11556*.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019c. Eli5: Long form question answering. *arXiv preprint arXiv:1907.09190*.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The WebNLG challenge: Generating text from RDF data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133, Santiago de Compostela, Spain. Association for Computational Linguistics.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862*.

Ana Valeria Gonzalez, Gagan Bansal, Angela Fan, Robin Jia, Yashar Mehdad, and Srinivasan Iyer.

2020. Human evaluation of spoken vs. visual explanations for open-domain qa. *arXiv preprint arXiv:2012.15075*.

Ben Goodrich, Vinay Rao, Peter J Liu, and Mohammad Saleh. 2019. Assessing the factual accuracy of generated text. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 166–175.

Eduardo Graells-Garrido, Mounia Lalmas, and Filippo Menczer. 2015. First women, second sex: Gender bias in wikipedia. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, pages 165–174.

Kelvin Guu, Tatsunori B Hashimoto, Yonatan Oren, and Percy Liang. 2018. Generating sentences by editing prototypes. *Transactions of the Association for Computational Linguistics*, 6:437–450.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*.

Peter Hase, Shiyue Zhang, Harry Xie, and Mohit Bansal. 2020. Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language? *arXiv preprint arXiv:2010.04119*.

Marit Hinnosaar. 2019. Gender inequality in new media: Evidence from wikipedia. *Journal of Economic Behavior & Organization*, 163:262–276.

David M Howcroft, Anja Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: Nlg needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182.

Encina Calvo Iglesias. 2020. Preparing biographies of stem women in the wikipedia format, a teaching experience. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, 15(3):211–214.

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.

Zhijing Jin, Qipeng Guo, Xipeng Qiu, and Zheng Zhang. 2020. Genwiki: A dataset of 1.3 million content-sharing text and graphs for unsupervised graph-to-text generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2398–2409.

Lucie-Aimée Kaffee, Hady Elsahar, Pavlos Vougiouklis, Christophe Gravier, Frédérique Laforest, Jonathon Hare, and Elena Simperl. 2018a. Learning to generate wikipedia summaries for underserved languages from wikidata. *arXiv preprint arXiv:1803.07116*.

Lucie-Aimée Kaffee, Hady Elsahar, Pavlos Vougiouklis, Christophe Gravier, Frédérique Laforest, Jonathon Hare, and Elena Simperl. 2018b. Mind the (language) gap: Generation of multilingual wikipedia summaries from wikidata for articleplaceholders. In *European Semantic Web Conference*, pages 319–334. Springer.

Lucie-Aimée Kaffee, Pavlos Vougiouklis, and Elena Simperl. 2020. Using natural language generation to bootstrap missing wikipedia articles: A human-centric perspective. *Semantic Web Journal*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Sawan Kumar and Partha Talukdar. 2020. Nile: Natural language inference with faithful natural language explanations. *arXiv preprint arXiv:2005.12116*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Kushal Lakhotia, Bhargavi Paranjape, Asish Ghoshal, Wen-tau Yih, Yashar Mehdad, and Srinivasan Iyer. 2020. Fid-ex: Improving sequence-to-sequence models for extractive rationale generation. *arXiv preprint arXiv:2012.15482*.

Matthew Lamm, Jennimaria Palomaki, Chris Alberti, Daniel Andor, Eunsol Choi, Livio Baldini Soares, and Michael Collins. 2020. Qed: A framework and dataset for explanations in question answering. *arXiv preprint arXiv:2009.06354*.

Isabelle Langrock and Sandra González-Bailón. 2020. The gender divide in wikipedia: A computational approach to assessing the impact of two feminist interventions. *Available at SSRN*.

Veronica Latcinnik and Jonathan Berant. 2020. Explaining question answering models through text generation. *arXiv preprint arXiv:2004.05569*.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. *arXiv preprint arXiv:1603.07771*.

Nayeon Lee, Andrea Madotto, and Pascale Fung. 2019. Exploring social bias in chatbots using stereotype knowledge. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 177–180.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy,

Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2005.11401*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Haochen Liu, Wentao Wang, Yiqi Wang, Hui Liu, Zitao Liu, and Jiliang Tang. 2020. Mitigating gender bias for neural dialogue generation with adversarial learning. *arXiv preprint arXiv:2009.13028*.

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

Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. 2021. A token-level reference-free hallucination detection benchmark for free-form text generation. *arXiv preprint arXiv:2104.08704*.

Xiaojiang Liu, Zaiqing Nie, Nenghai Yu, and Ji-Rong Wen. 2010. Biosnowball: automated population of wikis. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 969–978.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Wei Luo, Julia Adams, and Hannah Brueckner. 2018. The ladies vanish?: American sociology and the genealogy of its missing women on wikipedia. *Comparative Sociology*, 17(5):519–556.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.

Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M Khapra. 2018. Towards exploiting background knowledge for building conversation systems. *arXiv preprint arXiv:1809.08205*.

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. Wt5?! training text-to-text models to explain their predictions. *arXiv preprint arXiv:2004.14546*.

Feng Nie, Jin-Ge Yao, Jinpeng Wang, Rong Pan, and Chin-Yew Lin. 2019. A simple recipe towards reducing hallucination in neural surface realisation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2673–2679.

Ankur P Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. Totto: A controlled table-to-text generation dataset. *arXiv preprint arXiv:2004.14373*.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231*.

Stan Peshterliev, Barlas Oguz, Debojeet Chatterjee, Hakan Inan, and Vikas Bhardwaj. 2021. Conversational answer generation and factuality for reading comprehension question-answering. *arXiv preprint arXiv:2103.06500*.

Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Dmytro Okhonko, Samuel Broscheit, Gautier Izacard, Patrick Lewis, Barlas Oğuz, Edouard Grave, Wen-tau Yih, et al. 2021. The web is your oyster–knowledge-intensive nlp against a very large web corpus. *arXiv preprint arXiv:2112.09924*.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with content selection and planning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6908–6915.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Christina Joan Sauper and Regina Barzilay. 2009. Automatically generating wikipedia articles: A structure-aware approach. Association for Computational Linguistics.

Katja Geertruida Schmahl, Tom Julian Viering, Stavros Makrodimitris, Arman Naseri Jahfari, David Tax, and Marco Loog. 2020. Is wikipedia succeeding in reducing gender bias? assessing changes in gender bias in wikipedia using word embeddings. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 94–103.

Minjoon Seo, Jinhyuk Lee, Tom Kwiatkowski, Ankur P Parikh, Ali Farhadi, and Hannaneh Hajishirzi. 2019. Real-time open-domain question answering with dense-sparse phrase index. *arXiv preprint arXiv:1906.05807*.

Lei Sha, Lili Mou, Tianyu Liu, Pascal Poupart, Sujian Li, Baobao Chang, and Zhifang Sui. 2018. Order-planning neural text generation from structured data. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*.

Gabriel Stanovsky, Noah A Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. *arXiv preprint arXiv:1906.00591*.

Despina Stratigakos. 2016. Unforgetting women architects: From the pritzker to wikipedia. *Places Journal*.

Craig Thomson and Ehud Reiter. 2020. A gold standard methodology for evaluating accuracy in data-to-text systems. *arXiv preprint arXiv:2011.03992*.

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

Ran Tian, Shashi Narayan, Thibault Sellam, and Ankur P Parikh. 2019. Sticking to the facts: Confident decoding for faithful data-to-text generation. *arXiv preprint arXiv:1910.08684*.

Pavlos Vougiouklis, Hady Elsahar, Lucie-Aimée Kaffee, Christophe Gravier, Frédérique Laforest, Jonathon Hare, and Elena Simperl. 2018. Neural wikipedian: Generating textual summaries from knowledge base triples. *Journal of Web Semantics*, 52:1–15.

Zhenyi Wang, Xiaoyang Wang, Bang An, Dong Yu, and Changyou Chen. 2020. Towards faithful neural table-to-text generation with content-matching constraints. *arXiv preprint arXiv:2005.00969*.

Zena Worku, Taryn Bipat, David W McDonald, and Mark Zachry. 2020. Exploring systematic bias through article deletions on wikipedia from a behavioral perspective. In *Proceedings of the 16th International Symposium on Open Collaboration*, pages 1–22.

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

Catherine Yeo and Alyssa Chen. 2020. Defining and evaluating fair natural language generation. *arXiv preprint arXiv:2008.01548*.

Amber G Young, Ariel D Wigdor, and Gerald C Kane. 2020. The gender bias tug-of-war in a co-creation community: Core-periphery tension on wikipedia. *Journal of Management Information Systems*, 37(4):1047–1072.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Paco Guzman, Luke Zettlemoyer, and Marjan Ghazvininejad. 2020. Detecting hallucinated content in conditional neural sequence generation. *arXiv preprint arXiv:2011.02593*.

A Appendix

A.1 Model and Training Details

We use the BART-Large model as open sourced by Lewis et al. (2019). We train with learning rate $3e-05$ and a polynomial decay learning rate schedule, warming up for 500 updates, and end training after 50,000 updates. We train with dropout and attention dropout 0.1, label smoothing 0.1, and 0.01 weight decay. Our final model trains on 8 GPUs for three days. For experimentation, we train on 4 GPUs for 12 hours, which is about the time required for convergence.

A.2 Human Evaluation Details

Our evaluation is conducted on the Amazon Mechanical Turk platform. We pay evaluators approximately fifteen dollars an hour. Each section is evaluated independently, and evaluation tasks are not batched. The generated section and reference section are displayed side by side, segmented into separate sentences. To ease the challenge of human evaluation, we evaluate sentence by sentence. This is displayed by highlighting sentences independently, to reduce information overload.

A.3 Additional Examples

We present several examples of full generated articles in Figure 4.

A.4 Amount of Information Used from Retrieved Documents

Sequence-to-sequence models for text generation are able to utilize retrieval to augment generation, widely used in tasks such as question answering. Compared to these tasks, where the information to e.g. compose a written answer to a question is contained in a very specific paragraph, writing Wikipedia articles is much more freeform. For example, Wikipedia articles usually are written by human editors who have looked at a large amount of source material and paraphrased it, and articles are edited by many people over time. Thus, we find that it is difficult to directly retrieve a perfect provenance document that part of the Wikipedia article could be copy-pasted from.

We analyze how the model utilizes the retrieved information, and we find three main cases. In the first case, a small number of the web search documents are very useful (for example, biographical information about the person already on the web, such as on `biography.com`). In this case, the model utilizes this information very heavily, and often only retrieves content from this small number of documents. In the second case, there are a number of partially relevant documents, and web searches on the different predicted section headings change the web search results. Thus, models retrieve small amounts of information from multiple different sources. Finally, the third case is discussed in Section 7.2, and is potentially the most challenging to resolve: the situation where little information about the biographical subject is present on the web.

These three scenarios arise for all biographical articles, but differ in prevalence between different categories of people. For example, certain occupations more naturally come with some quantity of information available online compared to others. An example is Olympic athletes — at that level of notability, usually their athletic career is chronicled more by the media, thus making a larger quantity of evidence on the web available. Another example can extend to scientists, where we observed that scientists in the United States tend to have personal websites that collate a lot of information, compared to scientists in other locations.

toplevel phoebe legere is an american singer - songwriter, painter, actress, and musician. she is a member of the acadian - cajun renaissance, a group of artists who combine elements of americana, cajuns, blues, and jazz. she has released 17 cds of original and traditional music, and her latest album is heart of love, and on the charts on americana radio. she appears on hbo 's documentary it' s me hilary. her original song "hip hop frog ", a song about the environment, was licensed by hbo. leger 's paintings and drawings are in the collections of the metropolitan museum of art, the museum of fine arts, boston, and the metropolitan opera. [1,3,4,8,11]

---

toplevel joan paton (née paton) (c. 1883 – 1962) was a british ornithologist. she was the first woman to be elected a fellow of the linnean society of london and the royal society of edinburgh. she is best known for her work in the field of ornithology, particularly her work on the birds of wales and scandinavia, and for her contributions to the study of birds of the north of england and scotland. [2]
=early life= paton was born in london, england. she was educated at the university of cambridge, where she graduated with a bachelor of arts (ba) degree in zoology and a master of science (msc) degree. she went on to earn a doctor of philosophy (phd) degree from the london school of hygiene and tropical medicine (lse) and a postgraduate diploma in tropical medicine and hygiene from the royal college of physicians and surgeons of london (rcpsl). [2,5]
=career= paton began her career as an ornithologist at the royal botanic gardens, kew, where she was a member of the ornithological society of london. she was elected a fellow of the british ornithologists' union (f. a. e. u.) in 1954. she served as the society 's vice - president from 1958 to 1960. she became a fellow in 1962 and was elected to the royal society of edinburgh in 1964. she also served on the council of the society for the protection of birds of great britain and ireland. paton was elected an honorary fellow of st john 's college, cambridge in 1966. she retired from the society in 1972. she died in london in 1984. [1,2]

---

toplevel ashley mckenzie is a canadian film director, screenwriter and producer. she is the winner of the stella artois jay scott prize for emerging talent at the 2016 toronto international film festival. her first feature film, werewolf, premiered at the toronto film festival in 2016. she has also directed short films for the national film board of canada and the canadian screen actors guild. she was born in montreal, quebec, canada, and grew up in ottawa, ontario. [1,3,11,13,14]
=personal life= mckenzie was born in london, england. she is the daughter of alexander mckenzie, who was a member of the british rock band the beatles. she has a younger sister, jessica, who is also a singer. she was educated at st mary 's college, oxford, where she graduated with a bachelor of arts degree in english literature. she also studied at the university of london. she married fellow x factor contestant andrew davies in september 2006. they have two children, a son and a daughter. [3,4,7,8,10,11]
=career= mckenzie was a contestant on the third series of the x - factor in 2006. she was eliminated in the first week of the competition. in 2007, mckenzie released her debut single "don 't pretend you hadn' t, now..." which peaked at no .160; 2 on the uk singles chart. she also released a second single ," i 'm not afraid ", in 2008. in 2009, she released her third single ," don' t pretend you haven 't, now ". in 2010, she was a judge on the x factor uk. [2]

Figure 4: **Random Examples of Generated Articles.** Note that *toplevel* is an augmented special tag to indicate the start of the article and = surrounds section headings on Wikipedia. Text in brackets indicates the cited references.