WEBBRAIN: LEARNING TO GENERATE FACTUALLY CORRECT ARTICLES FOR QUERIES BY GROUNDING ON LARGE WEB CORPUS

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we introduce a new NLP task – generating short factual articles for queries by mining supporting evidence from the Web. In this task, called WEBBRAIN, the ultimate goal is to generate a fluent, informative, and factually-correct short article (*e.g.*, Wiki article) for a factual query unseen in Wikipedia. To enable experiments on WEBBRAIN, we construct a large-scale dataset WebBrain-Raw by extracting English Wikipedia articles and their crawlable Wiki references. WebBrain-Raw is ten times larger than the previous biggest peer dataset, which can greatly benefit the research community. Besides, we empirically analyze the performances of the current state-of-the-art NLP techniques on WEBBRAIN and introduce a new framework ReGen, which enhances the generation factualness by improved evidence retrieval and task-specific pre-training for generation. Experiment results show that ReGen outperforms all baselines in both automatic and human evaluations.

1 INTRODUCTION

Information acquisition is one of the fundamental daily needs of human beings. Acquiring information from the Web undoubtedly is a convenient and efficient way. However, with the exponential growth of the Web, information on the Web becomes scattered and evolves quickly, making it challenging for users to acquire the expected information quickly. As a result, the Wikipedia articles become the best bet for most users when searching answers for factual queries on the Web (Singer et al., 2017). The reason is that the Wikipedia articles provide credible contents in which most claims can be supported by references from reputable sources. While Wikipedia is a good source of answers for factual queries, the need for manual editing (crowd-sourcing and editor checking) curbs its growth of coverage on a broader range of information needs. What if the Wikipedia articles can be automatically generated?

In this paper, we introduce a new task, WEBBRAIN, exploring the capacity of generating short factual articles for queries via a large web corpus. **Given a factual query, the goal of the task is to enable a system to mine supporting evidence from the Web and generate a short factual article in which the claims are supported by the mined evidence** (defined in Section 3.1). One of the potential generation targets for WEBBRAIN is the first section of a new Wiki page, based on which we can further explore generating long factual articles (*e.g.*, a complete Wiki page). WEBBRAIN can be greatly helpful in various scenarios, including generating Wiki pages for new entities, intelligent writing assistance, knowledge-intensive QA, etc. WEBBRAIN's goal is considered one of the ultimate goals of the future search engine (Metzler et al., 2021). Figure 1 illustrates a case of our WEBBRAIN.[1]

To establish the data foundation of WEBBRAIN, we construct a large-scale dataset, WebBrain-Raw, from scratch by extracting all English Wikipedia articles and all the corresponding reference articles. To the best of our knowledge, WEBBRAIN-Raw is the biggest dataset sourced from Wikipedia (about 10× larger than the previous biggest peer WikiSum (Liu et al., 2018), introduced in Section 3.2). Along with WEBBRAIN, we empirically investigate the ability of the current state-of-the-art techniques and conclude that most current models lack the ability to correctly cite references and

[1]the text generation result is obtained via the OpanAI's GPT3 API: `https://beta.openai.com/`

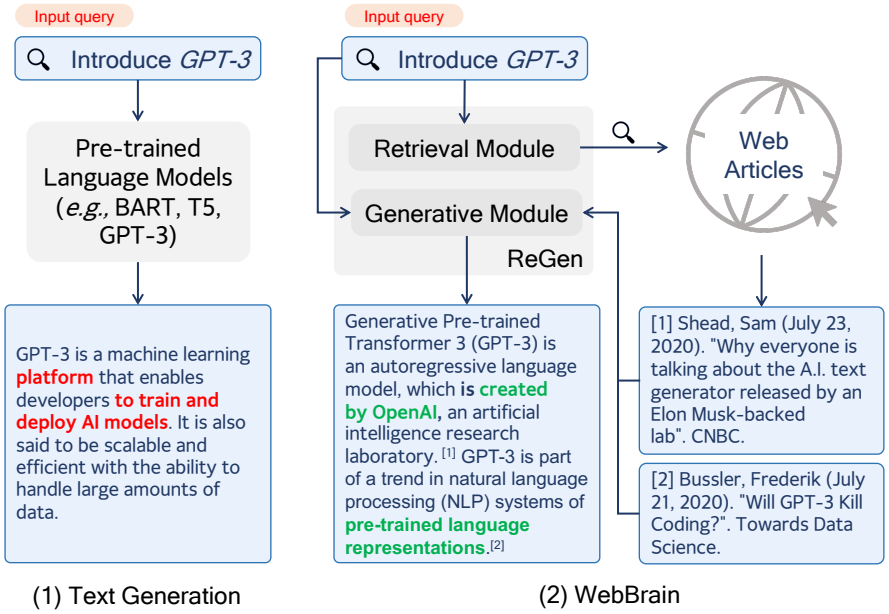

Figure 1: Comparison between text generation task and WEBBRAIN.

generate factual descriptions. Based on the best current models (SPLADE+FiD+BART) (Izacard & Grave, 2021; Formal et al., 2021a), we propose a new framework, ReGen, which enhances the factual correctness of generation by (1) controlling the topic consistency of the retrieved supporting evidence; (2) introduce the citation mark into the generation process to indicate the source of a referenced claim; (3) warm-up the generation model by pre-training with fact-oriented self-supervised tasks (described in Section 4).

Current tasks that are similar to WEBBRAIN include: (1) text generation with generative pre-trained language models (*e.g.*, GPT-3) (Brown et al., 2020); (2) retrieval-augmented QA (*e.g.*, RAG (Lewis et al., 2020b), GopherCite (Menick et al., 2022), REALM (Guu et al., 2020)); (3) multi-document summarization (MDS) (Liu et al., 2018; Liu & Lapata, 2019); (4) WebGPT (Nakano et al., 2021).

WEBBRAIN distinguishes itself from the existing tasks on the following aspects: (1) existing generative models generate text solely depending on the implicit knowledge stored in the model's parameters, which are prone to generating factually incorrect statements, a phenomenon commonly called *hallucination*. In contrast, our WEBBRAIN aims to generate factual statements depending on the supporting evidence mined from the Web; (2) Retrieval-augmented QA utilizes the retriever to enhance answering specific questions whose answer is usually a text span or a single sentence. Given a factual query, our WEBBRAIN explores capturing all the knowledge available on the Web to generate a comprehensive and accurate short article; (3) WEBBRAIN has a more complex pipeline than the MDS task, which focuses on summarizing documents that are already prepared. WEBBRAIN requires models to mine useful evidence before generating the factual article; (4) WebGPT mimics the behavior of a human by browsing the Web to answer a specific question. This involves many challenges that are difficult to solve (*e.g.*, collecting training data by recording human behaviors). Instead, WEBBRAIN grounds on a large number of Wikipedia articles that are already created and edited by crowds, which makes the task much more realistic.

In summary, our contributions are threefold: (1) we introduce a new task, WEBBRAIN, dedicated to answering factual queries by generating short factual articles based on evidence retrieved from a large web corpus; (2) we construct a large-scale dataset, WebBrain-Raw, to evaluate the potential of WEBBRAIN. WebBrain-Raw is ten times larger than the previous biggest peer dataset; and (3) we empirically analyze the performance of current state-of-the-art techniques on WEBBRAIN, and propose our factuality-enhanced framework, ReGen, which outperforms all baselines through both automatic and human evaluation.

2 RELATED WORK

Pre-trained Language Models Pre-trained language models (PLMs) have been widely applied to various natural language processing (NLP) tasks and achieved outstanding performance. In general, these models are first *pre-trained* on a large-scale corpus, and then *fine-tuned* on downstream datasets for specific tasks. By pre-training, the model can learn effective language representations, thus improving the performance on downstream tasks. Typical paradigms of PLMs include masked LM (Devlin et al., 2019; Zhang et al., 2019; Sun et al., 2019), left-to-right LM (Radford et al., 2019; Black et al., 2021), prefix LM (Dong et al., 2019; Bao et al., 2020), and encoder-decoder LM (Song et al., 2019; Raffel et al., 2020; Lewis et al., 2020a). Masked LMs are generally most suitable for natural language understanding or analysis tasks, *e.g.*, text classification and natural language inference. The other three types of PLMs can be naturally used for text generation.

Wikipedia-related Tasks Wikipedia is a popular knowledge source for NLP tasks that aim to utilize external knowledge. Examples include knowledge-grounded conversation (Dinan et al., 2019), fact checking (Thorne et al., 2018), open-domain QA Kwiatkowski et al. (2019), and slot filling El-Sahar et al. (2018) etc. Petroni et al. (2021) propose a benchmark for those knowledge-intensive language tasks (KILT) in which all tasks apply the snapshot of Wikipedia as the knowledge base. Compared to the KILT, other Wikipedia-related tasks more focus on the Wikipedia itself. Liu et al. (2018) construct a dataset WikiSum from Wikipedia to enable multi-document summarization (MDS). The MDS task is then further explored by Liu & Lapata (2019) with a hierarchical model and Perez-Beltrachini et al. (2019) with a topic-guided structure. WikiWrite (Banerjee & Mitra, 2016) propose a systematic solution to constructing Wiki pages by assigning retrieved content into different topical sections. Fruit (Iv et al., 2022) explores update information of existing Wiki pages. Recently, Piktus et al. (2021) construct a dataset Sphere from CCNet (Wenzek et al., 2020), using which they explore refining the citation quality of Wikipedia pages.

Retrieval-Augmented Text Generation Pre-trained language models are prone to generating hallucinations (*i.e.*, factually incorrect statements) (Vinyals & Le, 2015; Koehn & Knowles, 2017; Rohrbach et al., 2018; Raunak et al., 2021). Retrieval-augmented text generation, as a new text generation paradigm, can be a possible way to alleviate this problem. Compared with generation-based counterpart, this new paradigm can reduce the reliance of storing enormous knowledge in parameters by providing more references. Retrieval-augmented text generation has been widely applied to many NLP tasks, such as dialogue generation (Weston et al., 2018; Zhu et al., 2020), machine translation (Gu et al., 2018; Cai et al., 2021), and open-domain question answering (Lewis et al., 2020b; Guu et al., 2020). In this work, we formalize a new retrieval-augmented text generation task – WEBBRAIN with two new features: (1) The retrieval base is a large-scale open-domain web corpus; and (2) The target is to generate a natural, informative, and factual text rather than a short span. Both features make the task extremely challenging. Our experiments will show that though existing models can be applied to this task, there is still a large space for research and exploration.

3 TASK: WEBBRAIN

3.1 TASK DEFINITION

Formally, given an input query q, our task is to generate a short article T_q for q grounded by prior knowledge Θ. We have:

$$T_q^* = \arg\max p(T_q|q, \Theta). \tag{1}$$

For a pre-trained language model solely, the knowledge prior Θ is stored in the PLM's internal parameters. For WEBBRAIN, the knowledge prior Θ contains the implicit language knowledge θ stored in the PLM's parameters and the explicit knowledge γ defined by the supporting evidence V, which is mined from an external large web corpus C. In practice, C can be an off-line processed large web corpus, an online search engine API, or a combination of both. Intuitively, coupling θ and γ, $\Theta = (\theta, \gamma)$ can be characterized by the strong language capacity from the PLM and the factual constraints from the supporting evidence.

Supposing the generated T_q contains n sentences $(t_1, \cdots, t_n)$ and the retrieved supporting evidences $V = (v_1, \cdots, v_k) \in C$, we expect each sentence t would be supported by a V's subset V^τ, denoted

Table 1: Datasets built on Wikipedia. WEBBRAIN-Raw is significantly larger than existing datasets.

Dataset	# Wiki Pages	# Refs	Status	Storage Size
WikiSum (Liu et al., 2018)	2.3M	87M	Need crawling	300GB
WikiCatSum Perez-Beltrachini et al. (2019)	0.17M	23.5M	Ready	4.8GB
Hiersumm (Liu & Lapata, 2019)	1.66M	-	Ready	6.9GB
WEBBRAIN-Raw	14.86M	259.5M	Ready	2.8TB

as $t^{\leftarrow V^\tau}$. More precisely, we have:

$$T_q^* = (t_1^{\leftarrow V^{\tau_1}}, \cdots, t_n^{\leftarrow V^{\tau_n}}) = \arg\max p(T_q|q, \theta, \gamma), \quad \gamma := V \in C. \tag{2}$$

Note that when the generated sentence only relies on the PLM's language knowledge θ (*e.g.*, common sense generation), the corresponding supporting evidence is $V^\tau = \emptyset$.

3.2 DATA COLLECTION

Wikipedia is one of the most inclusive and widely-used encyclopedias on the web. By the nature of an encyclopedia, Wikipedia can be regarded as a large collection of factual articles of various topics (*e.g.*, entry of a Wiki page). The factual content in Wikipedia are generated by crowd-sourcing, in which most claims can be supported by the citations in the *References* section. Due to such nature, Wikipedia is the most appropriate resource to leverage for WEBBRAIN. The specific reasons are: (1) Wikipedia's format conforms to the WEBBRAIN task's definition in Eq. (2). We can view the Wiki page title as a factual query q, the citations (or parts of them) as the retrieved supporting evidence γ, and the Wiki article as the target factual article T_q^*; (2) Wikipedia has a huge volume of references which are usually reputable sources. Using the references, we can build a large web corpus of good quality as the retrieval source; (3) Wikipedia covers a wide range of topics of common interests, which provides a good knowledge basis for models to generalize to open-domain queries.

Many previous studies collected Wikipedia data for different tasks (Liu et al., 2018; Liu & Lapata, 2019; Perez-Beltrachini et al., 2019). However, none of these datasets provides full-size Wikipedia pages and the corresponding references. To lay a good data foundation to WEBBRAIN, we build a large-scale dataset from scratch by extracting all Wikipedia pages and all crawlable references. Table 1 shows the statistics of different datasets. In the Supplementary Material, we attach 500 raw data samples and the corresponding experiment data for demonstration. The full datasets we construct for WEBBRAIN will be released upon acceptance of this paper.

Table 2: Statistics of data for experiments.

	WebBrain-R	WebBrain-G
# Queries	2.74M	12.32M
# Ref. passages	3.20M	12.61M
# Tokens / Query	3.2	2.9
# Tokens / Passage	237.5	250.0
# Tokens / Target	-	108.6
# Training	4.46M	12.30M
# Validation	0.2M	0.5M
# Test	88,935	24,546

Data Cleaning As shown in Table 1, the WebBrain-Raw data contains full-size Wiki pages and a huge volume of references which are open-domain web pages. Regarding data quality, the Wiki pages are good due to their server stability and format consistency. However, the references contain much noise or are useless for our task. For example, they may be images, tables, or have empty content etc. To guarantee data quality, we process the WebBrain-RAW dataset in the following steps:

(1) **Wiki Article Cleaning**: for Wiki pages, we remove the Wiki template [2], special symbols (*e.g.*, ASCII symbols), and other invalid text[3].

(2) **Reference Cleaning**: for reference, we filter out articles that are shorter than 16 words or non-English token ratio > 0.3. We remove invalid text such as those containing only HTML tag, unusually long sentences (*e.g.*, > 256 tokens), text tables and URLs, etc. Furthermore, as the citations

[2]see details in `https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Templates`

[3]We observe some template text that may not be useful for model training (*e.g.*, `You can help Wikipedia by expanding it.`)

may be missing (uncrawlable or filtered out) or incorrectly cited (Piktus et al., 2021), we define the following metric to calculate the term recall of a reference to choose "reasonable" references:

$$P_{\text{ST}} = \frac{|(\mathcal{Q} \cup \mathcal{T}) \cap \mathcal{V} - \mathcal{S}|}{|\mathcal{T} - \mathcal{S}|}, \tag{3}$$

where $\mathcal{Q}$ is the term set of a query q, $\mathcal{T}$ is the term set of a Wiki article' sentence $t^{\leftarrow V^\tau}$ that cites a reference set V^τ, $\mathcal{V}$ is the term set of V^τ, and $\mathcal{S}$ is a set of stop words. Intuitively, if more terms in the sentence claim t can be found in the references V^τ, the references would be more useful. Therefore, we only keep references that have $P_{\text{ST}} > \rho$.

Reference Passage Selection The original Wiki articles and reference articles tend to be very long. To adapt the capacity of most pre-trained language models (*e.g.*, BERT has a 512 token limit), we use the first section of Wiki articles as the generation target, and we select the passage from the reference article with the highest P_{ST} value as the supporting passage. Specifically, we first split a reference article by sentences and concatenate the sentences into passages (stride $= 1$, max_length $= 256$) We then compute the P_{ST} value for each passage and keep the highest one.

Dataset Generation As defined in Eq. (2), given a query q, the WEBBRAIN task aims to retrieve the supporting evidence V from a web corpus C, so as to generate a factual article T_q. To enable training the in-domain retriever and generator for WEBBRAIN, we generate the WebBrain-R(etriever) and WebBrain-G(enerator) datasets from WebBrain-Raw. For WebBrain-G, we keep five supporting passages for each query [4], and retrospectively assign reference marks to all unmarked sentences in the Wiki article by measuring their P_{ST} value. We append a "`[0]`" mark to the sentence end for these fail to match any references. For WebBrain-R, we use the supporting passages as the positive and randomly select four negative passages from the Top30-50 retrieved results via a BM25 engine. Table 2 shows the statistics of the two datasets.

4 FRAMEWORK: REGEN

4.1 RETRIEVER

By the definition of WEBBRAIN in Eq. (2), ReGen starts from a retriever that mines the supporting evidence V from the large web corpus C. ReGen utilizes a parametric sparse lexical model SPLADE (Formal et al., 2021a;b) as the retriever, which encodes a sequence $s = (w_1, \cdots, w_n)$ into a sparse lexical representation $\mathbf{s} \in \mathbb{R}^{|D|}$ by predicting token-level importance in WordPiece vocabulary size (*e.g.*, $|D| = 30522$). Specifically, $\mathbf{s}$ is computed based on the dense representation $(\mathbf{h}_1, \cdots, \mathbf{h}_n)$ output by the underlying PLMs:

$$\mathbf{w}_i = \mathbf{E}\mathbf{W}\mathbf{h}_i^\top + \mathbf{b}, \quad i \in [1, n], \quad \mathbf{s} = \max_{i \in [1,n]} \log(1 + \text{ReLU}(\mathbf{w}_i)), \tag{4}$$

where $\mathbf{W} \in \mathbb{R}^{768 \times 768}$ and $\mathbf{b} \in \mathbb{R}^{|D|}$ are a trainable transition matrix and a bias, respectively, and $\mathbf{E} \in \mathbb{R}^{768 \times |D|}$ refers to the PLM's input embedding matrix. In the training phase, given a query q and a document d, the ranking score $g(q, d)$ is computed by the dot product of the their representations $\mathbf{s}_q$ and $\mathbf{s}_d$.

Previous studies that utilize a retriever in the generation loop usually apply the traditional term-match-based retriever (*e.g.*, BM25), dense passage retriever (*e.g.*, DPR), or a hybrid retriever (Karpukhin et al., 2020; Lewis et al., 2020b; Izacard & Grave, 2021; Glass et al., 2022) We build ReGen's retriever based on SPLADE because of the following reasons: (1) Factual queries in WEBBRAIN are usually open-domain and entity-centric, while recent work observed that DPR tend to perform poorly for entity-centric queries or the out-of-domain queries (Sciavolino et al., 2021; Xiong et al., 2021; Thakur et al., 2021). SPLADE, on the contrary, inherits the good properties of sparse retrievers such as exact-entity match and generalizability, making it suitable to WEBBRAIN. (2) SPLADE also inherits the contextualized semantics of the underlying PLMs, enabling the query encoder to perform contextualized query expansion (Formal et al., 2021a), which alleviates the vocabulary mismatch problem of BM25. (3) Empirical results also prove the effectiveness of SPLADE in WEBBRAIN (see Section 5.3).

[4] statistics shows that the first section of 96%+ Wiki pages has equal or less than five references.

Ranking Loss Generally, given a query q, a positive document d^+ and a set of negative documents d^-, the optimization goal of a retriever is to maximize the probability:

$$\mathbf{P}(q, d^+, \mathcal{D}) = \frac{\mathbf{e}^{f(q,d^+)}}{\mathbf{e}^{f(q,d^+)} + \sum_{d^- \in \mathcal{D}} \mathbf{e}^{f(q,d^-)}}, \tag{5}$$

where $\mathcal{D}$ represents the entire corpus. We mine the negative documents via a BM25 engine and also include in-batch negative samples into d^-. We train the model with a contrastive LCE loss (Gao et al., 2021).

Asymmetric Sparsity We introduce two sparsity constraints to control SPLADE's sparsity: L1 regularization and FLOPS regularization (Paria et al., 2020). For the query encoder and document encoder, we apply asymmetric sparsity regularization. We use the L1 regularization with a small weight λ_q for the query encoder and FLOPS regularization with a big weight λ_d for the doc encoder. Thus, the final loss for SPLADE becomes:

$$\mathcal{L} = \mathcal{L}_{\texttt{rank}} + \lambda_q \mathcal{L}_{\texttt{L1}} + \lambda_d \mathcal{L}_{\texttt{FLOPS}}. \tag{6}$$

We apply the asymmetric sparsity because: (1) a small λ_q lowers the sparsity of the query encoder, enabling the query encoder to generate semantically-rich query representation; (2) the FLOPS regularization is not sensitive to the number of documents in a batch so that we can compute $\mathcal{L}_{\texttt{FLOPS}}$ on both positive and negative samples; (3) a big λ_d increases the sparsity of doc encoder, making the off-line document index more efficient, and producing a self-denoising effect – more unimportant tokens tend to be ignored.

Topic Consistency and Diversity For a document pair (d_m, d_n) and their $|D|$-size sparse representations $(\mathbf{s}_m, \mathbf{s}_n)$, we measure their topic distance by:

$$d_T(d_m, d_n) = \texttt{distance}(\mathbf{s}_m, \mathbf{s}_n) = \frac{|\mathbf{s}'_m - \mathbf{s}'_n|}{|\mathbf{s}'_m| + |\mathbf{s}'_n|}, \quad \mathbf{s}'[s_j > \mu] = 1, \quad j \in |D|, \tag{7}$$

where $\mathbf{s} = (s_1, \cdots, s_j)$, $\mathbf{s}'$ is a binary variant of $\mathbf{s}$, and μ is a threshold to eliminate tokens with low importance. For (d_m, d_n), when $d_T \to 0$, the two documents tend to have the same important tokens and are likely to be redundant; whereas when $d_T \to 1$, the two documents tend to have non-overlapping important tokens and are likely topically irrelevant. In the ReGen framework, we propose a strategy to maintain the topic consistency and avoid redundancy: Considering k sorted retrieved evidence documents $(d_1, \cdots, d_k)$, we first filter out the documents d_i that have $d_T(d1, d_i) > 0.9, i \in [2, k]$ (topically-irrelevant to the top-1 document). We then remove documents d_i that have $d_T(d_i, d_j) < 0.1, j \in [1, i-1]$ (redundant to fore-rank documents).

4.2 Generator

ReGen employs a sequence-to-sequence model with a Fusion-in-Decoder (FiD) structure (Izacard & Grave, 2021) as the generator. The generator takes the query and references as input, and generates a factual description. More precisely, given a query q and a set of k references $\{v_1, v_2, \cdots, v_k\}$ returned by the retriever, we add special tokens before the query/reference and concatenate the query with each reference as:

$$x_i = \texttt{[query]}\, q\, \texttt{[ref_i]}\, v_i. \tag{8}$$

These sequences are processed independently by the encoder as:

$$\mathbf{h}_i = \texttt{Encoder}(x_i), \quad i = 1, 2, \cdots, k. \tag{9}$$

Finally, the decoder conducts attention over the concatenation of the representations of all the sequences, and outputs a text sequence T^* with auto-regressive mechanism:

$$T^* \sim \texttt{Decoder}(T^*, \mathbf{h}) = \prod_{n=1}^{|T^*|} p(T^* | T^*_{<n}, \mathbf{h}), \qquad \mathbf{h} = [\mathbf{h}_1; \mathbf{h}_2; \cdots; \mathbf{h}_k]. \tag{10}$$

The model can perform evidence fusion in the decoder and generate more factual claims with the references. The generator is optimized via minimizing the negative log-likelihood function of the ground-truth text:

$$\mathcal{L}_{\text{NLL}}(x, T^*) = -\sum_{n=1}^{|T^*|} \log p(T^*_n | T^*_{<n}, x). \tag{11}$$

Model Warm-Up for Sentence-Reference Matching In our task, the model is trained to generate claims for a query based on several references. This task is difficult as the model should first find clues from various references and then determine which part is the most relevant in the current generation status. To facilitate the model's capability of extracting key information from the reference, we design a model warm-up stage. Specifically, for a target text, we break it into multiple sentences and extract the ones that are marked with references. As a result, we can obtain several ⟨reference, sentence⟩ pairs. Then, we concatenate the query with the reference, and train the model to generate the corresponding sentence. This warm-up stage has **two advantages**: (1) The task is to generate a sentence with a single reference. Compared with the final generation task (*i.e.*, generating multiple sentences with multiple references), this warm-up task is much simpler. Therefore, the overall training follows a curriculum learning paradigm, *i.e.*, learning from easy task to hard task. This learning paradigm has shown to be effective in improving performance (Bengio et al., 2009; Hacohen & Weinshall, 2019). (2) Existing models often apply general tasks for pre-training. For example, BART (Lewis et al., 2020a) uses several sequence denoising tasks for pre-training. These tasks have different objectives from text generation. Our proposed warm-up task is also a text generation task, so it can help model mitigate the gap of tasks between pre-training and fine-tuning.

5 EXPERIMENTS

5.1 SETTINGS

ReGen applies SPLADE as the basis of retriever Formal et al. (2021a), improves the vanilla SPLADE with a training strategy of asymmetric sparsity and uses a ranking strategy considering documents' topic consistency and diversity. For comparison, we report the performance of the vanilla SPLADE, DPR (Karpukhin et al., 2020) and BM25. We use RetroMAE as the underlying PLM of DPR, which is shown to be a state-of-the-art model in many retrieval benchmarks (Liu & Shao, 2022). For the generator, we use FiD+BART as the backbone model. As a comparison, we report the performance of BART (Lewis et al., 2020a), GPT-2 (Radford et al., 2019), and vanilla FiD+BART (Izacard & Grave, 2021). All models are initialized with the off-the-shelf checkpoints provided by HuggingFace (Wolf et al., 2019). The implementation details are in Appendix A.

To evaluate the performance, we employ several automatic metrics as follows and perform a human annotation: **BLEU** (Papineni et al., 2002), **METEOR** (Banerjee & Lavie, 2005), **ROUGE** (Lin, 2004), and **CIDEr** (Vedantam et al., 2015): These are metrics based on n-gram overlapping between generated text and ground-truth text. A higher value indicates the generated text is more similar to the ground truth. We compute the metrics using the `nlg-eval` package (Sharma et al., 2017). **QAGS** (Wang et al., 2020) and **Triple Score** (Goodrich et al., 2019) are two metrics for evaluating factualness. We also compute them between generated text and ground-truth text.[5] For QAGS, several questions are generated from the generated text, which is then answered by both the ground-truth text and the generated text. A higher value reflects that more identical answers are obtained, which further suggests that the generated text is more factual. To compute Triple Score, the relations in both generated text and ground-truth text are first extracted by OpenIE (Angeli et al., 2015). Then, the score is computed by the precision of relations in the generated text. A higher score indicates that more relations are correctly generated. We compute QAGS and triple score by `FactSumm` package (Heo, 2021). As for human annotation, we compare the result of our ReGen with that of baselines with respect to fluency, informativeness, and faithfulness. Detailed description of the annotation criteria is given in Appendix B.

5.2 OVERALL RESULTS

The results of various models are shown in Table 3 and Table 5. As can be seen, our ReGen model (in both sizes) achieves the best results on all metrics. This clearly demonstrates the effectiveness of our framework. Specifically, (1) compared with BART and GPT-2, FiD and ReGen achieve better performance, especially on factualness metrics (QAGS and triple score). This indicates that the knowledge from Web corpus is critical for WEBBRAIN task. (2) Large models can always perform better than base models, but the main improvement is made on language quality (*e.g.*, BLEU

[5]These scores can also be computed between generated text and reference, but how to combine the scores based on multiple references is still under-explored.

Table 3: Performance of different models on WEBBRAIN corpus. The best results are in bold.

Model	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr	QAGS	TripleScore
BART (Base)	8.16	2.98	9.28	23.08	38.30	28.49	10.91
BART (Large)	11.63	4.55	10.49	24.45	45.54	28.61	12.47
GPT-2	9.35	2.85	8.57	20.22	25.09	25.61	7.30
FiD+BART (Base)	11.20	4.72	12.60	27.42	51.32	39.88	15.89
FiD+BART (Large)	15.90	6.89	14.03	28.64	57.06	39.52	17.36
ReGen (Base)	12.11	5.26	12.98	27.81	52.92	40.00	16.19
ReGen (Large)	**17.97**	**8.82**	**15.31**	**30.09**	**62.39**	**40.32**	**19.43**

Table 4: Retrieval (left) and generation (right) performance of ReGen with different retrievers. The best results are in bold.

Model	R@1	R@5	MAP	BLEU-1	METEOR	ROUGE-L	CIDEr	TripleScore
ReGen	**0.399**	**0.746**	**0.550**	24.71	12.99	**21.19**	**20.12**	**10.97**
+ BM25	0.290	0.558	0.410	**25.76**	13.73	19.94	14.85	10.43
+ DPR	0.228	0.390	0.301	20.42	12.71	16.11	6.44	6.05
+ SPLADE	0.389	0.724	0.536	23.84	**14.24**	18.98	10.80	9.48
Golden	-	-	-	17.97	15.31	30.09	62.39	19.43

and ROUGE) rather than factualness (*e.g.*, QAGS). This is consistent with our assumption that the external knowledge is hard to be "remembered" in model parameters. (3) Compared with FiD, we design a warm-up stage for ReGen model. This strategy is effective regardless of model size, but works better for larger model. It suggests that larger models with more capacity are promising for better results, but their potential needs to be activated through well-tailored tasks.

5.3 DISCUSSION

Impact of Retriever To explore how different retrievers impact the end-to-end performance of ReGen, we replace ReGen's underlying retriever with BM25, DPR (RetroMAE), and SPLADE. From the results shown in Table 4, we can make the following observations: (1) Compared to using golden references, the BLEU-1 score increases while the other metrics decrease, indicating that the retrieved knowledge enriches the informativeness of generation but undermines the generation quality. This implies that the retrieval quality is crucial for the factuality of generation and how to retrieve helpful knowledge remains a challenge in the WEBBRAIN task. (2) ReGen surpasses the other three retrievers regarding ROUGE-L, CIDEr, and TripleScore, which implies that ReGen could retrieve more helpful knowledge. With such knowledge, ReGen could generate better answers regarding factuality, confirming the effectiveness of the fact-enhance strategies applied in ReGen. (3) Regarding retrieval performance, our Regen outperforms all baselines on all metrics (see full retrieval results in Table 10).

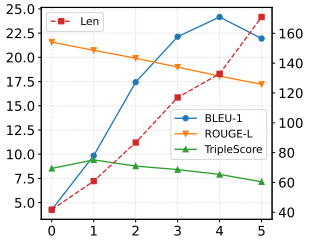

Figure 2: Performance of ReGen with different numbers of retrieved references.

We also evaluate ReGen's performance with different numbers of retrieved references. Specifically, we randomly sample 100 queries from the test set and feed them into ReGen's generator together with zero to five references retrieved by ReGen's retriever. Figure 2 illustrates the results. Generally, with more references, BLEU score and text length increase, but ROUGE score and triple score decrease. This implies that more content is generated but the quality cannot be guaranteed. The BLEU score achieves the best with four references. Combining with the retriever's performance, we attribute this effect to the noisy nature of retrieved references. Finally, we note that, even with references, generating long and faithful text is still a challenging problem.

Impact of Number of References As references provide evidence for generating factual content, their quantity will impact the quality of generation. To investigate this, we randomly sample 500

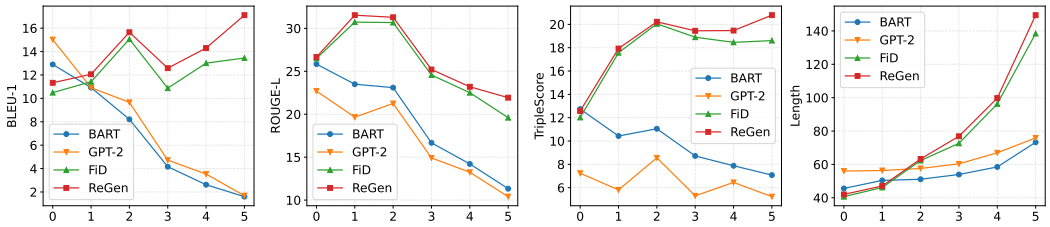

Figure 3: Performance of different models with different numbers of ground-truth references.

queries from the test set with zero to five references respectively (3,000 in total). Different from the previous analysis, we use ground-truth references in this experiment. Figure 3 shows the results. We observe: (1) For models using no reference (*i.e.*, BART and GPT-2), their performance is decreasing ($0 \rightarrow 5$ references). Intuitively, target texts using more references are much more difficult for generation, as they require more knowledge to construct the content. This result validates our assumption that writing human-like descriptions for queries heavily relies on external knowledge, which is difficult for pre-trained models to store as parameters. (2) The performance of FiD and ReGen is improving in terms of BLEU and TripleScore. This is because target texts with more references are usually longer, and the references can provide more information for generation. However, as suggested by ROUGE, more noise also appears in the generated text.

Performance vs. GPT-3 GPT-3 (Brown et al., 2020) is a large-scale pre-trained language model with 175B parameters. The model demonstrates strong few- or zero-shot performance on many text-based tasks, as vast knowledge has been stored in its parameters. To compare this strong model with ours, we first randomly sample 100 queries from the test set. Then, for GPT-3, we use the prompt `Introduce [X]:` to generate zero-shot results.[6] For our model, we use the query and evidence provided by the retriever as input. The results on automatic metrics and human annotation are shown in Figure 4 and the last row of Table 5. Our model outperforms GPT-3 by a large margin in terms of BLEU, METEOR, and ROUGE, This imply that our model can generate texts that are closer to the ground-truth (written by human), indicating that the explicit knowledge in web corpus is essential for generating natural and informative text. Only increasing the size of the language model would fail to capture such knowledge. We also observe a clear improvement of our model on factualness metrics. Even with a large amount of parameters, it is evident that GPT-3 cannot guarantee the factualness. In contrast, evident from web corpus can reduce these hallucinations.

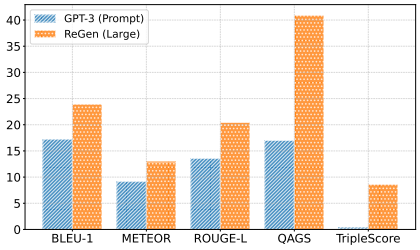

Figure 4: Performance of GPT-3 (Prompt) and our model (Large) on 100 test samples.

Due to limited space, we present (1) a reference mark correction strategy in Appendix C; (2) a case study in Appendix D; (3) a system demonstration in Appendix E; and (4) future work in Appendix F.

6 Conclusion

In this paper, we introduce a new NLP task, WebBrain, which aims to generate a fluent, informative, and factually-correct description for factual queries via mining supporting evidence from a large web corpus. We construct a large-scale dataset WebBrain-raw as the data foundation of WebBrain. WebBrain-raw is constructed by crawling all Wikipedia articles and the corresponding reference articles. Through empirical analysis, we found that most current NLP techniques lack the ability to maintain the factualness on the WebBrain task. To improve the performance on this task, we propose a new framework ReGen which uses several specifically designed strategies. It outperforms all baselines on automatic and human evaluation. We believe that WebBrain would highlight a valuable research pathway in which AI models could acquire knowledge from the Web world by themselves and serve human beings by fulfilling a broader range of fact-oriented information needs.

[6]We use the model `text-davinci-002` provided in `https://beta.openai.com/playground`.

REFERENCES

Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pp. 344–354. The Association for Computer Linguistics, 2015. doi: 10.3115/v1/p15-1034. URL https://doi.org/10.3115/v1/p15-1034.

Satanjeev Banerjee and Alon Lavie. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare R. Voss (eds.), *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, pp. 65–72. Association for Computational Linguistics, 2005. URL https://aclanthology.org/W05-0909/.

Siddhartha Banerjee and Prasenjit Mitra. Wikiwrite: Generating wikipedia articles automatically. In Subbarao Kambhampati (ed.), *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pp. 2740–2746. IJCAI/AAAI Press, 2016. URL http://www.ijcai.org/Abstract/16/389.

Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, and Hsiao-Wuen Hon. Unilmv2: Pseudo-masked language models for unified language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 642–652. PMLR, 2020. URL http://proceedings.mlr.press/v119/bao20a.html.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In Andrea Pohoreckyj Danyluk, Léon Bottou, and Michael L. Littman (eds.), *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, volume 382 of *ACM International Conference Proceeding Series*, pp. 41–48. ACM, 2009. doi: 10.1145/1553374.1553380. URL https://doi.org/10.1145/1553374.1553380.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow, March 2021. URL https://doi.org/10.5281/zenodo.5297715. If you use this software, please cite it using these metadata.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Deng Cai, Yan Wang, Huayang Li, Wai Lam, and Lemao Liu. Neural machine translation with monolingual translation memory. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pp. 7307–7318. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.567. URL https://doi.org/10.18653/v1/2021.acl-long.567.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and

Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-1423. URL https://doi.org/10.18653/v1/n19-1423.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of wikipedia: Knowledge-powered conversational agents. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=r1l73iRqKm.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 13042–13054, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/c20bb2d9a50d5ac1f713f8b34d9aac5a-Abstract.html.

Hady ElSahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon S. Hare, Frédérique Laforest, and Elena Simperl. T-rex: A large scale alignment of natural language with knowledge base triples. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Kôiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga (eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA), 2018. URL http://www.lrec-conf.org/proceedings/lrec2018/summaries/632.html.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. SPLADE v2: Sparse lexical and expansion model for information retrieval. *CoRR*, abs/2109.10086, 2021a. URL https://arxiv.org/abs/2109.10086.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. SPLADE: sparse lexical and expansion model for first stage ranking. In Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai (eds.), *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pp. 2288–2292. ACM, 2021b. doi: 10.1145/3404835.3463098. URL https://doi.org/10.1145/3404835.3463098.

Luyu Gao, Zhuyun Dai, and Jamie Callan. Rethink training of BERT rerankers in multi-stage retrieval pipeline. In Djoerd Hiemstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani (eds.), *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part II*, volume 12657 of *Lecture Notes in Computer Science*, pp. 280–286. Springer, 2021. doi: 10.1007/978-3-030-72240-1_26. URL https://doi.org/10.1007/978-3-030-72240-1_26.

Michael R. Glass, Gaetano Rossiello, Md. Faisal Mahbub Chowdhury, Ankita Naik, Pengshan Cai, and Alfio Gliozzo. Re2g: Retrieve, rerank, generate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pp. 2701–2715. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.naacl-main.194. URL https://doi.org/10.18653/v1/2022.naacl-main.194.

Ben Goodrich, Vinay Rao, Peter J. Liu, and Mohammad Saleh. Assessing the factual accuracy of generated text. In Ankur Teredesai, Vipin Kumar, Ying Li, Rómer Rosales, Evimaria Terzi, and George Karypis (eds.), *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, pp. 166–175. ACM, 2019. doi: 10.1145/3292500.3330955. URL https://doi.org/10.1145/3292500.3330955.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor O. K. Li. Search engine guided neural machine translation. In Sheila A. McIlraith and Kilian Q. Weinberger (eds.), *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 5133–5140. AAAI Press, 2018. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17282.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. REALM: retrieval-augmented language model pre-training. *CoRR*, abs/2002.08909, 2020. URL https://arxiv.org/abs/2002.08909.

Guy Hacohen and Daphna Weinshall. On the power of curriculum learning in training deep networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2535–2544. PMLR, 2019. URL http://proceedings.mlr.press/v97/hacohen19a.html.

Hoon Heo. Factsumm: Factual consistency scorer for abstractive summarization. https://github.com/Huffon/factsumm, 2021.

Robert Iv, Alexandre Passos, Sameer Singh, and Ming-Wei Chang. FRUIT: Faithfully reflecting updated information in text. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3670–3686, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.269. URL https://aclanthology.org/2022.naacl-main.269.

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pp. 874–880. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.eacl-main.74. URL https://doi.org/10.18653/v1/2021.eacl-main.74.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 6769–6781. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.550. URL https://doi.org/10.18653/v1/2020.emnlp-main.550.

Philipp Koehn and Rebecca Knowles. Six challenges for neural machine translation. In Thang Luong, Alexandra Birch, Graham Neubig, and Andrew M. Finch (eds.), *Proceedings of the First Workshop on Neural Machine Translation, NMT@ACL 2017, Vancouver, Canada, August 4, 2017*, pp. 28–39. Association for Computational Linguistics, 2017. doi: 10.18653/v1/w17-3204. URL https://doi.org/10.18653/v1/w17-3204.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://doi.org/10.1162/tacl_a_00276.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 7871–7880. Association for Computational Linguistics, 2020a. doi: 10.18653/v1/2020.acl-main.703. URL https://doi.org/10.18653/v1/2020.acl-main.703.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020b. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

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. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=Hyg0vbWC-.

Yang Liu and Mirella Lapata. Hierarchical transformers for multi-document summarization. In Anna Korhonen, David R. Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 5070–5081. Association for Computational Linguistics, 2019. doi: 10.18653/v1/p19-1500. URL https://doi.org/10.18653/v1/p19-1500.

Zheng Liu and Yingxia Shao. Retromae: Pre-training retrieval-oriented transformers via masked auto-encoder. *CoRR*, abs/2205.12035, 2022. doi: 10.48550/arXiv.2205.12035. URL https://doi.org/10.48550/arXiv.2205.12035.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, H. Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese. Teaching language models to support answers with verified quotes. *CoRR*, abs/2203.11147, 2022. doi: 10.48550/arXiv.2203.11147. URL https://doi.org/10.48550/arXiv.2203.11147.

Donald Metzler, Yi Tay, Dara Bahri, and Marc Najork. Rethinking search: making domain experts out of dilettantes. *SIGIR Forum*, 55(1):13:1–13:27, 2021. doi: 10.1145/3476415.3476428. URL https://doi.org/10.1145/3476415.3476428.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. Webgpt: Browser-assisted question-answering with human feedback. *CoRR*, abs/2112.09332, 2021. URL https://arxiv.org/abs/2112.09332.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pp. 311–318. ACL, 2002. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Biswajit Paria, Chih-Kuan Yeh, Ian En-Hsu Yen, Ning Xu, Pradeep Ravikumar, and Barnabás Póczos. Minimizing flops to learn efficient sparse representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=SygpC6Ntvr.

Laura Perez-Beltrachini, Yang Liu, and Mirella Lapata. Generating summaries with topic templates and structured convolutional decoders. In Anna Korhonen, David R. Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 5107–5116. Association for Computational Linguistics, 2019. doi: 10.18653/v1/p19-1504. URL https://doi.org/10.18653/v1/p19-1504.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick S. H. 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 Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pp. 2523–2544. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.200. URL https://doi.org/10.18653/v1/2021.naacl-main.200.

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.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. The curious case of hallucinations in neural machine translation. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pp. 1172–1183. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.92. URL https://doi.org/10.18653/v1/2021.naacl-main.92.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pp. 4035–4045. Association for Computational Linguistics, 2018. doi: 10.18653/v1/d18-1437. URL https://doi.org/10.18653/v1/d18-1437.

Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. Simple entity-centric questions challenge dense retrievers. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pp. 6138–6148. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.496. URL https://doi.org/10.18653/v1/2021.emnlp-main.496.

Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *CoRR*, abs/1706.09799, 2017. URL http://arxiv.org/abs/1706.09799.

Philipp Singer, Florian Lemmerich, Robert West, Leila Zia, Ellery Wulczyn, Markus Strohmaier, and Jure Leskovec. Why we read wikipedia. In Rick Barrett, Rick Cummings, Eugene Agichtein, and Evgeniy Gabrilovich (eds.), *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, pp. 1591–1600. ACM, 2017. doi: 10.1145/3038912.3052716. URL https://doi.org/10.1145/3038912.3052716.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MASS: masked sequence to sequence pre-training for language generation. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5926–5936. PMLR, 2019. URL http://proceedings.mlr.press/v97/song19d.html.

Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. ERNIE: enhanced representation through knowledge integration. *CoRR*, abs/1904.09223, 2019. URL http://arxiv.org/abs/1904.09223.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *CoRR*, abs/2104.08663, 2021. URL https://arxiv.org/abs/2104.08663.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and verification. In Marilyn A. Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pp. 809–819. Association for Computational Linguistics, 2018. doi: 10.18653/v1/n18-1074. URL https://doi.org/10.18653/v1/n18-1074.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 4566–4575. IEEE Computer Society, 2015. doi: 10.1109/CVPR.2015.7299087. URL https://doi.org/10.1109/CVPR.2015.7299087.

Oriol Vinyals and Quoc V. Le. A neural conversational model. *CoRR*, abs/1506.05869, 2015. URL http://arxiv.org/abs/1506.05869.

Alex Wang, Kyunghyun Cho, and Mike Lewis. Asking and answering questions to evaluate the factual consistency of summaries. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 5008–5020. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.450. URL https://doi.org/10.18653/v1/2020.acl-main.450.

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. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pp. 4003–4012. European Language Resources Association, 2020. URL https://aclanthology.org/2020.lrec-1.494/.

Jason Weston, Emily Dinan, and Alexander H. Miller. Retrieve and refine: Improved sequence generation models for dialogue. In Aleksandr Chuklin, Jeff Dalton, Julia Kiseleva, Alexey Borisov, and Mikhail S. Burtsev (eds.), *Proceedings of the 2nd International Workshop on Search-Oriented Conversational AI, SCAI@EMNLP 2018, Brussels, Belgium, October 31, 2018*, pp. 87–92. Association for Computational Linguistics, 2018. doi: 10.18653/v1/w18-5713. URL https://doi.org/10.18653/v1/w18-5713.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771, 2019. URL http://arxiv.org/abs/1910.03771.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual*

Event, Austria, May 3-7, 2021. OpenReview.net, 2021. URL https://openreview.net/forum?id=zeFrfgyZln.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. ERNIE: enhanced language representation with informative entities. In Anna Korhonen, David R. Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 1441–1451. Association for Computational Linguistics, 2019. doi: 10.18653/v1/p19-1139. URL https://doi.org/10.18653/v1/p19-1139.

Yutao Zhu, Zhicheng Dou, Jian-Yun Nie, and Ji-Rong Wen. Reboost: a retrieval-boosted sequence-to-sequence model for neural response generation. *Inf. Retr. J.*, 23(1):27–48, 2020. doi: 10.1007/s10791-019-09364-x. URL https://doi.org/10.1007/s10791-019-09364-x.

Table 5: Human annotation results. The first three pairs are all base models.

	Fluency			Informativeness			Faithfulness		
Comparison	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose
ReGen vs. BART	39%	52%	9%	29%	63%	8%	42%	46%	12%
ReGen vs. GPT-2	42%	43%	15%	30%	59%	11%	54%	30%	16%
ReGen vs. FiD+BART	22%	63%	15%	17%	74%	9%	37%	51%	12%
ReGen (Large) vs. GPT-3 (Prompt)	48%	22%	30%	34%	51%	15%	72%	4%	24%

Table 6: Performance of models on samples with different numbers of retrieved references.

#Refs.	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr	QAGS	TripleScore	Len
0	4.27	1.48	7.67	21.58	46.99	17.14	8.53	41.71
1	9.85	3.19	8.99	20.72	32.89	50.93	9.41	60.87
2	17.43	5.68	11.37	19.91	16.49	48.21	8.77	86.75
3	22.13	5.84	12.14	18.99	18.03	43.54	8.41	116.98
4	24.16	6.15	12.94	18.06	8.20	40.34	7.91	132.88
5	21.94	4.68	13.13	17.21	4.75	40.73	7.15	170.93

A Implementation Details

In our ReGen, the retriever is training on the WebBrain-R dataset for five epochs, initializing from `efficient-splade-V-large`[7]. We set the batch size to 192, the max query length to 16, and the max passage length to 256. We use the AdamW optimizer with a learning rate of 2e-5. We set $(\lambda_q, \lambda_d) = (5e-4, 5e-3)$ for the Eq. 6. For baseline models, we train SPLADE and DPR with their official settings on the WebBrain-R dataset for five epochs. The underlying RetroMAE of DPR is initialized from the BEIR checkpoint [8]. We implement the models with Pytorch and Pytorch-Lightning [9]. We train the models on 8 Tesla V100 32G GPUs. For the BM25 retriever, we use the Pyserini tool with its default parameter setting (Lin et al., 2021). As for the generator, we use BART-based and BART-large checkpoints provided by HuggingFace (Wolf et al., 2019) to initialize the model. For training, we use 32 and 64 Tesla V100 32G GPUs for base and large models, respectively. Correspondingly, the batch size of each GPU is set as 8 and 4. AdamW optimizer (Loshchilov & Hutter, 2019) is used for optimization with a learning rate of 5e-5. In inference stage, we apply beam search (width=5), and the maximum generation length is set as 512. This generation setting is also used for GPT-3 (prompt).

B Human Evaluation

To further evaluate our ReGen model, we perform human evaluation in terms of three metrics: (1) **Fluency** evaluates the linguistic coherency of the generated text;(2) **Informativeness** measures whether the generated text gives a response in specific details based on the query and references, (3) **Faithfulness** reflects the proportion of the generated text that can be supported by the retrieved supporting evidence. To reduce the influence of objective factors in the manual annotation, we perform a pairwise evaluation method: by giving the results generated by ReGen and other baseline models, three annotators are asked to judge which text is better or make a tie. For fairness, the results from different models are randomly shuffled, and the annotators would not know the underlying model in advance. The evaluation results are shown in Table 5.

[7] `https://huggingface.co/naver/efficient-splade-V-large-query` and `https://huggingface.co/naver/efficient-splade-V-large-doc`

[8] `https://huggingface.co/SamuelYang/SentMAE_BEIR`

[9] `https://pytorch-lightning.readthedocs.io/en/stable/`

Table 7: Performance of models on samples with different numbers of ground-truth references.

Model	#Refs.	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr	QAGS	TripleScore	Len
BART	0	12.90	4.87	11.28	25.84	49.31	25.22	12.75	45.60
	1	10.92	4.11	10.04	23.50	39.45	29.40	10.43	50.42
	2	8.21	3.59	10.11	23.10	40.13	30.70	11.05	51.10
	3	4.16	1.20	7.39	16.67	13.16	27.75	8.72	53.96
	4	2.62	1.08	7.20	14.21	10.91	25.03	7.88	58.49
	5	1.60	0.41	6.13	11.33	1.96	25.22	7.08	73.29
GPT-2	0	15.02	4.11	10.34	22.70	27.94	23.88	7.24	56.01
	1	10.91	2.96	8.75	19.66	18.84	25.32	5.81	56.35
	2	9.66	3.80	9.84	21.25	33.77	30.09	8.56	57.57
	3	4.72	1.19	6.70	14.91	8.99	25.89	5.30	60.34
	4	3.53	1.29	6.82	13.24	10.36	22.89	6.45	66.95
	5	1.68	0.33	5.60	10.40	1.49	23.08	5.23	75.90
FiD	0	10.49	4.18	10.84	26.50	49.18	20.68	12.02	40.68
	1	11.42	5.16	13.06	30.73	70.52	59.63	17.57	46.17
	2	15.06	7.18	15.51	30.68	66.91	57.91	20.05	62.31
	3	10.87	4.04	13.48	24.57	29.23	59.46	18.90	72.64
	4	13.01	5.58	14.96	22.51	17.81	55.26	18.46	96.29
	5	13.45	4.97	14.36	19.60	14.30	50.64	18.61	138.41
ReGen	0	11.33	4.39	11.11	26.66	50.04	19.62	12.57	42.02
	1	12.06	5.67	13.36	31.54	76.35	59.05	17.93	47.10
	2	15.65	7.76	15.78	31.29	67.90	59.39	20.22	63.27
	3	12.58	4.84	14.09	25.21	36.29	58.84	19.45	76.87
	4	14.30	6.17	6.17	23.20	20.48	55.13	19.47	99.75
	5	17.11	7.24	15.95	21.93	19.33	50.60	20.79	149.45

Table 8: Influence of reference mark-enhanced decoding.

Model	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr	QAGS	TripleScore
ReGen (Base)	12.11	5.26	12.98	27.81	52.92	40.00	16.19
ReGen (Base) + RED	11.89	5.18	12.51	27.94	52.79	40.27	16.19
ReGen (large)	**17.97**	**8.82**	**15.31**	30.09	**62.39**	40.32	**19.43**
ReGen (Large) + RED	17.77	8.71	14.90	**30.23**	62.20	**40.42**	19.41

C Reference Mark Correction

Reference mark is used to indicate the source of a referenced text. Correct reference marks can directly improve the factualness and readability of the generated text. However, as the FiD model simultaneously performs attention on the representations of all tokens in multiple references, it is hard to identify the correct source of a generated sentence. We try to alleviate this problem by a **Reference mark-Enhanced Decoding** (RED) strategy: in inference stage, if a reference mark (other than `[0]`) has the largest generation probability at a decoding step, we extract the last generated sentence and match it with each reference. Then, the reference with the largest word overlapping ratio is selected as the source of the sentence (*i.e.*, its reference mark is used for this sentence). To test this strategy, for each generated text, we compute a `refer score` as: $P_r = \frac{1}{n}\sum\frac{|\mathcal{T}\cap\mathcal{V}-\mathcal{S}|}{|\mathcal{V}-\mathcal{S}|}$, where n is the number of sentences with references; $\mathcal{T}$ is the set of words in the sentence; $\mathcal{V}$ is the set of words in the corresponding reference; and $\mathcal{S}$ is the set of stopwords. This score is similar to STP score defined in Eq. 3 but focuses on how much content of sentence is quoted from the corresponding reference. From the results shown in Figure 5, we can see this strategy improves refer score greatly but has less influence on other metrics. This indicates that the reference mark is labeled more accurately. We apply this strategy in our demo (described in Appendix E), which also improves the demo's readability. However, our strategy is just an early exploration on reference mark accuracy. It is only involved in the inference stage, but not in model training. Moreover, how to evaluate the reference mark accuracy is also an unsolved problem. We leave these problems as our future work.

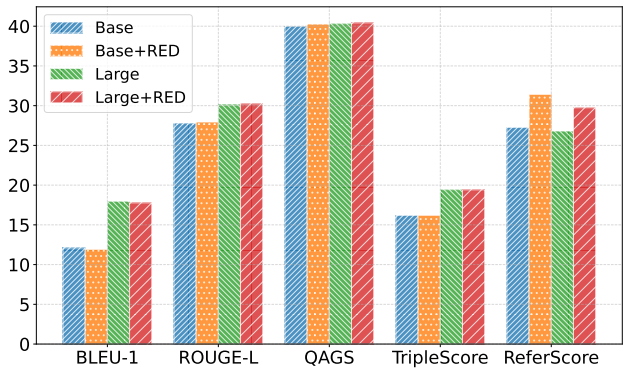

Figure 5: Influence of reference mark-enhanced decoding.

Table 9: Performance of GPT-3 (Prompt) and our ReGen (Large) on 100 samples.

Model	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr	QAGS	TripleScore
GPT-3 (Prompt)	17.24	2.19	9.14	13.57	6.7	17.00	0.29
ReGen (Large)	23.93	6.03	12.90	20.43	30.57	40.90	8.61

Table 10: Performance of various retrievers.

Model	R@1	P@1	R@5	R@10	R@20	MRR	MAP
BM25	0.290	0.292	0.558	0.645	0.716	0.411	0.410
DPR (BERT)	0.192	0.193	0.348	0.403	0.457	0.263	0.263
DPR (RetroMAE)	0.228	0.229	0.390	0.438	0.478	0.301	0.301
SPLADE	0.389	0.392	0.724	0.815	0.877	0.537	0.536
ReGen	**0.399**	**0.401**	**0.746**	**0.839**	**0.896**	**0.551**	**0.550**

D Case Study

To make a qualitative analysis of our model, we conduct a brief case study, as shown in Table 11. The green texts are correct information, while red ones are incorrect. In the first case, as no reference is given, FiD, BART, and ReGen generates similar results, but ReGen can provide more details. We attribute this to the warm-up strategy, it can effectively mitigate the gap of tasks between BART's pre-training and our fine-tuning. In the second case, with one reference provided, both FiD and ReGen can generate more faithful text than BART. This clearly reflects the important of external knowledge. Intriguingly, we can also observe that ReGen can better capture clues from the reference. The result also demonstrates the effectiveness of our proposed warm-up strategy, as it can facilitate our model's capability of sentence-reference matching.

E System Demonstration

Though there is a gap between our study and our ultimate goal of generating Wikipedia articles, we still explore potential application scenarios for our current model. We build a demonstration system as shown in Figure 6. This system can accept a user query and extract the keywords. Then, several references are retrieved from ReGen's retriever, based on which the generator can create a short article for the query. Furthermore, with the reference mark-enhanced decoding strategy, the reference mark is generated more accurately, which greatly enhance the readability of the generated text. This system can help us obtain some basic knowledge of entities.

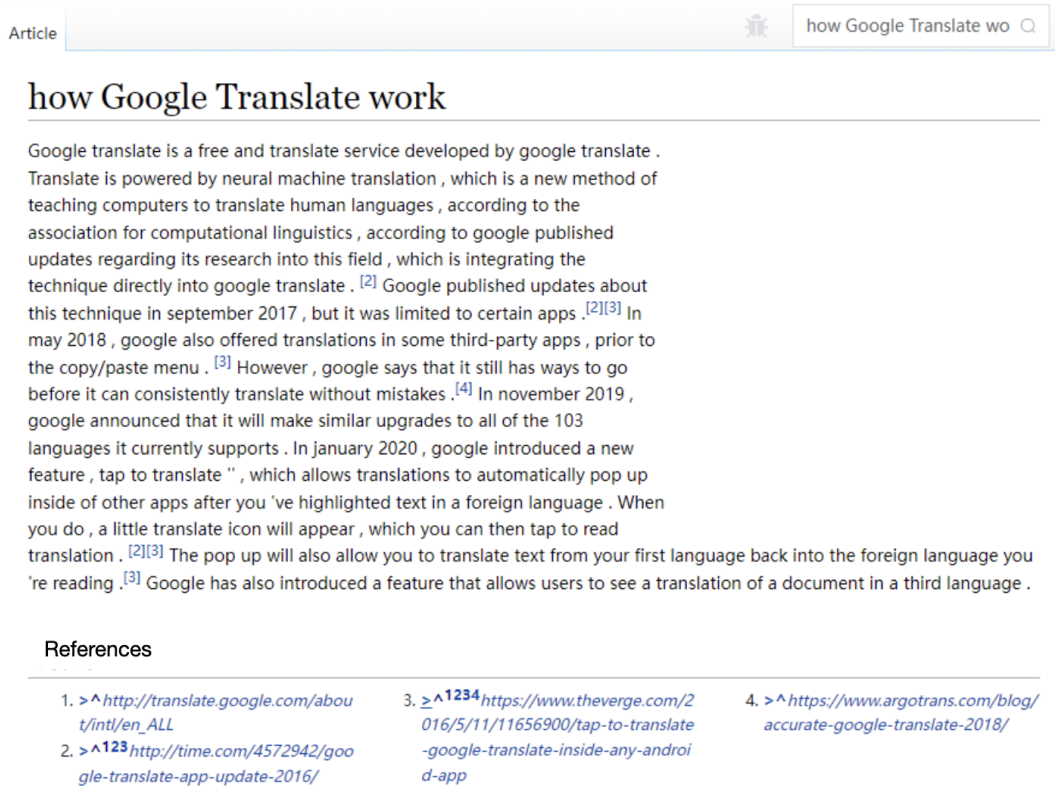

Figure 6: Screen shot of our system.

F Limitation and Future Work

In this work, we explore generating short factual articles (like Wikipedia page) for queries by collecting evidence from the Web. This is a very early step towards our ultimate goal: automatically building Wikipedia pages for unseen topics. We collect a new dataset to support future research, evaluate several existing models on the new task, and propose some effective strategies. However, there are still many limitations in our study, which needs further investigation in the future.

(1) The retriever is far less than perfect. As shown in our experiments, ReGen's performance with retrieved evidence is worse than that with golden references. So, two problems appear: (a) how to improve the retriever's performance; and (b) how to generate better result with non-perfect references.

(2) As we use FiD as our backbone model, it is hard to determine the reference mark accurately. Moreover, the reference mark accuracy cannot be measured by existing metrics. We plan to design new attention mechanism to facilitate the model's judgement on these reference marks.

(3) In this study, we try to generate a short article for given queries. However, the original Wikipedia page contains more information than a short article. Generating long articles will be another challenge in the future.

Table 11: Some cases generated by different models.

Query: 2016 Pateros local elections
Ground-truth reference: None
Ground-truth target: Local elections in pateros were held on may 9, 2016 within the philippine general election. The voters elected for the elective local posts in the city: the mayor, vice mayor, district representative, and councilors, six in each of the city's two legislative districts.
BART: Local local elections were held in pateros on may 9, 2016 within the philippine general election. The voters elected for the elective local posts in the city: the mayor, vice mayor, and ten councilors.
FiD: Local elections was held in pateros on may 9, 2016 within the philippine general election. The voters elected for the elective local posts in the city: the mayor, vice mayor, and eight councilors.
ReGen: Local elections was held in pateros on may 9, 2016 within the philippine general election. The voters elected for the elective local posts in the city: the mayor, vice mayor, the two congressmen, and the councilors, six in each of the city's two legislative districts.

Query: Charles Lovell (trade unionist)
Ground-truth reference 1: The Guardian Charles Lovell, who has died aged 91, led the plumbers' trade union the ptu (later the eetpu) from the 1960s to the 1980s, a period of great turmoil and change in the world of industrial relations. Charles developed partnerships between his union and employer bodies representing the industries in which his members served. The outcome was enormously beneficial to both parties and to the industry. Son of Michael, a warehouseman, and Isabella, Charles was born in Newcastle upon tyne and began work in 1939. He was offered an apprenticeship as a plumber and joined the plumbing trade union, but as the second world war approached, he was drafted to swan hunter's shipyard under an essential works order. He was never happy there and returned to construction as soon as he was able. After the war, he also immersed himself in trade union activities and became chairman of the wallsend trades council and a labour councillor, as well as learning esperanto and providing for his beloved wife, Irene (nee lamb), whom he married in 1944, and their three daughters and two sons. In 1966 he was elected to the executive of the ptu and in 1968 became its general secretary. To make his union more secure, one of his first acts was to merge it with the electrical trade union to form the eetpu.
Ground-truth target: Charles Lovell (1923 – 13 october 2014) was a british trade unionist. Born in Newcastle-upon-Tyne, Lovell left school in 1939 and began an apprenticeship as a plumber, also joining the plumbing trades union. Soon after, he was drafted to work at the swan hunter shipyard, but returned to plumbing immediately after world war ii.[1] Lovell joined the labour party and was elected as a councillor, and also chaired wallsend trades council. He was elected to his union's executive council in 1966, and two years later won election as its general secretary.[1] He immediately led it into a merger with the electrical trades union, becoming national officer of the new electrical, electronic, telecommunications and plumbing union. In this role, he created a national pension scheme for plumbers, and championed more flexible apprenticeships, which could be completed under a variety of employers.[1] Lovell retired in 1987, and was made an officer of the order of the British empire. In his spare time, he learned Esperanto.[1]
BART: Charles Lovell was a British trade unionist. Born in london, Lovell worked as a coal miner, and became active in the miners' federation of great Britain. He was elected as the union's general secretary, and also served on the general council of the national union of mineworkers. Lovell also served as a member of the executive of the trades union congress, and was a delegate to the general conference of the international labour organization. He stood unsuccessfully for the labour party at the 1918 united kingdom general election, but was elected at the 1922 general election. In 1923, he was appointed as general secretary of the union, a post he held until his death.
FiD: Charles Lovell (1921 – 30 June 2011) was a British trade unionist. Born in Newcastle upon Tyne, Lovell began working as a plumber in 1939. He joined the plumbing trade union, and was elected to its executive in 1953. He was also elected to the wallsend trades council, and served as its chairman in the 1960s.[1] In 1966, Lovell was elected as the general secretary of the plumbers' trade union, and led the union through a period of turmoil and change in the world of industrial relations.[1] Lovell was also active in the labour party, serving as a councillor on wallsend borough council and as a member of the national executive committee of the trades union congress.[1] He retired from the union in 1986, and died in 2011, aged 91.[1]
ReGen: Charles Lovell (1928 – 15 July 2016) was a British trade unionist. Born in Newcastle upon Tyne, Lovell began work in 1939. He joined the plumbing trade union, but as the second world war approached, he was drafted to swan hunter's shipyard under an essential works order. He was never happy there and returned to construction as soon as he was able.[1] After the war, he also immersed himself in trade union activities and became chairman of the wallsend trades council and a labour councillor, as well as learning Esperanto and providing for his beloved wife, Irene (nee lamb), whom he married In 1944, and their three daughters and two sons. In 1966 he was elected to the general council of the trades union congress.[1] He led the ptu from the 1960s to the 1980s, a period of great turmoil and change in the world of industrial relations. Charles developed partnerships between his union and employer bodies representing the industries in which his members served.[1] The outcome was enormously beneficial to both parties and to the industry.[1]