Tackling Query-Focused Summarization as A Knowledge-Intensive Task: A Pilot Study

Weijia Zhang
w.zhang2@uva.nl
University of Amsterdam
Netherlands

Svitlana Vakulenko
svitlana.vakulenko@gmail.com
Amazon Alexa AI
Spain

Thilina Rajapakse
t.c.r.rajapakse@uva.nl
University of Amsterdam
Netherlands

Yumo Xu
yumo.xu@ed.ac.uk
University of Edinburgh
Edinburgh

Evangelos Kanoulas
E.Kanoulas@uva.nl
University of Amsterdam
Netherlands

ABSTRACT

Query-focused summarization (QFS) requires generating a summary given a query using a set of relevant documents. However, such relevant documents should be annotated manually and thus are not readily available in realistic scenarios. To address this limitation, we tackle the QFS task as a knowledge-intensive (KI) task without access to any relevant documents. Instead, we assume that these documents are present in a large-scale knowledge corpus and should be retrieved first. To explore this new setting, we build a new dataset (KI-QFS) by adapting existing QFS datasets. In this dataset, answering the query requires document retrieval from a knowledge corpus. We construct three different knowledge corpora, and we further provide relevance annotations to enable retrieval evaluation. Finally, we benchmark the dataset with state-of-the-art QFS models and retrieval-enhanced models. The experimental results demonstrate that QFS models perform significantly worse on KI-QFS compared to the original QFS task, indicating that the knowledge-intensive setting is much more challenging and offers substantial room for improvement. We believe that our investigation will inspire further research into addressing QFS in more realistic scenarios.

CCS CONCEPTS

• **Information systems** → **Summarization**; • **Computing methodologies** → **Natural language processing**.

KEYWORDS

query-focused summarization; retrieval-enhanced generation

1 INTRODUCTION

Query-focused summarization (QFS) [4] aims to produce a multi-sentence summary from a set of topic-related documents to answer a given query. This task has important applications in various domains such as search engines and report generation [28]. Although recent years have witnessed the advancement of QFS methods
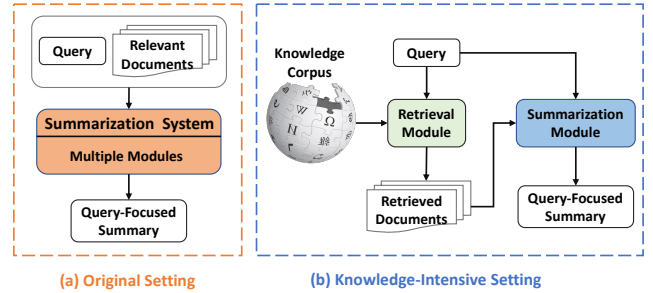
Figure 1: The Comparison between (a) the original QFS and (b) our knowledge-intensive setting. The original QFS assumes relevant documents are available. While our setting aims to generate the summary without relying on such relevant documents.

[21, 37, 38] thanks to large pre-trained language models [5, 22], most existing QFS datasets, such as widely-used DUC 2005-2007 datasets [4], are still relatively small in size and mainly used as evaluation-only sets, due to the labor-intensive nature of creating QFS datasets. For instance, in the DUC datasets, human annotators should first collect relevant documents and then carefully write a summary to answer a given query based on the documents. More importantly, the assumption that relevant documents are readily available is often impractical, which hinders the further development of QFS methods.

In this work, we tackle the QFS task as a knowledge-intensive (KI) task [32], where relevant documents should be retrieved from a large-scale knowledge corpus given a query. Figure 1 shows the difference between the original QFS and our knowledge-intensive setting. In the original setting, relevant documents are available as inputs. Note that the summarization system is often complex and consists of different modules. For instance, it may contain a relevance module that selects salient sentences and a generation module that produces the summary based on the sentences. While our setting addresses QFS in a more realistic scenario, where relevant documents are present in a knowledge corpus. In our setting, we employ an explicit retrieval module to retrieve more concise documents from a large-scale knowledge corpus and feed them into a summarization module to obtain the summary. In order to

explore this new knowledge-intensive setting, we have adapted existing QFS datasets to create a new dataset, KI-QFS. For dataset collection, we first take query-summary pairs from the original QFS datasets. Instead of only summarizing manually-collected relevant documents to answer the query, we explore generating such a summary based on a much larger knowledge corpus that better fits realistic scenarios. In this work, we consider three types of knowledge corpora. The first is an internal collection, which is formed by combining all documents in the QFS datasets. However, the size of this collection is relatively limited as the size of the QFS datasets is small, which makes the new dataset less realistic. We then use commonly-used Wikipedia as an external large-scale collection. As no explicit evidence is included in this external collection, we combine both collections to form an augmented collection. As retrieval is necessary for the new setting, we further provide relevance annotations to enable retrieval evaluation by recruiting crowdsourcing workers to annotate relevance labels.

To benchmark the proposed dataset KI-QFS, we consider two families of models: QFS models [37, 38] and retrieval-enhanced models [15, 23]. The performance of the former would indicate new challenges brought by the new setting compared with the original task. The latter have achieved state-of-the-art performance across many knowledge-intensive tasks such as open-domain question answering (QA) [2, 20] and fact verification [36]. Thus, their performance would indicate the efficacy of the existing methods in the new setting. The experimental results on KI-QFS demonstrate that the summarization performance of QFS models is much lower than their performance on the original QFS task, showing that the new setting is much more challenging than the original QFS task and offers substantial room for improvement. Besides, we show that the retrieval-enhanced model outperforms QFS methods, indicating their better efficacy in realistic scenarios where knowledge comes from large-scale collections.

The contributions of this work are summarized as follows: (i) To our knowledge, we make the very first attempts to tackle QFS as a knowledge-intensive task, where relevant documents should be retrieved from a large-scale knowledge corpus. (ii) We construct a corresponding dataset KI-QFS to study the new setting, and relevance annotations are provided to enable retrieval evaluation as well. (iii) We benchmark the dataset with a variety of models, and the experimental results indicate the new challenges brought by the knowledge-intensive setting.

2 RELATED WORK

Query-focused summarization. Due to the low resource setting, the previous work on QFS relied mainly on unsupervised approaches to extract salient sentences to produce the output summary [13, 24, 27]. With the increasing popularity of weak or distant supervision, recent studies [21, 37, 38] have explored the adaption of neural models trained on other domains where large-scale resources are available. For instance, Laskar et al. [21] incorporated an answer selection model trained on a large-scale QA dataset, MS MARCO [31]. Xu and Lapata [38] adopt a summarization model trained on a large-scale generic summarization dataset, CNN/DailyMail [14]. Another line of research aims to alleviate data scarcity by creating larger QFS datasets [19, 30]. For instance, Kulkarni et al. [19] build

QFS datasets without any manual annotation using unsupervised semantic matching techniques. However, the existing datasets either focus on short summaries (i.e. one sentence) or are adapted from the datasets in other domains (i.e., QA). In contrast to all previous work, our work does not rely on any large resources from other domains. Instead, we tackle the QFS task in a more realistic knowledge-intensive setting.

Knowledge-intensive tasks. Knowledge-intensive tasks, including open-domain QA [2, 20], dialogue [6], and fact-checking [36], involve answering a given question from a large-scale knowledge base [32]. The closest KI task is called long-form question answering (LFQA) [8, 35], where outputs are informative answers to given questions. However, existing LFQA tasks either suffer from poor answer grounding [18] or are restricted to ambiguous questions only [29]. Compared with them, our study has two main advantages: 1) expected answers are multi-document summaries, which are highly grounded in the documents; 2) queries reflect multiple information needs, and thus summaries are naturally longer in order to sufficiently cover various information.

Open-domain multi-document summarization. More recently, Giorgi et al. [11] extends multi-document summarization (MDS) into an open-domain setting. This work conducts a similar investigation on MDS datasets [7, 10]. Compared with it, our study has two significant differences: 1) it uses gold summaries as pseudo-queries as queries are unavailable in MDS datasets, while we use high-quality queries written by human experts from QFS datasets; 2) It only considers internal documents within individual dataset, which restricts their analysis on a specific domain (e.g., news). While we consider both internal documents and large-scale external knowledge corpus, which makes our analysis better fit realistic scenarios.

3 TASK FORMULATION

We briefly introduce the original QFS task and then describe our proposed knowledge-intensive setting.

Let the tuple $(q, \mathcal{D}, s)$ denote a sample in the QFS dataset (a *cluster*), where q is a given query, $\mathcal{D} = \{d_1, \ldots, d_M\}$ is a set of relevant documents, and s is the gold summary. The goal of QFS is to produce the summary given the documents and the query as input: $(q, \mathcal{D}) \rightarrow s$. Note that $\mathcal{D}$ is collected manually by human annotators and usually includes tens of documents.

In this work, we reformulate the QFS task to a knowledge-intensive task as $q \rightarrow s$. As shown in Figure 1, we do not rely on the relevant document set $\mathcal{D}$. Instead, we assume that a large-scale knowledge corpus can be accessed: $\mathcal{K} = \{k_1, \ldots k_N\}$. Normally, $\mathcal{K}$ consists of millions of documents, i.e., $N \gg M$. Note that our goal remains to generate the summary s with respect to the query q. However, different from the original setup, there is no evidence that the given documents are relevant to the query. More importantly, there are millions of candidate documents so it is impossible to consider all of them. Thus, an effective information retrieval model is necessary to supplement the extended setting, which is described in Section 5.1.

4 KI-QFS DATASET

In this section, we describe dataset collection and relevance annotation process. The dataset collection includes a collection of query-summary pairs and three knowledge corpora. The purpose of relevance annotation is to enable retrieval evaluation.

4.1 Dataset Collection

Collecting query-summary pairs. We build the dataset on the top of query-summary pairs (q, s) on existing QFS resources. Specifically, we adopt DUC 2005-07 [4], three standard QFS benchmark datasets. [1] The DUC datasets consist of three subsets collected for the Document Understanding Conferences (DUC) from 2005 to 2007. Each subset contains 45-50 clusters, where each cluster contains a query, a set of topic-related documents, and multiple reference summaries. As we assume that relevant documents are inaccessible, we only collect query-summary pairs. The documents will be used to build a knowledge corpus, which is described in the next paragraph. For data split, we follow previous work [21, 25] to use the pairs of the first two years as the training set and randomly select 10% from the training set as the validation set. We leave the subset of 2007 as the test set. Finally, the training set, validation set, and test set contain 90, 10, and 45 examples, respectively.

Building knowledge corpora. We explore three alternatives of knowledge corpora for the query-summary pairs above. Specifically, we follow the standard data processing for large-scale knowledge sources [16] to collect documents, where we split each origin document into non-overlapping context documents with 100 words maximum. [2] The first knowledge corpus, $\text{Corpus}_{\text{int}}$, is an internal collection, where we take all clusters of DUC documents from the three years to form this collection, which results in about 32K documents in total. As the queries have an explicit connection with the documents, this collection can be considered an in-domain knowledge corpus, where relevant documents are relatively easy to find. However, as our main goal is to explore knowledge-intensive QFS on a large-scale knowledge base, we consider the second external knowledge corpus named $\text{Corpus}_{\text{ext}}$. We use the Wikipedia dump in the KILT benchmark [32] to form this corpus, resulting in about 21 million documents in total. However, as $\text{Corpus}_{\text{ext}}$ is a Wiki-based corpus, there is no guarantee that the collection contains sufficient evidence to answer the query since the reference summary is derived from the content of the original DUC documents. To this end, we build an augmented knowledge corpus called $\text{Corpus}_{\text{aug}}$ by combining the previous two collections. We merge all the documents of $\text{Corpus}_{\text{int}}$ and $\text{Corpus}_{\text{ext}}$ into this collection ensuring that summaries can be grounded in the collection and that the collection is large-scale to make the task challenging.

4.2 Relevance Annotation

As the retrieval is performed for our knowledge-intensive setting, it is important to evaluate the performance so we can better analyze its effect on summarization. However, the DUC datasets do

---

[1]Note that we take DUC as an example, but nothing prohibits exploring other QFS resources.
[2]Some studies also use passages to name the processed text, but we keep using documents to maintain the coherence of the paper.

Table 1: Comparisons of retrieval models on the test set of KI-QFS with three knowledge corpora. * means that the improvement is statistically significant (a two-paired t-test with p-value < 0.01).

Corpus	Model	P@10	P@50	R@10	R@50
$\text{Corpus}_{\text{int}}$	BM25	**16.00***	**11.96**	**8.24***	**30.19***
	DPR	11.56	11.51	5.97	27.80
$\text{Corpus}_{\text{ext}}$	BM25	12.67	8.58	**7.00**	22.22
	DPR	**13.78***	**12.58***	6.95	**29.69***
$\text{Corpus}_{\text{aug}}$	BM25	12.22	8.98	6.84	23.21
	DPR	**14.44***	**12.53***	**7.41***	**30.32***

not come with relevance labels to designate which document provides evidence towards the final summary. A possible solution to that would be to do a lexical match between the individual summary sentences and the available document collection, however, such an automatic evaluation has its disadvantages (e.g. a summary sentence may appear in a document but outside the right context, lexical overlap may pay attention to insignificant words in the summary, and there may be a lexical gap between semantically similar sentences). Therefore, we collect human annotations through Amazon Mechanical Turk (MTurk) to create our evaluation set. Following the TREC document retrieval task [3], we apply retrieval models to choose top-ranked documents that are used fro annotation. The MTurk workers are asked to label a relevance score for each query-document pair using a 4-point scale. More details on the annotation process are provided in Appendix A.

5 EXPERIMENTS

5.1 Models

QFS models. To illustrate new challenges brought by the proposed knowledge-intensive setting, we first evaluate the models designed for original QFS tasks: (i) QUERYSUM [37] is an extractive QFS model leveraging distant supervision signals from trained QA models to select salient sentences as the summary. (ii) MARGESUM [38] is an abstractive QFS model that employs generative models trained on generic summarization resources to boost sentence ranking and summary generation, achieving competitive performance in the weakly supervised setting. Implementation details about QFS models are provided in Appendix B.1

Table 2: Comparisons of QFS models on the original document sets (ORIGIN) and our knowledge corpus ($\text{Corpus}_{\text{int}}$). The results inside the bracket denote the percentage decrease ($\downarrow$) when the models are adapted to the knowledge-intensive setting.

Corpus	Model	ROUGE-1	ROUGE-2	ROUGE-SU4
ORIGIN	QUERYSUM	43.25	11.58	16.76
	MARGESUM	42.52	11.77	16.73
$\text{Corpus}_{\text{int}}$	QUERYSUM	36.08 (17%)	7.54 (35%)	12.69 (24%)
	MARGESUM	37.99 (11%)	9.13 (22%)	14.29 (15%)

Table 3: Comparison of summarization performance on the three knowledge corpora. R-1, R-2 and R-SU4 stand for ROUGE-1, ROUGE-2 and ROUGE-SU4, repectively. Note that the results of BARTScore are normalized with the exponential function. * indicates the improvement to the best QFS model is statistically significant (a two-paired t-test with p-value < 0.01).

Model	$\text{Corpus}_{\text{int}}$					$\text{Corpus}_{\text{ext}}$					$\text{Corpus}_{\text{aug}}$				
	R-1	R-2	R-SU4	$\text{BERT}_{\text{score}}$	$\text{BART}_{\text{score}}$	R-1	R-2	R-SU4	$\text{BERT}_{\text{score}}$	$\text{BART}_{\text{score}}$	R-1	R-2	R-SU4	$\text{BERT}_{\text{score}}$	$\text{BART}_{\text{score}}$
Oracle+FiD	44.14	12.63	17.50	24.35	3.77	38.79	8.38	14.27	15.77	3.25	44.08	12.58	17.42	24.98	3.77
QuerySum	36.08	7.54	12.69	8.53	3.08	31.07	4.52	10.19	2.03	2.72	32.64	5.47	11.09	2.96	2.84
MargeSum	37.99	9.13	14.29	11.49	3.09	34.37	6.52	12.20	5.86	2.84	36.69	8.07	13.52	8.69	3.01
RAG	28.92	5.72	10.07	12.60	3.19	32.28	5.21	10.76	7.99	3.03	27.06	4.57	9.00	8.28	3.03
BM25+FiD	42.44*	11.25*	16.54*	21.42*	3.61	38.80*	8.36*	14.22*	15.69*	3.25	41.40*	10.83*	16.07*	19.99*	3.47
DPR+FiD	41.46	10.72	15.91	21.41	3.60	38.56	8.01	14.10	15.50	3.22	39.99	9.42	15.06	18.04	3.38

Retrieval-enhanced models. We also adopt retrieval-enhanced generation models in the new setting: (i) Retrieval-Augmented Generation (RAG) [23] is a generative model in which document retrieval and summary generation are learned jointly. As RAG models have two variants including RAG-Token and RAG-Sequence, we mainly report the results of RAG-Sequence as it performs better in open-domain QA. (ii) Fusion-in-Decoder (FiD) [15] is an improved encoder-decoder architecture which has achieved state-of-art performance in some knowledge-intensive tasks [1]. In this model, encoded representations for top-k retrieved documents are first concatenated and then fed into the decoder to generate the output answer. We explore the effect of two commonly-used retrieval models: BM25 [34] and Dense Passage Retrieval (DPR) [16]. More implementation details about the models are provided in Appendix B.2

5.2 Evaluation Metrics

As our knowledge-intensive setting includes a retrieval module and a summarization module, we evaluate both modules using different metrics. For retrieval evaluation, we use precision@k (P@k) and recall@k (R@k) where k denotes the cut-off. For summarization evaluation, we use standard ROUGE-1, ROUGE-2, and ROUGE-SU4 [26]. We also report semantic metrics including BERTScore [40] and BARTScore [39]. More details on the metrics are given in Appendix C.

5.3 Results and Analysis

Our experiments mainly are designed to answer the following research questions (RQs): (i) **RQ1:** What is the performance of retrieval models in the knowledge-intensive setting? (ii) **RQ2:** Is this setting much more challenging than the original QFS? (iii) **RQ3:** How do QFS models and retrieved-enhanced models perform in the KI-QFS dataset?

5.3.1 Retrieval Results. To answer **RQ1**, we compare the retrieval performance of BM25 and DPR on the test set of KI-QFS (see Appendix A for details of relevance annotation). As shown in Table 1, BM25 outperforms DPR on $\text{Corpus}_{\text{int}}$ than $\text{Corpus}_{\text{aug}}$. One possible explanation is that relevant documents in internal knowledge corpus $\text{Corpus}_{\text{int}}$ contain more relevant keywords. These keywords can be easier retrieved by BM25 which is a term-based IR method. While DPR performs better on $\text{Corpus}_{\text{ext}}$, Since it is pre-trained on the dump of Wikipedia so semantic representations of documents

are learned better, which leads to retrieval improvement. The fairly low precision and recall scores across all three collections show that knowledge retrieval is challenging, and more effort is required to boot the performance.

5.3.2 Summarization Results. To answer **RQ2**, we compare the start-of-the-art QFS models (distantly supervised) between original document sets (Origin) and our $\text{Corpus}_{\text{int}}$. Note that the only difference between them is that Origin only contains a small set of relevant documents while $\text{Corpus}_{\text{int}}$ consists of a large number of candidate documents in the datasets. As shown in Table 2, we find that the performance of QFS models drops largely when they are performed in the knowledge-intensive setting. This is because QFS systems are developed for closed-domain retrieval and suffer from the distributional shift when they perform knowledge retrieval in an open-domain setting. This shows the proposed dataset is challenging, and more research effort is needed to adapt close-domain QFS models to our setting.

To answer **RQ3**, we benchmark our dataset with QFS models and retrieval-enhanced models. As shown in Table 3, We first report the results of an Oracle+FiD model which serves as an upper-bound performance. To obtain Oracle documents, we first compute ROUGE scores between all documents in the corresponding knowledge corpus and gold summary, and then we select top-k documents as Oracle documents based on their scores. We observe a significant increase in the performance of the Oracle+FID especially on the $\text{Corpus}_{\text{aug}}$. This shows that the retrieval performance in large-scale collections really affects the summarization performance. We find that FiD-based models outperform QFS models, showing benefits from fine-tuning on the datasets. We notice that the performance of the RAG-based model shows inferior performance, potentially due to the small size of the training data.

6 CONCLUSION

In this paper, we address the QFS task in a knowledge-intensive setting, where relevant documents are not readily available and should be retrieved from a large-scale knowledge corpus. To this end, we build a new dataset based on the existing DUC datasets. In this dataset, we construct three knowledge corpora to enable document retrieval. To evaluate retrieval performance, we further provide relevance annotations. We benchmark the dataset with QFS models and retrieval-enhanced models, the experimental results

show that the knowledge-intensive setting is much more challenging than the original setting. In future work, we consider generating relevance labels using some domain adaption methods, in which human annotation may not be required.

REFERENCES

[1] Akari Asai, Matt Gardner, and Hannaneh Hajishirzi. 2022. Evidentiality-Guided Generation for Knowledge-Intensive NLP Tasks. 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*. Association for Computational Linguistics, 2226–2243. https://doi.org/10.18653/v1/2022.naacl-main.162

[2] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to Answer Open-Domain Questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1870–1879.

[3] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2020. Overview of the TREC 2020 Deep Learning Track. In *Proceedings of the Twenty-Ninth Text REtrieval Conference, TREC 2020, Virtual Event [Gaithersburg, Maryland, USA], November 16-20, 2020 (NIST Special Publication, Vol. 1266)*, Ellen M. Voorhees and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST).

[4] Hoa Trang Dang. 2006. DUC 2005: Evaluation of question-focused summarization systems. In *Proceedings of the Workshop on Task-Focused Summarization and Question Answering*. 48–55.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/n19-1423

[6] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. 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.

[7] Alexander R. Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R. Radev. 2019. Multi-News: A Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 1074–1084. https://doi.org/10.18653/v1/p19-1102

[8] Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long Form Question Answering. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Anna Korhonen, David R Traum, and Llu\'{i}s Màrquez (Eds.). Association for Computational Linguistics, 3558–3567. https://doi.org/10.18653/v1/p19-1346

[9] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378.

[10] Demian Gholipour Ghalandari, Chris Hokamp, Nghia The Pham, John Glover, and Georgiana Ifrim. 2020. A Large-Scale Multi-Document Summarization Dataset from the Wikipedia Current Events Portal. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 1302–1308. https://doi.org/10.18653/v1/2020.acl-main.120

[11] John M. Giorgi, Luca Soldaini, Bo Wang, Gary D. Bader, Kyle Lo, Lucy Lu Wang, and Arman Cohan. 2022. Exploring the Challenges of Open Domain Multi-Document Summarization. abs/2212.10526 (2022). https://doi.org/10.48550/arXiv.2212.10526 arXiv:2212.10526

[12] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. [n. d.]. Deberta: Decoding-Enhanced Bert with Disentangled Attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021* (2021). OpenReview.net.

[13] Zhanying He, Chun Chen, Jiajun Bu, Can Wang, Lijun Zhang, Deng Cai, and Xiaofei He. 2012. Document Summarization Based on Data Reconstruction. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, July 22-26, 2012, Toronto, Ontario, Canada*, Jörg Hoffmann and Bart Selman (Eds.). AAAI Press. http://www.aaai.org/ocs/index.php/AAAI/AAAI12/paper/view/4991

[14] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems* 28 (2015), 1693–1701.

[15] Gautier Izacard and Edouard Grave. 2021. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In *Proceedings of the*

[16] 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*. 6769–6781.

[17] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1412.6980

[18] Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to Progress in Long-form Question Answering. In *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*, Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (Eds.). Association for Computational Linguistics, 4940–4957. https://doi.org/10.18653/v1/2021.naacl-main.393

[19] Sayali Kulkarni, Sheide Chammas, Wan Zhu, Fei Sha, and Eugene Ie. 2020. Aquamuse: Automatically generating datasets for query-based multi-document summarization. *arXiv preprint arXiv:2010.12694* (2020).

[20] 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 (2019), 453–466.

[21] Md. Tahmid Rahman Laskar, Enamul Hoque, and Jimmy Xiangji Huang. 2020. WSL-DS: Weakly Supervised Learning with Distant Supervision for Query Focused Multi-Document Abstractive Summarization. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, Donia Scott, Núria Bel, and Chengqing Zong (Eds.). International Committee on Computational Linguistics, 5647–5654. https://doi.org/10.18653/v1/2020.coling-main.495

[22] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-Training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703

[23] 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. 2020. 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*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.).

[24] Piji Li, Zihao Wang, Wai Lam, Zhaochun Ren, and Lidong Bing. 2017. Salience estimation via variational auto-encoders for multi-document summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31.

[25] Weikang Li, Xingxing Zhang, Yunfang Wu, Furu Wei, and Ming Zhou. 2019. Document-Based Question Answering Improves Query-Focused Multi-document Summarization. In *CCF International Conference on Natural Language Processing and Chinese Computing*. Springer, 41–52.

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

[27] He Liu, Hongliang Yu, and Zhi-Hong Deng. 2015. Multi-document summarization based on two-level sparse representation model. In *Twenty-ninth AAAI conference on artificial intelligence*.

[28] Congbo Ma, Wei Emma Zhang, Mingyu Guo, Hu Wang, and Quan Z Sheng. 2023. Multi-Document Summarization via Deep Learning Techniques: A Survey. 55, 5 (2023), 102:1–102:37. https://doi.org/10.1145/3529754

[29] Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering Ambiguous Open-domain Questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 5783–5797. https://doi.org/10.18653/v1/2020.emnlp-main.466

[30] Preksha Nema, Mitesh M. Khapra, Anirban Laha, and Balaraman Ravindran. 2017. Diversity Driven Attention Model for Query-Based Abstractive Summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, Regina Barzilay and Min-Yen Kan (Eds.). Association for Computational Linguistics, 1063–1072. https://doi.org/10.18653/v1/P17-1098

[31] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine

Reading COmprehension Dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches 2016 Co-Located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016 (CEUR Workshop Proceedings, Vol. 1773)*, Tarek Richard Besold and etc (Eds.). CEUR-WS.org. https://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf

[32] 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. 2021. KILT: A Benchmark for Knowledge Intensive Language Tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (Eds.). Association for Computational Linguistics, 2523–2544. https://doi.org/10.18653/v1/2021.naacl-main.200

[33] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. 21 (2020), 140:1–140:67. http://jmlr.org/papers/v21/20-074.html

[34] Stephen E. Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. 3, 4 (2009), 333–389. https://doi.org/10.1561/1500000019

[35] Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. ASQA: Factoid Questions Meet Long-Form Answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, 8273–8288. https://aclanthology.org/2022.emnlp-main.566

[36] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: A Large-Scale Dataset for Fact Extraction and VERification. In *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)*, Marilyn A. Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, 809–819. https://doi.org/10.18653/v1/n18-1074

[37] Yumo Xu and Mirella Lapata. 2020. Coarse-to-Fine Query Focused Multi-Document Summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 3632–3645. https://doi.org/10.18653/v1/2020.emnlp-main.296

[38] Yumo Xu and Mirella Lapata. 2021. Generating Query Focused Summaries from Query-Free Resources. In *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*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 6096–6109. https://doi.org/10.18653/v1/2021.acl-long.475

[39] Weizhe Yuan, Graham Neubig, and Pengfei Liu. [n. d.]. BARTScore: Evaluating Generated Text as Text Generation. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, Virtual* (2021), Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N Dauphin, Percy Liang, and Jennifer Wortman Vaughan (Eds.). 27263–27277.

[40] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

A DETAILS ON ANNOTATION PROCESS

As the main goal of the annotation process is to identify relevant documents that support the summary, we begin with filtering out irrelevant documents, which are the majority of millions of candidate documents in the knowledge corpus. Following the TREC document retrieval task [3], we apply retrieval models to choose top-ranked documents. In particular, we consider two different but effective retrieval approaches, BM25 [34] and Dense Passage Retrieval (DPR) [16]. We first perform both BM25 and DPR on Corpus$_{int}$, Corpus$_{ext}$ and Corpus$_{aug}$, and construct top-50 pools, which result in a maximum of 300 candidate documents for each query. Not that duplicate documents may exist in the pool and would be removed.

As there are 45 queries in our test set, we end up with 7761 query-document pairs to be annotated. The MTurk workers are asked to label a relevance score for each pair using a 4-point scale. We design a web annotation interface shown to MTurk workers, which contains a query, a reference summary, and at most five documents. We ask workers to label whether a document contains either part of the summary or evidence to a query, adapting the annotation instructions in Craswell et al. [3]. The judgments are on a 4-point scale from 0 to 3:

- **Irrelevant (0)**: the document has nothing to do with the query.
- **Weakly related (1)**: the document seems related to the query but fails to contain any evidence in the summary.
- **Related (2)**: the document provides unclear information related to the query, but human inference may be needed.
- **Relevant (3)**: the document explicitly contains evidence that is part of the summary.

Prior to the collection of the annotations we performed two pilot studies to improve our annotation interface and the instructions. We collect three annotations per document and use a majority vote to define the label. The distribution of relevance labels is 837 relevant, 5811 related, 1097 weakly related, and 16 irrelevant documents. We also measure inter-annotator agreement (IAA) using Fleiss Kappa [9]. A Fleiss Kappa score of 0.245 was obtained, which indicates a fair agreement.

From the annotation results, we find that although most retrieved documents are related to the queries to some extent, there are only 837 documents (10.8%) that contained explicit fragments of the summary. To validate whether the documents with relevance labels equal to 3 are sufficient to generate the summaries, we compared the summarization performance when documents annotated as perfectly relevant were provided as input and the performance when documents annotated either as perfectly relevant or relevant, with the latter offering no extra performance gains over the former.

B IMPLEMENTATION DETAILS

B.1 QFS Models

As QFS models have achieved competitive performance in the original DUC datasets, we directly use them to generate summaries without any further training. For inference, as they include a retrieval module that performs in-domain evidence estimation, we can directly implement the models on the collection Corpus$_{int}$. However, their retrieval models are not designed to handle large-scale, open-domain knowledge bases, such as Corpus$_{ext}$. Therefore, we first retrieve top-1000 documents using BM25 for each query and then feed them as inputs to their customized retrieval models. We follow the original papers to set the hyper-parameters. All models are conducted on a single Titan-XP GPU.

B.2 Retrieval-Enhanced Models

As retrieval-enhanced generation models are designed for open-domain QA where answers are much shorter. For fair comparisons, we fine-tune them in our training set. To be specific, there are 4 gold summaries in the datasets. we randomly choose a summary to construct the input-output pairs to construct each training batch with a batch size of 1. For the hyper-parameters of FiD, we set k

for top-k retrieved documents to 50 after a small validation study. Following the origin paper, we use T5$_{base}$ [33] as the base model. For DPR, we use the version trained on the KILT benchmark [32]. As for RAG, we follow the original paper to set the number of retrieved documents to 10 [23]. For fine-tuning both models, we use Adam [17] with a learning rate of 10^{-4} and the maximum number of training epochs to 30. We also apply early-stopping strategies to avoid over-fitting when we find there is no improvement in the validation set within 10 training steps. For inference, we set the maximum decoded length and beam size to 250 and 5, respectively. All experiments are conducted on a single Titan-XP GPU.

C   DETAILS ON EVALUATION METRICS

C.1   Retrieval Evaluation

To evaluate retrieval effectiveness, i.e., the ability of the QA systems to identify and retrieve evidence to be used to compose the final answer, we use standard information retrieval (IR) metrics, and in particular precision@k (P@k) and recall@k (R@k) where k denotes the cut-off. For P@k and R@k, we mapped the judgment level 3 to relevant and judgment levels 0-2 to irrelevant and reported the corresponding results for $k \in \{10, 50\}$.

C.2   Summarization Evaluation

In our experiments, we measure the accuracy of generated summaries against gold summaries. Following previous work in the QFS task [21, 37], we employ the standard ROUGE-1, ROUGE-2, and ROUGE-SU4 [26]. They evaluate the accuracy in terms of unigrams, bi-grams, and bi-grams with a maximum skip distance of 4, respectively. We report the F_1 score with a maximum length of 250 words for the metrics. Note that there are multiple reference summaries in DUC datasets, we take the average of the ROUGE scores for all summaries as the final reported ROUGE score.

Although ROUGE-based metrics are commonly used, they fail to capture semantic similarities since they only measure lexical overlap. Recent semantic metrics, such as BERTScore [40] and BARTScore [39] have been shown to be strongly correlated with human judgments. Therefore, we also report both metrics in our experiments. Specifically, we report the F1 score of BERTscore and use recommended `deberta-xlarge-mnli` [12] as the backbone of BERTScore. For BARTScore, we use the recall score of BARTScore as it is more suitable for summarization tasks [39], where we use `bart-large-cnn` as the base model. As the outputs of BARTScore are negative log probabilities that are difficult to explain, we normalize them with the exponential function.