BioGen: Generating Biography Summary under Table Guidance on Wikipedia

Shen Gao [♠,*], **Xiuying Chen** [♠♣,*], **Chang Liu** [♠♣], **Dongyan Zhao** [♠♣♡,†], **Rui Yan** [♢]
[♠]Wangxuan Institute of Computer Technology, Peking University, Beijing, China
[♣]Center for Data Science, Peking University, Beijing, China
[♢]Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
[♡]State Key Laboratory of Media Convergence Production Technology and Systems
{shengao, xy-chen, liuchang97, zhaody}@pku.edu.cn
ruiyan@ruc.edu.cn

Abstract

Capturing the salient information from an input article has been a long-standing challenge for summarization. On Wikipedia, most of the wiki pages about people contain a factual table that lists the basic properties of the people. Illuminatingly, a factual table can be regarded as a natural summary of the key information in the corresponding article. Thus, in this paper we propose the task of table-guided abstractive biography summarization, which utilizes factual tables to capture important information and then generate a summary of a biography. We first introduce the TaGS (Table-Guided Summarization) dataset[1], the first large-scale biography summarization dataset with tables. Next, we report some statistics about this dataset to validate the quality of the dataset. We also benchmark several commonly used summarization methods on TaGS and hope this will inspire more exciting methods.

1 Introduction

Text summarization generates a short text version of a long passage which retains the most important information. Recently, two kinds of approaches have been proposed for automatic text summarization. One is extractive summarization (Nallapati et al., 2017; Liu and Lapata, 2019), which directly selects salient sentences from the passage to create a summary. The other is abstractive summarization (See et al., 2017; Hsu et al., 2018a), which aims to concisely paraphrase the input article. In both methods, the summary should always focus on important information, though a document may include trivial facts.

To focus on the main information when generating summaries, some researchers propose to incorporate manifold information to improve the performance. Narayan et al. (2017) proposed to incorporate the figures and Gao et al. (2019b) investigated the using of reader comments for more effective summarization. As another type of side information, factual tables provide a natural summary of the biography document. On Wikipedia, in each wiki page about people, there is a factual table (infobox) on the right side of the page summarizing the main properties. Clearly, infobox is helpful for capturing the salient information during summarizing the biography. However, no existing work takes advantage of tables, though are widely available in the biography on Wikipedia.

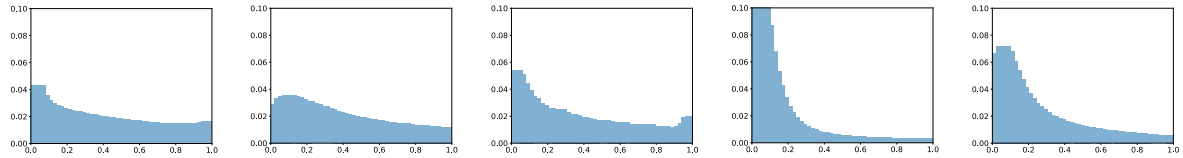
In this paper, we propose *Table-Guided Summarization* (TaGS) dataset, the first large-scale biography summarization dataset with tables. And we report some statistics and three important characteristics of this dataset to verify its quality. The first one is it has the weak lead bias that makes it suitable for training both abstractive and extractive summarization methods. The second one is it has strong abstractness that is helpful for generating a more condensed summary. The most important characteristic is that the summary of the biography is guided by a table which contains the most salient facts described in the biography.

To verify the quality of this dataset, we employ some commonly used state-of-the-art summarization methods to conduct experiments on our proposed dataset. From these experimental results, we can see that the methods which simply incorporate the table information outperform the methods which do not use the table information. That demonstrates the effectiveness of incorporating table guidance when generating summaries of documents which have a factual table in it.

Our contributions are summarized as follows:

[*]Equal contribution. Ordering is decided by a coin flip.
[†]Corresponding Author: Dongyan Zhao
[1]https://github.com/gsh199449/table-summ

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 4752–4757



(a) TaGS (b) XSUM (c) Reddit (d) Newsroom (e) CNN/DailyMail

Figure 1: Relative locations of bigrams of ground truth summaries in the source text across different datasets.

• To the best of our knowledge, we are the first to use factual tables to guide the summarization procedure so as to generate better summaries.

• We release a large-scale abstractive biography summarization dataset with tables. Experiments conducted on this dataset demonstrate the effectiveness of incorporating table information in generating summaries.

2 Related Work

2.1 Text Summarization

Text summarization is an important task which can be classified into extractive and abstractive approaches. Extractive summarization (Narayan et al., 2018b; Chen et al., 2018; Jadhav and Rajan, 2018) tends to generate a summary by integrating the most salient sentences in the document. Cheng and Lapata (2016) first propose using recurrent neural network (RNN) to extract salient sentences. After that researchers explore many the neural based method (Nallapati et al., 2017; Liu and Lapata, 2019; Chen et al., 2018; Zhang et al., 2018; Zhou et al., 2018; Liu et al., 2019), and achieve the state-of-the-art performance (Liu and Lapata, 2019) on the benchmark dataset CNN/DailyMail. In the mean time, the Nallapati et al. (2016) firstly apply this text generation method to the abstractive summarization task and Gehrmann et al. (2018) achieve the state-of-the-art performance by using a data-efficient content selector.

2.2 Summarization with Side Information

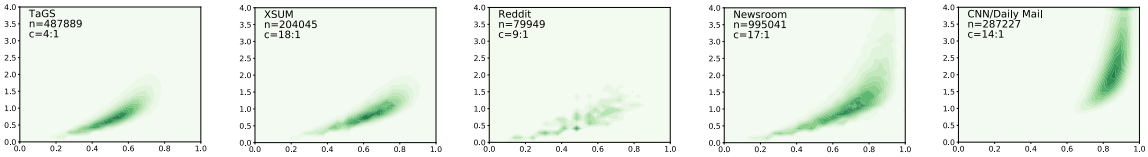
Traditional text summarization methods only use the document as input. However, the gist of the document may lie in side information, such as the title, image captions, or comments which are often available for news-wire articles. As such, various studies (Gao et al., 2020; Hu et al., 2008) have tried to use such information for more efficient and accurate summarization. However, to the best of our knowledge, no existing works consider the use of tables to guide biography summarization.

3 TaGS Dataset

Our dataset, named Table-Guide Summarization (TaGS), consists of over 500,000 document-summary pairs, along with their corresponding factual tables collected from Wikipedia. Concretely, following Chen et al. (2019), we use the leading paragraphs before the content outline as the summary, and following paragraphs as the document. The infobox in the right part of the webpage is extracted as the guided table.

Some key statistics of the factual table are described below. 7.31% of words from a document and 29.41% of words from a summary are included in a factual table. The average number of fields in a table is 12.89, and there are 46.83 words in a table in average. We show some detailed statistics of document-summary pairs in TaGS and compare them with other popular text summarization datasets in Table 1. We next discuss some abstractive characteristics of TaGS compared to existing summarization datasets.

Weak Lead Bias. Lead bias means that directly using the leading sentences of a document can produce a good performance in terms of the summarization evaluation metric ROUGE (Lin, 2004). This is a common problem in text summarization datasets, which mostly occurs in news-based documents. Figure 1 plots the density histograms for the relative locations of words from the ground truth summary in the input document. In the CNN/DailyMail and Newsroom datasets, the words are highly concentrated at the leading parts of the input document. In contrast, our TaGS dataset shows more uniform distributions across words in the document. This characteristic can be also found from the `LEAD` score shown in Table 2, where `LEAD` is a baseline method that selects the first few sentences in the input document as the summary. A high `LEAD` score implicitly indicates a strong lead bias. From Table 2, we find that TaGS has a much lower `LEAD` score than the CNN/DailyMail and Newsroom datasets, and



(a) TaGS (b) XSUM (c) Reddit (d) Newsroom (e) CNN/DailyMail

Figure 2: Density estimation of extractive diversity scores. Large variability along the y-axis suggests variation in the average length of source sequences present in the summary, while x-axis shows variability in the average length of extractive fragments to which summary words belong.

Table 1: Comparison of summarization datasets with respect to corpus size, average document (source) and summary (target) length (in terms of words and sentences) on both source and target.

Datasets	# docs (train/val/test)	avg. document length		avg. summary length	
		words	sentences	words	sentences
CNN/DailyMail (Hermann et al., 2015)	287,227/13,368/11,490	810.57	39.78	56.20	3.68
XSum (Narayan et al., 2018a)	204,045/11,332/11,334	431.07	19.77	23.26	1.00
Reddit (Kim et al., 2019)	79,949	342.00	18.02	9.33	1.03
Newsroom (Grusky et al., 2018)	995,041/108,837/108,862	658.60	31.43	26.70	1.42
TaGS	487,889/5,000/5,000	226.13	7.81	49.25	1.88

thus prevents the model from directly learning the salient information by locational bias.

Strong Abstractness. Table 2 reports the percentage of novel n-grams in the ground truth summary that do not appear in the input document. The result shows that our dataset comprises of 43.38% novel unigrams in the summary, 122.46% higher than the commonly-used benchmark dataset CNN/DailyMail. This indicates that summaries in TaGS are more abstractive. Besides, other two metrics, density and coverage, proposed by Grusky et al. (2018), are commonly used when evaluating the summarization dataset (Kim et al., 2019; Grusky et al., 2018). We plot the distributions of these two metrics in Figure 2, where small density and coverage reflects the summary has strong abstractness. The result shows that TaGS is similar to XSUM and Reddit in terms of density and coverage, and these datasets all have strong abstractness.

PG/LEAD in Table 2 is the ROUGE-L ratios of PG to LEAD, which quantifies the difficulty for extractive methods and the suitability for abstractive methods. CNN/DailyMail and Newsroom achieve low PG/LEAD scores, demonstrating that these datasets are more suitable for extractive based model. On the contrary, XSUM and TaGS have high PG/LEAD, showing that TaGS is potentially an excellent benchmark for evaluation of abstractive summarization systems.

Table Guided. The intrinsic difference between TaGS and other datasets lies in that each document-summary pair in TaGS is associated with a factual table. We judge whether the table is good guidance from two aspects: (1) word-level overlap percentage and (2) sentence-level mapping relation between document and table. From the word-level aspect, 7.31% words in the document are included in the table; from the sentence-level aspect, we found that 99.86% of document sentences have words in common with the factual table. This demonstrates that the table captures the majority of facts in the document. Thus, it is safe to conclude that the tables in TaGS can be utilized as a good guidance.

4 Experimental Setup

4.1 Comparison Methods

To evaluate the effectiveness of incorporating table, we conduct experiments using the following baselines: (1) **LEAD3**: selects the first three sentences of a document as the summary. (2) **S2S**: is the traditional sequence-to-sequence framework in (Sutskever et al., 2014) which has been used in many text generation tasks (Gao et al., 2019c, 2021; Chan et al., 2019a, 2020, 2019b). (3) **PG**: combines S2S with copy mechanism in See et al. (2017). (4) **Unified**: proposed by Hsu et al. (2018b), combines the strength of extractive and abstractive summarization. (5) **Transformer**: is solely based on attention mechanism

Table 2: Corpus bias towards extractive methods in datasets. We show the proportion of novel n-grams in gold summaries. We also report ROUGE scores for the LEAD baseline and the abstractive summarization method PG. Results are computed on the test set.

Datasets	% of novel n-grams in gold summary				LEAD			PG			PG/LEAD
	unigram	bigrams	trigrams	4-grams	R1	R2	RL	R1	R2	RL	Ratio (R-L)
CNN/DM	19.50	56.88	74.41	82.83	39.60	17.70	36.20	36.4	15.7	33.4	0.92x
XSum	35.76	83.45	95.50	98.49	16.30	1.61	11.95	29.7	9.2	23.2	1.93x
Reddit	48.97	84.42	94.66	97.82	3.40	0.00	3.30	19.0	3.7	15.1	5.59x
Newsroom	19.53	48.39	59.38	64.06	30.50	21.30	28.40	14.7	2.2	10.3	0.92x
TaGS	43.38	84.30	94.09	97.11	23.10	4.80	19.57	29.3	11.2	27.9	1.43x

proposed in Vaswani et al. (2017). (6) **CopyTransformer**: is a state-of-the-art generative summarization model (Gehrmann et al., 2018), which combines the `Transformer` with copy mechanism. (7) **TabWords**: just concatenates all the words in table as the summary.

Additionally, we select two best baselines, `Unified` and `CopyTransformer`, concatenating the original input document with tables as input, denoted as (9) **Unified+T** and (10) **CopyTransformer+T**, to determine whether the improvement of TGSG simply arises from the table information.

4.2 Evaluation Metrics

For evaluation metrics, we adopt the ROUGE scores (Lin, 2004) which is widely applied for summarization evaluation (Sun et al., 2018; Chen et al., 2018; Gao et al., 2019a). The ROUGE metrics compare the generated summary with the reference summary by computing overlapping lexical units, and include ROUGE-1, ROUGE-2 and ROUGE-L.

5 Experimental Results

We first examine the performance of these baselines, as shown in Table 3. Firstly, among models without table information, `Unified` achieves the highest performance. `PG` and `CopyTransformer` achieve the second best performance. Secondly, tables are indeed helpful for the summarization process. For models with additional table information, the ROUGE-1 score of `Unified+T` and `CopyTransformer+T` improves by 4.22 and 14.3, respectively. This observation demonstrates that factual tables can help the summarization model to capture the main idea of the table by emphasizing the key facts in the document. However, the performance of `TabWords` is much lower than `Unified+T` and `CopyTransformer+T`, which demonstrates only using the information from table is not sufficient for generating a good summary.

Table 3: ROUGE scores comparison between baselines and TGSG. All our ROUGE scores have a 95% confidence interval of at most ±0.23 as reported by the official ROUGE script.

	ROUGE-1	ROUGE-2	ROUGE-L
w/o table information			
LEAD3	23.10	4.80	19.57
S2S	26.31	10.07	25.13
PG	29.33	11.24	27.91
Unified	32.23	13.29	29.36
Transformer	27.67	14.87	27.02
CopyTransformer	29.17	16.15	28.43
with table information			
TabWords	30.43	12.58	23.73
Unified+T	36.45	15.80	32.93
CopyTransformer+T	43.47	29.14	42.49

6 Conclusion

In this paper, we proposed to use factual tables to guide biography summarization. To demonstrate the effectiveness of incorporating table information in generating biography summaries, we developed the first large-scale abstractive biography summarization dataset with tables. We employ several state-of-the-art summarization methods, and adapt these methods to table guided biography summarization task. These methods outperformed other summarization methods in terms of ROUGE which only use the document as input and ignore the table guidance.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (No. 2020AAA0106600), the National Science Founda-

tion of China (NSFC No. 61876196 and NSFC No. 61672058). Rui Yan is partially supported as a Young Fellow of Beijing Institute of Artificial Intelligence (BAAI).

Ethical Impact

In this paper, we propose a table guided biography summarization dataset on Wikipedia. In the real-world application which provided automatic biography summarization service, we will employ human editors to double-check the generated summary to ensure the correctness of content and grammar before publish the summary.

References

Zhangming Chan, Xiuying Chen, Yongliang Wang, Juntao Li, Zhiqiang Zhang, Kun Gai, Dongyan Zhao, and Rui Yan. 2019a. Stick to the facts: Learning towards a fidelity-oriented e-commerce product description generation. In *EMNLP*.

Zhangming Chan, Juntao Li, Xiaopeng Yang, Xiuying Chen, Wenpeng Hu, Dongyan Zhao, and Rui Yan. 2019b. Modeling personalization in continuous space for response generation via augmented wasserstein autoencoders. In *EMNLP*.

Zhangming Chan, Yuchi Zhang, Xiuying Chen, Shen Gao, Zhiqiang Zhang, Dongyan Zhao, and Rui Yan. 2020. Selection and generation: Learning towards multi-product advertisement post generation. In *EMNLP*.

Xiuying Chen, Zhangming Chan, Shen Gao, Meng-Hsuan Yu, Dongyan Zhao, and Rui Yan. 2019. Learning towards abstractive timeline summarization. In *IJCAI*.

Xiuying Chen, Shen Gao, Chongyang Tao, Yan Song, Dongyan Zhao, and Rui Yan. 2018. Iterative document representation learning towards summarization with polishing. In *EMNLP*, pages 4088–4097.

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 484–494.

Shen Gao, Xiuying Chen, Piji Li, Zhangming Chan, Dongyan Zhao, and Rui Yan. 2019a. How to write summaries with patterns? learning towards abstractive summarization through prototype editing. In *EMNLP*.

Shen Gao, Xiuying Chen, Piji Li, Zhaochun Ren, Lidong Bing, Dongyan Zhao, and Rui Yan. 2019b. Abstractive text summarization by incorporating reader comments. In *AAAI*.

Shen Gao, Xiuying Chen, Zhaochun Ren, and Dongyan Zhao Rui Yan. 2020. From standard summarization to new tasks and beyond: Summarization with manifold information. In *IJCAI*.

Shen Gao, Xiuying Chen, Zhaochun Ren, Dongyan Zhao, and Rui Yan. 2021. Meaningful answer generation of e-commerce question-answering. *ACM Trans. Inf. Syst.*, 39(2).

Shen Gao, Zhaochun Ren, Yihong Eric Zhao, Dongyan Zhao, Dawei Yin, and Rui Yan. 2019c. Product-aware answer generation in e-commerce question-answering. In *WSDM*.

Sebastian Gehrmann, Yuntian Deng, and Alexander M Rush. 2018. Bottom-up abstractive summarization. *EMNLP*.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *NAACL-HLT*.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS*.

Wan Ting Hsu, Chieh-Kai Lin, Ming-Ying Lee, Kerui Min, J. J. Tang, and Min Sun. 2018a. A unified model for extractive and abstractive summarization using inconsistency loss. In *ACL*.

Wan-Ting Hsu, Chieh-Kai Lin, Ming-Ying Lee, Kerui Min, Jing Tang, and Min Sun. 2018b. A unified model for extractive and abstractive summarization using inconsistency loss. *ACL*.

Meishan Hu, Aixin Sun, and Ee-Peng Lim. 2008. Comments-oriented document summarization: understanding documents with readers' feedback. In *SIGIR*.

Aishwarya Jadhav and Vaibhav Rajan. 2018. Extractive summarization with swap-net: Sentences and words from alternating pointer networks. In *EMNLP*, pages 142–151.

Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. 2019. Abstractive summarization of reddit posts with multi-level memory networks. In *NAACL*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *EMNLP*.

Yang Liu, Ivan Titov, and Mirella Lapata. 2019. Single document summarization as tree induction. In *NAACL-HLT*.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *AAAI*.

Ramesh Nallapati, Bowen Zhou, Cícero Nogueira dos Santos, Çaglar Gülçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *CoNLL*.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018a. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *EMNLP*.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018b. Ranking sentences for extractive summarization with reinforcement learning. In *NAACL-HLT*.

Shashi Narayan, Nikos Papasarantopoulos, Mirella Lapata, and Shay B. Cohen. 2017. Neural extractive summarization with side information. *CoRR*, abs/1704.04530.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Min Sun, Wan Ting Hsu, Chieh-Kai Lin, Ming-Ying Lee, Kerui Min, and Jing Tang. 2018. A unified model for extractive and abstractive summarization using inconsistency loss. In *ACL*.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*, pages 5998–6008.

Xingxing Zhang, Mirella Lapata, Furu Wei, and Ming Zhou. 2018. Neural latent extractive document summarization. In *EMNLP*.

Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. 2018. Neural document summarization by jointly learning to score and select sentences. *ArXiv*, abs/1807.02305.