Recent Advances in Retrieval-Augmented Text Generation

Deng Cai
The Chinese University of Hong Kong
China
thisisjcykcd@gmail.com

Yan Wang
Tencent AI Lab
China
brandenwang@tecent.com

Lemao Liu
Tencent AI Lab
China
redmondliu@tecent.com

Shuming Shi
Tencent AI Lab
China
shumingshi@tecent.com

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**.

KEYWORDS

information retrieval; text generation

ACM Reference Format:
Deng Cai, Yan Wang, Lemao Liu, and Shuming Shi. 2022. Recent Advances in Retrieval-Augmented Text Generation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3477495.3532682

COVER SHEET

1 TITLE

Recent Advances in Retrieval-Augmented Text Generation

2 LENGTH

The proposed length of this tutorial is expected to be half day. The organization of this tutorial is outlined in Section 8.

3 FORMAT

Online.

4 INTENDED AUDIENCE AND PREREQUISITE KNOWLEDGE

Retrieval-augmented text generation has already attracted increasing attention from both the NLP and IR community. Any audience who may be interested in recent advances of natural language generation, information retrieval, dialogue systems, machine translation, etc, would find it very inspiring and valuable in attending this tutorial.

Although no specific knowledge is required, audiences with basic concepts about information retrieval or deep learning will find it more beneficial in understanding the techniques and analysis to be discussed in this tutorial. To quickly get the main idea of this tutorial, we refer the participants to our survey paper [5]. Moreover, we maintain a paper list[1] for further reading on this topic, which will be dynamically updated to include forthcoming papers.

5 PRESENTERS

Deng Cai (`thisisjcykcd@gmail.com`*).* Deng Cai is a senior Ph.D. student (final-year) at The Chinese University of Hong Kong. Previously, he received his M.Sc. in computer science from Shanghai Jiao Tong University. His research interests include semantic parsing, dialogue systems, and text generation. He has published research papers at prestigious conferences and journals, such as ACL, EMNLP, NAACL, AAAI, and TASLP. He received an outstanding paper award in ACL 2021 for one of his work on retrieval-augmented text generation. He served regularly as program committee members in leading NLP conferences including ACL, EMNLP, NAACL, etc, and was selected as an outstanding reviewer in EMNLP 2020. He was invited to give talks about retrieval-augmented text generation in research institutes such as Amazon AWS AI and Chinese Academy of Sciences. Website: https://jcyk.github.io/

Yan Wang (`brandenwang@tecent.com`*).* Yan Wang is a senior researcher of Natural Language Processing Center, Tencent AI Lab. His research interests include dialogue systems, text generation, and question answering. He has published over 30 research papers in leading conferences and journals, such as ACL, EMNLP, NAACL, AAAI, and TASLP. He received an outstanding paper award in ACL 2021 for one of his work on retrieval-augmented text generation. He served in the program committee of some conferences including ACL, EMNLP, NAACL, AAAI, etc, and was selected as a session chair in ACL 2021 and senior program committee member in AAAI 2022. Website: https://libertywing.github.io/yanwang.github.io/

Lemao Liu (`redmondliu@tecent.com`*).* Lemao Liu is a senior researcher of Natural Language Processing Center, Tencent AI Lab, China. Previously, He was with National Institute of Information and Communications Technology (NICT), Japan. His research interests include machine translation, syntactic parsing, and natural language understanding. He has published more than 40 research papers in leading conferences and journals, such as ACL, EMNLP,

SIGIR '22, July 11–15, 2022, Madrid, Spain

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3532682

[1]The paper list is available at https://github.com/lemaoliu/retrieval-generation-reading-list.

NAACL, COLING, ICLR, AAAI, and JAIR. He received an outstanding paper award in ACL 2021. He served as a publication co-chair in EMNLP 2020 (Findings), a session chair in IJCAI 2019 and ACL 2021, and a senior program committee member in IJCAI 2021. Additionally, he had a tutorial entitled "Scalable Large-Margin Structured Learning: Theory and Algorithms." in ACL 2014. Website: https://lemaoliu.github.io/homepage/

*Shuming Shi (*shumingshi@tecent.com*).* Shuming Shi is a principal researcher of Tencent and Director of Natural Language Processing Center, Tencent AI Lab. His research interests include knowledge mining, natural language understanding, natural language generation, and dialogue systems. He has published over 100 research papers in leading conferences and journals, such as ACL, EMNLP, AAAI, IJCAI, WWW, SIGIR, and TACL. He served as a co-chair of the EMNLP 2021 demonstration track and served in the program committee of some conferences including ACL, EMNLP, WWW, AAAI, etc.

EXTENDED ABSTRACT

Recently retrieval-augmented text generation has achieved state-of-the-art performance in many NLP tasks and has attracted increasing attention of the NLP and IR community, this tutorial thereby aims to present recent advances in retrieval-augmented text generation comprehensively and comparatively. It firstly highlights the generic paradigm of retrieval-augmented text generation, then reviews notable works for different text generation tasks including dialogue generation, machine translation, and other generation tasks, and finally points out some limitations and shortcomings to facilitate future research.

6 MOTIVATION AND OBJECTIVES

Text generation is an important field of NLP and IR that has a wide range of applications. Retrieval-augmented text generation, as a new text generation paradigm that fuses deep learning and information retrieval technology, has achieved state-of-the-art performance in many NLP tasks as well as brought advances in terms of explainable and green AI. This tutorial is supposed to be of great interest to the board NLP and IR community.

The recent developments in this paradigm are distributed in many sub-fields of text generation, such as dialogue response generation, machine translation, and text style transfer. While it demonstrates the universality of retrieval-augmented text generation, it also increases the difficulty for newcomers to get started. They are required to be not only familiar with recent work in both NLP and retrieval technology, but also aware of the characteristics of downstream tasks. We expect that this tutorial would help the audience more deeply understand the development and highlights of retrieval-augmented text generation.

7 RELEVANCE TO THE IR COMMUNITY

Retrieval-augmented text generation, an emerging direction for more efficient, scalable, explainable, and adaptive text generation, has a great impact on the NLP and IR community. Retrieval-augmented text generation has a wide range of application scenarios such as dialog systems and machine translation. This tutorial aims to provide a comprehensive review of recent approaches for retrieval-augmented text generation, including works in dialogue response generation [25], machine translation [14] and others [15]. We introduce the background, motivation, and typical applications of retrieval-augmented text generation, summarize the generic paradigm of retrieval-augmented text generation and present a comparative analysis on three key components of retrieval-augmented text generation, which are retrieval sources, retrieval metrics, and integration methods.

In the main body of this tutorial, we review notable research papers about retrieval-augmented text generation and organize the content with respect to different tasks. Specifically, on the dialogue response generation task, exemplar/template retrieval as an intermediate step has been shown beneficial to informative response generation [25, 26][1, 2] and personalized response generation [9]. In addition, there has been growing interest in knowledge-grounded generation exploring different forms of knowledge such as knowledge bases and external documents [13, 20, 21, 23, 27, 30, 31]. On the machine translation task, we quickly summarize the early work on how the retrieved sentences (called translation memory) are used to improve statistical machine translation (SMT) models [17, 18, 24][6, 7]. Since neural machine translation (NMT) [12] delivers dominant advantages compared with SMT thanks to its end-to-end modeling and sufficient training data, in particular, we intensively highlight several popular methods to integrating translation memory to NMT models [14, 28, 29][3, 4, 10]. We also review the applications of retrieval-augmented text generation in other generation tasks such as abstractive summarization [22], text style transfer [11], code generation [15], paraphrase [16][8], and knowledge-intensive generation [19]. Finally, as the conclusion, we also point out some limitations and shortcomings for recent approaches such that it will be easier for participants to push forward the research about retrieval-augmented text generation.

8 DETAILED SCHEDULE OF THE TUTORIAL

This tutorial is organized as follows:

- Background (15 mins): Background of text generation, the limitations of existing generation models, and the motivation of the retrieval-augmented text generation paradigm
- A New Paradigm: retrieval-augmented text generation (20 mins)
  (a) Retrieval Sources: training corpus, external datasets, and unsupervised corpus
  (b) Retrieval Metrics: sparse-vector retrieval, dense-vector retrieval, and task-specific retrieval
  (c) Integration: how to combine retrieval and generation
- Dialogue Response Generation (40 mins)
  (a) Background: retrieval-based and generation-based dialogue systems
  (b) Shallow Integration: retrieval results as auxiliary input
  (c) Deep Integration: retrieval results as response skeleton or prototype
- Machine Translation (40 mins)

(a) Background: the definitions of translation memory in statistical machine translation (SMT) and neural machine translation (NMT)
(b) Integrating Translation Memory in Inference Phase
(c) Integrating Translation Memory in Training Phase

- A short break (10 mins)
- Other Generation Tasks (40 mins)

(a) Exemplar-driven Generation: Style Transfer, Summarization and Paraphrase Generation
(b) Fact-driven Generation: Language Modeling and Data-to-Text Generation

- Discussion on Main Issues & Conclusion (30 mins)

(a) Retrieval Sensitivity: How to make the performance of retrieval augmented text generation less sensitive to the retrieval quality?
(b) Retrieval Efficiency: How to balance the trade-off between retrieval memory size and retrieval efficiency?
(c) Multi-Modalities: Is it possible to extend retrieval memory to other modalities?

9 SUPPORT MATERIALS

We maintain a reading list for this tutorial at https://github.com/lemaoliu/retrieval-generation-reading-list and more details about this topic can be found in our recent survey paper [5]. In addition, the webpage of this tutorial is https://github.com/lemaoliu/retrieval-generation-tutorial.

RELATED PUBLICATIONS BY THE PRESENTERS

[1] Deng Cai, Yan Wang, Wei Bi, Zhaopeng Tu, Xiaojiang Liu, Wai Lam, and Shuming Shi. 2019. Skeleton-to-Response: Dialogue Generation Guided by Retrieval Memory. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 1219–1228.

[2] Deng Cai, Yan Wang, Wei Bi, Zhaopeng Tu, Xiaojiang Liu, and Shuming Shi. 2019. Retrieval-guided Dialogue Response Generation via a Matching-to-Generation Framework. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 1866–1875.

[3] Deng Cai, Yan Wang, Huayang Li, Wai Lam, and Lemao Liu. 2021. Neural Machine Translation with Monolingual Translation Memory. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 7307–7318. https://doi.org/10.18653/v1/2021.acl-long.567

[4] Qiuxiang He, Guoping Huang, Qu Cui, Li Li, and Lemao Liu. 2021. Fast and accurate neural machine translation with translation memory. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 3170–3180.

[5] Huayang Li, Yixuan Su, Deng Cai, Yan Wang, and Lemao Liu. 2022. A Survey on Retrieval-Augmented Text Generation. arXiv:2202.01110 [cs.CL]

[6] Lemao Liu, Hailong Cao, Taro Watanabe, Tiejun Zhao, Mo Yu, and Conghui Zhu. 2012. Locally training the log-linear model for SMT. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. 402–411.

[7] Lemao Liu, Tiejun Zhao, Taro Watanabe, Hailong Cao, and Conghui Zhu. 2014. Discriminative Training for Log-Linear Based SMT: Global or Local Methods. *ACM Transactions on Asian Language Information Processing (TALIP)* 13, 4 (2014), 1–25.

[8] Yixuan Su, David Vandyke, Simon Baker, Yan Wang, and Nigel Collier. 2021. Keep the Primary, Rewrite the Secondary: A Two-Stage Approach for Paraphrase Generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, Online, 560–569. https://doi.org/10.18653/v1/2021.findings-acl.50

[9] Yixuan Su, Wang Yan, Deng Cai, Simon Baker, Anna Korhonen, and Nigel Collier. 2021. Prototype-to-style: Dialogue generation with style-aware editing on retrieval memory. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2021).

[10] Mengzhou Xia, Guoping Huang, Lemao Liu, and Shuming Shi. 2019. Graph based translation memory for neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 7297–7304.

[11] Fei Xiao, Liang Pang, Yanyan Lan, Yan Wang, Huawei Shen, and Xueqi Cheng. 2021. Transductive Learning for Unsupervised Text Style Transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2510–2521.

REFERENCES

[12] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

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

[14] Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor OK Li. 2018. Search engine guided neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[15] Tatsunori B Hashimoto, Kelvin Guu, Yonatan Oren, and Percy S Liang. 2018. A retrieve-and-edit framework for predicting structured outputs. In *Advances in Neural Information Processing Systems*. 10052–10062.

[16] Amirhossein Kazemnejad, Mohammadreza Salehi, and Mahdieh Soleymani Baghshah. 2020. Paraphrase Generation by Learning How to Edit from Samples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 6010–6021. https://doi.org/10.18653/v1/2020.acl-main.535

[17] Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*. 127–133. https://aclanthology.org/N03-1017

[18] Philipp Koehn and Jean Senellart. 2010. Convergence of translation memory and statistical machine translation. In *Proceedings of AMTA Workshop on MT Research and the Translation Industry*. 21–31.

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

[20] Zekang Li, Cheng Niu, Fandong Meng, Yang Feng, Qian Li, and Jie Zhou. 2019. Incremental Transformer with Deliberation Decoder for Document Grounded Conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 12–21.

[21] Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. Learning to select knowledge for response generation in dialog systems. *arXiv preprint arXiv:1902.04911* (2019).

[22] Hao Peng, Ankur P. Parikh, Manaal Faruqui, Bhuwan Dhingra, and Das Dipanjan. 2019. Text Generation with Exemplar-based Adaptive Decoding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

[23] Lianhui Qin, Michel Galley, Chris Brockett, Xiaodong Liu, Xiang Gao, William B Dolan, Yejin Choi, and Jianfeng Gao. 2019. Conversing by Reading: Contentful Neural Conversation with On-demand Machine Reading. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 5427–5436.

[24] Michel Simard and Pierre Isabelle. 2009. Phrase-based machine translation in a computer-assisted translation environment. *Proceedings of the Twelfth Machine Translation Summit (MT Summit XII)* (2009), 120–127.

[25] Jason Weston, Emily Dinan, and Alexander Miller. 2018. Retrieve and Refine: Improved Sequence Generation Models For Dialogue. In *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*. 87–92.

[26] Yu Wu, Furu Wei, Shaohan Huang, Yunli Wang, Zhoujun Li, and Ming Zhou. 2019. Response generation by context-aware prototype editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 7281–7288.

[27] Zeqiu Wu, Michel Galley, Chris Brockett, Yizhe Zhang, Xiang Gao, Chris Quirk, Rik Koncel-Kedziorski, Jianfeng Gao, Hannaneh Hajishirzi, Mari Ostendorf, et al. 2021. A Controllable Model of Grounded Response Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 14085–14093.

[28] Jitao Xu, Josep M Crego, and Jean Senellart. 2020. Boosting neural machine translation with similar translations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 1580–1590.

[29] Jingyi Zhang, Masao Utiyama, Eiichro Sumita, Graham Neubig, and Satoshi Nakamura. 2018. Guiding neural machine translation with retrieved translation pieces. *arXiv preprint arXiv:1804.02559* (2018).

[30] Yizhe Zhang, Siqi Sun, Xiang Gao, Yuwei Fang, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. 2021. Joint retrieval and generation training for grounded text generation. *arXiv preprint arXiv:2105.06597* (2021).

[31] Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A dataset for document grounded conversations. *arXiv preprint arXiv:1809.07358* (2018).