IJCAI22 Tutorial Proposal - Recent Advances in Retrieval-Augmented Text Generation

Deng Cai[♡], **Yan Wang**[♠], **Lemao Liu**[♠], **Shuming Shi**[♠]
[♡]The Chinese University of Hong Kong
[♠]Tencent AI Lab
thisisjcykcd@gmail.com
{brandenwang,redmondliu,shumingshi}@tencent.com

1 Title

Recent Advances in Retrieval-Augmented Text Generation

2 A two-sentence description of the tutorial

Recently, retrieval-augmented text generation has achieved state-of-the-art performance in many NLP tasks and attracted increasing attention of the computational linguistics community. Since retrieval-augmented generation is distributed in many subdomains of information retrieval and text generation, it increases the difficulty for newcomers to get started. Therefore, this tutorial aims to introduce recent advances in retrieval-augmented text generation. It firstly highlights the generic paradigm of retrieval-augmented generation, and then it reviews notable approaches according to different tasks including dialogue generation, machine translation, and other generation tasks. Finally, it points out some limitations and shortcomings for recent approaches to facilitate future research.

3 A two-paragraph description of the tutorial

This tutorial aims to review many representative approaches for retrieval-augmented text generation tasks including dialogue response generation (Weston et al., 2018), machine translation (Gu et al., 2018) and others (Hashimoto et al., 2018). In this tutorial, we firstly introduce the generic paradigm of retrieval-augmented generation and briefly present three key components under this paradigm, which are retrieval sources, retrieval metrics (Ramos et al., 2003; Johnson et al., 2017) and generation models (Bahdanau et al., 2014), respectively.

Then, we review notable research papers about retrieval-augmented 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 (Weston et al., 2018; Wu et al., 2019; Cai et al., 2019a,b) and personalized response generation (Su et al., 2021b). In addition, there has been growing interest in knowledge-grounded generation exploring different forms of knowledge such as knowledge bases and external documents (Dinan et al., 2018; Zhou et al., 2018; Lian et al., 2019; Li et al., 2019; Qin et al., 2019; Wu et al., 2021; Zhang et al., 2021). 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) (Koehn et al., 2003) models (Simard and Isabelle, 2009; Koehn and Senellart, 2010; Liu et al., 2012). Since neural machine translation (NMT) (Bahdanau et al., 2014) 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 (Gu et al., 2018; Zhang et al., 2018; Xu et al., 2020; He et al., 2021; Cai et al., 2021). We also review the applications of retrieval-augmented generation in other generation tasks such as abstractive summarization (Peng et al., 2019), code generation (Hashimoto et al., 2018), paraphrase (Kazemnejad et al., 2020; Su et al., 2021a), and knowledge-intensive generation (Lewis et al., 2020).

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 generation. The detailed organization of this tutorial is outlined in Section 6.

4 Proposed length of the tutorial

1/4 or 1/2 day (consisting of one or two 1:45h slots respectively)

5 Outline of the tutorial

This tutorial is organized as follows:

- Background
- Paradigm: Retrieval augmented Generation
  (a) Retrieval Sources
  (b) Retrieval Metrics
  (c) Generation Models
- Dialogue Response Generation
  (a) Exemplar/Template Guided Generation
  (b) Knowledge Grounded Generation
- Machine Translation
  (a) Translation Memory for SMT
  (b) Translation Memory for NMT
- Other Generation Tasks
  (a) Exemplar-driven Generation
  (b) Fact-driven Generation
- Conclusion

6 Outline of the tutorial

This tutorial is organized as follows:

- Background: the limitation of pre-trained models, and the motivation of the retrieval-augmented paradigm
- Paradigm: Retrieval augmented Generation
  (a) Retrieval Sources: training corpus, external datasets, and large-scale unsupervised corpus;
  (b) Retrieval Metrics: sparse-vector retrieval, dense-vector retrieval, and training-based retrieval
  (c) Integration of retrieval results and generation models
- Dialogue Response Generation
  (a) Background: retrieval-based dialogue systems and generation-based dialogue systems
  (b) Shallow Integration: retrieval results as an auxiliary guidance
  (c) Deep Integration: retrieval results as an response skeleton or prototype
- Machine Translation
  (a) Background: the definition of translation memory in statistical machine translation (SMT) and neural machine translation (NMT)
  (b) Translation Memory for statistical machine translation (SMT) and neural machine translation (NMT)
- Other Generation Tasks
  (a) Exemplar-driven Generation
  (b) Fact-driven Generation
- Conclusion

7 Target audience for the tutorial

A brief characterization of the potential target audience for the tutorial, including prerequisite knowledge.

We would assume acquaintance with basic concepts about search engines and neural networks, such as those included in most introductory NLP courses.

To quickly get the main idea of this tutorial, we refer the participants to the papers mentioned in section 1. Moreover, we maintain a paper list for further reading on this topic, [1] which will be dynamically updated to include forthcoming papers.

8 Ethical concerns

9 Why this tutorial

A brief description of why the tutorial topic should be of interest to a substantial part of the IJCAI-ECAI audience, and which of the above objectives are best served by the tutorial. Retrieval-augmented text generation such as dialogue response generation and machine translation[2], as a new text generation paradigm that fuses emerging deep learning technology and traditional retrieval technology, has achieved state-of-the-art (SOTA) performance in many NLP tasks and attracted the attention of the computational linguistics community (Dinan et al., 2018; Cai et al.,

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

[2] Throughout this tutorial, machine translation is considered to be a kind of text generation task, although it is a very popular task.

2021). Compared with generation-based counterpart, this new paradigm has some unique advantages: 1) The knowledge is no longer implicitly stored in model parameters, but is explicitly acquired in a plug-and-play manner; 2) Instead of generating from scratch, the paradigm generating text from some retrieved human-written reference, which potentially alleviates the difficulty of text generation.

The recent developments in this paradigm are distributed in many sub-domains of text generation, such as dialogue response generation, machine translation, and style transfer. It demonstrates the universality of retrieval-augmented text generation but also increases the difficulty for newcomers to get started. They are required to be not only familiar with recent work in both neural NLP and retrieval technology, but also aware of the characteristics of downstream tasks. A comprehensive tutorial may fill this gap and introduce the nascent field of retrieval-augmented text generation.

10 Presenters

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 as a regular program committee member 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 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 AAAI. 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, WWW, 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 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, 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 as "Scalable Large-Margin Structured Learning: Theory and Algorithms." in ACL 2014. Website: https: //lemaoliu.github.io/homepage/

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.

References

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

Deng Cai, Yan Wang, Wei Bi, Zhaopeng Tu, Xiaojiang Liu, Wai Lam, and Shuming Shi. 2019a. 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)*, pages 1219–1228.

Deng Cai, Yan Wang, Wei Bi, Zhaopeng Tu, Xiaojiang Liu, and Shuming Shi. 2019b. 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)*, pages 1866–1875.

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)*, pages 7307–7318, Online. Association for Computational Linguistics.

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*.

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*, volume 32.

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*, pages 10052–10062.

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)*, pages 3170–3180.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*.

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*, pages 6010–6021, Online. Association for Computational Linguistics.

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*, pages 127–133.

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*, pages 21–31.

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*.

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*, pages 12–21.

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*.

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*, pages 402–411.

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*.

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*, pages 5427–5436.

Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. Citeseer.

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)*, pages 120–127.

Yixuan Su, David Vandyke, Simon Baker, Yan Wang, and Nigel Collier. 2021a. Keep the primary, rewrite the secondary: A two-stage approach for paraphrase generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 560–569, Online. Association for Computational Linguistics.

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

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, pages 87–92.

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*, volume 33, pages 7281–7288.

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*, volume 35, pages 14085–14093.

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*, pages 1580–1590.

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*.

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*.

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