Data Efficient Summarization by Data Sampling, Linguistic Inductive Bias, and Unsupervised Learning

Overview

Summarization breaks the barrier of information overload by empowering readers to quickly gain information and acquire knowledge from documents. While state-of-the-art summarization models have demonstrated remarkable performance, they hinge on the availability of large amounts of training data, limiting their applications to high-resource domains such as news articles [12, 9, 4, 14]. However, creating large-scale data for every new domain is labor-intensive and highly costly. Therefore, current summarization models still fall short in real world applications due to the lack of abundant training data.

My long-term research agenda is to create *Efficient*, *Controllable*, and *Trustworthy* summarization systems in diverse, realistic, and meaningful scenarios such as healthcare, education, and legal domains. The goal of this proposal is to focus on the first of the three challenges: improving data efficiency by reducing the number of annotated data required for training. To this end, our research plan enhances summarization models through three aspects including training data sampling, linguistic inductive bias, and unsupervised learning paradigms.

- **Training Data Sampling.** Our hypothesis is that not all training examples are equally useful for finetuning and in-context learning of language models for summarization. Therefore, we propose data sampling strategies by jointly considering informativeness and diversity of examples to improve data efficiency [8, 13]. To this end, we apply domain adaptation and active learning to the summarization model under a transfer learning setting. We will adapt models trained on abundant out-of-domain datasets, and we then select the most informative in-domain instances to label for the model to continuously train on.
- **Linguistic Inductive Bias for Constrained Decoding.** Our hypothesis is that current deep learning language models are driven by empirical approaches without explicit notions of *Information Importance* [7, 10], and thus they rely on a huge number of examples to learn to summarize. Therefore, we propose to accelerate model learning by injecting inductive bias informed by the linguist theory of summarization into deep learning models through constrained decoding. Specifically, we adopt the information-theoretic framework [11] to decompose importance into relevance, informativeness, and redundancy. Each is measured by cross-entropy of semantic unit distributions among source, summary, and user background knowledge. We use this formal definition of importance to guide model decoders by constrained beam search [6]. This encourages generating summaries that contain relevant semantic units from sources without redundancy while bringing most new information to a user with certain background knowledge.
- **Unsupervised Learning Paradigms.** Current deep learning summarization models lack data efficiency because of their supervised learning nature [16, 1, 3]. Therefore, we propose to shift the model training paradigm to an unsupervised framework based on reinforced contrastive learning without using human-annotated summaries [15]. Our framework train an abstractive summarizer to produce multiple candidate summaries, and we measure the quality of summaries using our importance notion as the reward signal for reinforcement learning. Furthermore, we will also enhance summarization quality and coverage by adding contrastive learning supervision by matching the produced summaries with the source document and vice versa. We will also explore other unsupervised learning models such as diffusion language models [2, 5] by performing denoising training over words to reconstruct the whole document in both continuous and discrete spaces, and the intermediate representations will serve as the summaries of different granularities.

Intellectual Merit
Our proposal delivers fundamental solutions to improve data efficiency when creating summarization systems in realistic and meaningful settings across different domains (e.g., healthcare, science, legal). It reduces the number of training examples by redesigning data sampling strategies, injecting linguistic theory into deep learning models, and innovating unsupervised learning paradigms. These research innovations also offer promising solutions to other NLG tasks. Further, our research outcomes significantly contribute to our long-term pursuit of efficient, trustworthy, and human-centered NLP and AI.

Broader Impacts
The research findings from this proposal will have potential broader societal applicability for summarizing scientific reports to foster interdisciplinary research and pubic scientific literacy, summarizing policy reports to support community and government decision-making, and summarizing civil rights lawsuit reports for promoting democracy and legal construction.

References

[1] Alexander Fabbri, Simeng Han, Haoyuan Li, Haoran Li, Marjan Ghazvininejad, Shafiq Joty, Dragomir Radev, and Yashar Mehdad. Improving zero and few-shot abstractive summarization with intermediate fine-tuning and data augmentation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 704–717, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.57. URL `https://aclanthology.org/2021.naacl-main.57`.

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[3] Philippe Laban, Andrew Hsi, John Canny, and Marti A Hearst. The summary loop: Learning to write abstractive summaries without examples. *arXiv preprint arXiv:2105.05361*, 2021.

[4] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL `https://aclanthology.org/2020.acl-main.703`.

[5] Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *arXiv preprint arXiv:2205.14217*, 2022.

[6] Ximing Lu, Sean Welleck, Peter West, Liwei Jiang, Jungo Kasai, Daniel Khashabi, Ronan Le Bras, Lianhui Qin, Youngjae Yu, Rowan Zellers, Noah Smith, and Yejin Choi. NeuroLogic a*esque decoding: Constrained text generation with lookahead heuristics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 780–799, Seattle, United States, July 2022. Association for Computational Linguistics. URL `https://aclanthology.org/2022.naacl-main.57`.

[7] Hans Peter Luhn. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165, 1958.

[8] Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. Active learning by acquiring contrastive examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 650–663, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.51. URL `https://aclanthology.org/2021.emnlp-main.51`.

[9] Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/K16-1028. URL `https://aclanthology.org/K16-1028`.

[10] Ani Nenkova and Kathleen McKeown. A survey of text summarization techniques. In *Mining text data*, pages 43–76. Springer, 2012.

[11] Maxime Peyrard. A simple theoretical model of importance for summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1059–1073, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1101. URL `https://aclanthology.org/P19-1101`.

[12] Evan Sandhaus. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752, 2008.

[13] Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. Selective annotation makes language models better few-shot learners. *arXiv preprint arXiv:2209.01975*, 2022.

[14] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR, 2020.

[15] Rui Zhang, Yangfeng Ji, Yue Zhang, and Rebecca J Passonneau. Contrastive data and learning for natural language processing. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts*, pages 39–47, 2022.

[16] Jiawei Zhou and Alexander Rush. Simple unsupervised summarization by contextual matching. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5101–5106, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1503. URL `https://aclanthology.org/P19-1503`.