A Survey of Very Recent Text Summarization Approaches

Karthik Venkat Ramanan
kv16@illinois.edu

1 Introduction

Text summarization aims to produce a coherent, informative and factual summary of a piece of text. With the advent of transformer[30] models, text summarization in the last couple of years has improved substantially across a number of dimensions. But transformers bring with them their own set of challenges such as their large memory and compute requirements. In this survey, we'll be looking at some very recent text summarization approaches from the last couple of years. We'll also study the challenges and avenues for future work.

2 A Brief Taxonomy of Modern Text Summarization Approaches

Modern text summarization approaches can be classified across a variety of axes- Abstractive vs Extractive, Single-Document vs Multi-Document, Long vs Short Document, Fully Differentiable vs Reinforcement Learning Based, etc. Abstractive approaches[17, 33, 8, 32, 34, 28] aim to rephrase the source text into a coherent summary. Extractive Approaches[18, 13, 12, 31] on the other hand, pick key sentences of importance from a source text. In abstractive approaches, having large ngram overlaps between the source text and the summary is frowned upon and the metric 'abstractivity' explicitly studies this. Multi-Document approaches[35, 22, 16, 19] approaches aim to produce cogent summaries by collating information from multiple documents that provide multiple views of the data. Most text summarization approaches typically truncate the input document to 400 words. Long-Document Summarization approaches[9, 2, 23, 10] either use hierarchical extractive approaches, 2-step approaches or sparse self-attention to scale to documents that are up to 7000 words long. Reinforcement Learning based approaches[5, 11, 39, 21] usually use policy-gradient approaches to maximize a reward function on a text generation metric like the ROUGE score. In the following sections we shall briefly summarize a few cutting-edge approaches to text summarization.

3 Summarizing Text on Any Aspects: A Knowledge-Informed Weakly-Supervised Approach

This work[28] tackles the problem of summarizing text on certain arbitrary aspects that the user can specify. The paper performs Named Entity Recognition using SpaCy's[1] WikiNER on abstractive summaries to extract sentences that have particular entities. This way, we end up with a supervised dataset for summarization based on key aspects. The authors then augment this dataset using ConceptNet[27] by also including entities that are near the identified entities semantically. They then feed the concatenation of the extracted entities and the source document to BART[15] for summarization. At inference time the trained model can be fed any arbitrary entity along with a document to be summarized.

4 Friendly Topic Assistant for Transformer Based Abstractive Summarization

This work[32] proposes a novel approach to use topic modelling to keep a summarization model aware of global semantics while summarizing a document. The authors use Poisson Factor Analysis[40] using either Variational AutoEncoders[38] or Gibbs sampling to get a distribution over the topics of a model. They then obtain topic embeddings by multiplying the topic-word matrix and a word embeddings matrix. They use BERT[7] as their encoder-decoder model and attend to the topic distribution along with the input representations jointly in both the encoder and decoder. So this model essentially bootstraps additional global information from a document and uses it to enhance summarization.

5 Multi-hop Inference for Question-driven Summarization

Deng et al. propose a pointer-generator[26] based approach to summarize documents based on specific questions. So in a way, this work also falls under the umbrella of Question-Answering techniques. Question-driven summarization is different from factoid-based question answering in that the questions require deeper reasoning drawing from different portions of the source document. The authors use a BiLSTM to generate document and sentence embeddings. The generator is a pointer-generator LSTM network that attends to a) the question, b) the source document and c) the sentence representations.

[1] https://spacy.io/

6 Enhancing Factual Consistency of Abstractive Summarization

This work[41] builds upon previou work[14] showing that the summaries produced by abstractive summarization approaches are riddled with factual errors. The approach uses a two step process- fact-based summary generation followed by summary correction. OpenIE[1] is used to extract facts in the form of relation triples from the source document. A graph attention network is used to obtain node embeddings from the induced OpenIE knowledge graph. The authors use BART as the encoder-decoder network. Node embeddings generated using GAT are integrated into the generation process. Fact correction is simply a refining step using another transformer.

7 On Extractive and Abstractive Neural Document Summarization with Transformer Language Models

Pilault et al. propose a two-step approach to long-document summarization. The extractive step is a hierarchical LSTM that uses a pointer network to point to relevant sentences from the source document to summarize. For the extractive step, ground-truth extractive summaries are generated using ROUGE score similarity. The abstractive step feeds these sentences concatenated with as much of the source document as fits in memory to GPT-2[25]. The paper achieves competitive ROUGE scores even without using an explicit copy-mechanism.

8 Conclusion and Avenues for Future Work

In this survey we have taken a general look at the landscape of very recent summarization approaches as well as the ways some recent papers solve summarization problems. There is a lot of room for improvement in neural summarization. Recent works[3, 37] have shown that local + random global attention is adequate to obtain performance close to the state of the art. So we need to take a closer look at the expensive full-attention of modern transformer models and use domain knowledge to restrict this attention to specific neighbourhoods. Doing so could also address the scalability issues inherent to transformers. Also, there is only limited work on enhancing the factual accuracy of abstractive summaries. Using a combination of KG embeddings [4, 29, 36] and rules generated from rule-based techniques[20] to enforce relations at train time is a vital direction to look at.

References

[1] Angeli, G., M. J. Johnson Premkumar, and C. D. Manning (2015, July). Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Beijing, China, pp. 344–354. Association for Computational Linguistics.

[2] Bajaj, A., P. Dangati, K. Krishna, P. Ashok Kumar, R. Uppaal, B. Windsor, E. Brenner, D. Dotterrer, R. Das, and A. McCallum (2021, August). Long document summarization in a low resource setting using pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, Online, pp. 71–80. Association for Computational Linguistics.

[3] Beltagy, I., M. E. Peters, and A. Cohan (2020). Longformer: The long-document transformer. *arXiv:2004.05150*.

[4] Bordes, A., N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko (2013). Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems*, pp. 2787–2795.

[5] Chen, Y.-C. and M. Bansal (2018). Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of ACL*.

[6] Deng, Y., W. Zhang, and W. Lam (2020, November). Multi-hop inference for question-driven summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, pp. 6734–6744. Association for Computational Linguistics.

[7] Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[8] Fabbri, A., F. Rahman, I. Rizvi, B. Wang, H. Li, Y. Mehdad, and D. Radev (2021, August). ConvoSumm: Conversation summarization benchmark and improved abstractive summarization with argument mining. 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)*, Online, pp. 6866–6880. Association for Computational Linguistics.

[9] Gidiotis, A. and G. Tsoumakas (2020). A divide-and-conquer approach to the summarization of long documents. *IEEE/ACM Transactions on Audio, Speech, and Language Processing 28*, 3029–3040.

[10] Huang, L., S. Cao, N. Parulian, H. Ji, and L. Wang (2021, June). Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online, pp. 1419–1436. Association for Computational Linguistics.

[11] Huang, L., L. Wu, and L. Wang (2020, 01). Knowledge graph-augmented abstractive summarization with semantic-driven cloze reward. pp. 5094–5107.

[12] Jia, R., Y. Cao, H. Tang, F. Fang, C. Cao, and S. Wang (2020, November). Neural extractive summarization with hierarchical attentive heterogeneous graph network. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, pp. 3622–3631. Association for Computational Linguistics.

[13] Jin, H., T. Wang, and X. Wan (2020, July). Multi-granularity interaction network for extractive and abstractive multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, pp. 6244–6254. Association for Computational Linguistics.

[14] Kryscinski, W., B. McCann, C. Xiong, and R. Socher (2020, November). Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, pp. 9332–9346. Association for Computational Linguistics.

[15] Lewis, M., Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer (2019). BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR abs/1910.13461*.

[16] Li, W., X. Xiao, J. Liu, H. Wu, H. Wang, and J. Du (2020). Leveraging graph to improve abstractive multi-document summarization.

[17] Liu, Y. (2019). Fine-tune BERT for extractive summarization. *CoRR abs/1903.10318*.

[18] Liu, Y. and M. Lapata (2019). Text summarization with pretrained encoders. *CoRR abs/1908.08345*.

[19] Mao, Y., Y. Qu, Y. Xie, X. Ren, and J. Han (2020, November). Multi-document summarization with maximal marginal relevance-guided reinforcement learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, pp. 1737–1751. Association for Computational Linguistics.

[20] Meilicke, C., M. Fink, Y. Wang, D. Ruffinelli, R. Gemulla, and H. Stuckenschmidt (2018). Fine-grained evaluation of rule- and embedding-based systems for knowledge graph completion. In *SEMWEB*.

[21] Narayan, S., S. Cohen, and M. Lapata (2018, 01). Ranking sentences for extractive summarization with reinforcement learning. pp. 1747–1759.

[22] Pasunuru, R., M. Liu, M. Bansal, S. Ravi, and M. Dreyer (2021, June). Efficiently summarizing text and graph encodings of multi-document clusters. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online, pp. 4768–4779. Association for Computational Linguistics.

[23] Pilault, J., R. Li, S. Subramanian, and C. Pal (2020a, November). On extractive and abstractive neural document summarization with transformer language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, pp. 9308–9319. Association for Computational Linguistics.

[24] Pilault, J., R. Li, S. Subramanian, and C. Pal (2020b, November). On extractive and abstractive neural document summarization with transformer language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, pp. 9308–9319. Association for Computational Linguistics.

[25] Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever (2019). Language models are unsupervised multitask learners.

[26] See, A., P. Liu, and C. Manning (2017). Get to the point: Summarization with pointer-generator networks. In *Association for Computational Linguistics*.

[27] Speer, R., J. Chin, and C. Havasi (2019, 05). Conceptnet 5.5: An open multilingual graph of general knowledge.

[28] Tan, B., L. Qin, E. Xing, and Z. Hu (2020, November). Summarizing text on any aspects: A knowledge-informed weakly-supervised approach. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, pp. 6301–6309. Association for Computational Linguistics.

[29] Trouillon, T., J. Welbl, S. Riedel, E. Gaussier, and G. Bouchard (2016). Complex embeddings for simple link prediction. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pp. 2071–2080. JMLR.org.

[30] Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin (2017). Attention is all you need. In I. Guyon, U. V.

Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30*, pp. 5998–6008. Curran Associates, Inc.

[31] Wang, D., P. Liu, Y. Zheng, X. Qiu, and X. Huang (2020, July). Heterogeneous graph neural networks for extractive document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, pp. 6209–6219. Association for Computational Linguistics.

[32] Wang, Z., Z. Duan, H. Zhang, C. Wang, L. Tian, B. Chen, and M. Zhou (2020, November). Friendly topic assistant for transformer based abstractive summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, pp. 485–497. Association for Computational Linguistics.

[33] Wu, W., W. Li, X. Xiao, J. Liu, Z. Cao, S. Li, H. Wu, and H. Wang (2021, August). BASS: Boosting abstractive summarization with unified semantic graph. 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)*, Online, pp. 6052–6067. Association for Computational Linguistics.

[34] Xu, S., H. Li, P. Yuan, Y. Wu, X. He, and B. Zhou (2020, July). Self-attention guided copy mechanism for abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, pp. 1355–1362. Association for Computational Linguistics.

[35] Xu, Y. and M. Lapata (2020, November). Coarse-to-fine query focused multi-document summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, pp. 3632–3645. Association for Computational Linguistics.

[36] Yang, B., W. Yih, X. He, J. Gao, and L. Deng (2015). Embedding entities and relations for learning and inference in knowledge bases. In Y. Bengio and Y. LeCun (Eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

[37] Zaheer, M., G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, et al. (2020). Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems 33*.

[38] Zhang, H., B. Chen, D. Guo, and M. Zhou (2018). Whai: Weibull hybrid autoencoding inference for deep topic modeling. In *ICLR 2018*.

[39] Zhang, H., Y. Gong, Y. Yan, N. Duan, J. Xu, J. Wang, M. Gong, and M. Zhou (2019). Pretraining-based natural language generation for text summarization. In *CoNLL*.

[40] Zhou, M., L. Hannah, D. Dunson, and L. Carin (2011, 12). Beta-negative binomial process and poisson factor analysis.

[41] Zhu, C., W. Hinthorn, R. Xu, Q. Zeng, M. Zeng, X. Huang, and M. Jiang (2021, June). Enhancing factual consistency of abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online, pp. 718–733. Association for Computational Linguistics.