A Survey on the Limitations of Text Summarization Models

Team Members

David Gary	801325583
Michael Smalenberger	800984973
Matthew Seman	801156143
Ryan Marsh	800552800

Sentiment Analysis Datasets

Reddit TIFU Dataset[4]
Multi-News Dataset[2]
S2ORC Dataset[8]

Models

XLNet Model[15]	T5 Model[13]
BART Model[6]	BART-X Model[6]
Pegasus Model[16]	Pegasus-X Model[12]

Summarization Datasets

Gigaword Dataset[3, 14]
CNN/DailyMail Dataset[9]
XSum Dataset[10]

Introduction

With the increase in the availability of data in many fields, the need for summarizing texts has become increasingly important. Condensing large amounts of information into concise and consumable bits of information can have significant personal and business implications. For example, helping readers parse through complex topics by reducing the information to its salient points can help individuals quickly apply new information and gain expertise. Similarly, businesses have begun to use social media posts to gauge public sentiment toward cryptocurrencies in order to hedge their portfolios[1, 5]. Wading through this enormous amount of information would not be possible without the ability to summarize posts and accurately gauge sentiment.

Artificial intelligence, and specifically Intelligent Systems (IS) using natural language processing (NLP), is increasingly used in order to accomplish this task. As progress continues to be made in developing and fine-tuning NLP applications for text summarization, there has become an increasing number of models available from which to choose. These models are often created in order to accomplish a specific objective and are not intended to be highly effective in every scenario. Hence, when measuring the efficacy of different models, one may arrive at different conclusions. While this does certainly not mean that one model is necessarily better than another, it is essential to understand what leads to these different outcomes. Levels of efficacy when applied in different scenarios.

When intelligent systems summarize text, they may arrive at different conclusions based on the model they implement. For example, two methods of scoring are ROUGE[7] and Pyramid[11]. These two scoring systems produce different evaluations of the intelligent system. In order to continue to make progress in this field, it is crucial to investigate what causes these evaluations to be different and why.

Problem Statement

Our system implements text analysis and summarization following the current standards of generalizability. We make our platform available so that anyone can use it to summarize text or assess the tone of text either using several publicly available datasets implemented by our system or by the user loading a different dataset of their choice. However, since the primary purpose of this project is to show the limitations of generalized text summarization models, the user should be aware that we do not strive to achieve perfect accuracy in summarization or sentiment analysis.

Instead, we highlight the differences of the models implemented, and therefore the majority of the remaining analysis and discussion will focus on these differences.

Literature Survey

Next, we will build on the XLNet framework to produce a generalized text summarizer. This will include a score report that shows how well the model performs according to the Pyramid and ROUGE schemes.

Our System

Something here.

References

[1] Colianni, S. G., Rosales, S. M., and Signorotti, M. Algorithmic trading of cryptocurrency based on twitter sentiment analysis, 2015.

[2] Fabbri, A. R., Li, I., She, T., Li, S., and Radev, D. R. Multi-news: a large-scale multi-document summarization dataset and abstractive hierarchical model, 2019.

[3] Graff, D., Kong, J., Chen, K., and Maeda, K. English gigaword. *Linguistic Data Consortium, Philadelphia 4*, 1 (2003), 34.

[4] Kim, B., Kim, H., and Kim, G. Abstractive summarization of reddit posts with multi-level memory networks, 2018.

[5] Kim, Y. B., Kim, J., Kim, W., Im, J., Kim, T., Kang, S., and Kim, C.-H. Predicting fluctuations in cryptocurrency transactions based on user comments and replies. *PLOS ONE 11* (08 2016), e0161197.

[6] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019.

[7] Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out* (Barcelona, Spain, July 2004), Association for Computational Linguistics, pp. 74–81.

[8] Lo, K., Wang, L. L., Neumann, M., Kinney, R., and Weld, D. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Online, July 2020), Association for Computational Linguistics, pp. 4969–4983.

[9] Nallapati, R., Xiang, B., and Zhou, B. Sequence-to-sequence rnns for text summarization. *CoRR abs/1602.06023* (2016).

[10] Narayan, S., Cohen, S. B., and Lapata, M. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization, 2018.

[11] Nenkova, A., Passonneau, R., and McKeown, K. The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Trans. Speech Lang. Process. 4*, 2 (may 2007), 4–es.

[12] Phang, J., Zhao, Y., and Liu, P. J. Investigating efficiently extending transformers for long input summarization, 2022.

[13] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer, 2019.

[14] Rush, A. M., Chopra, S., and Weston, J. A neural attention model for abstractive sentence summarization. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (2015).

[15] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. V. Xlnet: Generalized autoregressive pretraining for language understanding, 2019.

[16] Zhang, J., Zhao, Y., Saleh, M., and Liu, P. J. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization, 2019.