Abstractive News Article Summarization Using Sequence-to-Sequence Models

The Anson L. Clark Summer Research Program

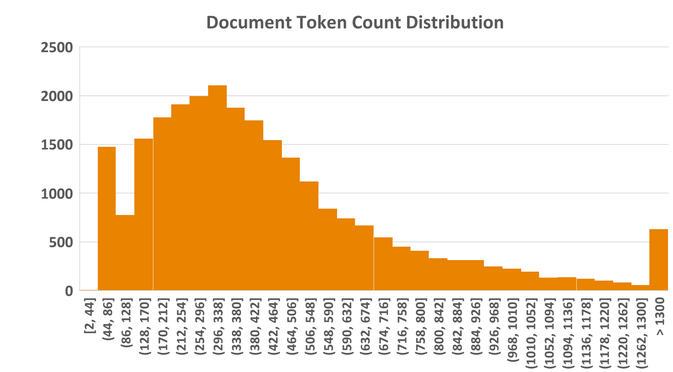
Aditya Gunvantray Rathod, Dr. Anurag Nagar, Department of Computer Science

Introduction

Thousands of temporally-sensitive news articles are published daily by various news outlets. The high volume of news published prevents the average user from staying reasonably up-to-date on current events. We introduce Reportik, an attention-based sequence-to-sequence model which utilizes recent advances in neural machine translation[3] and neural text summarization[1, 2] to produce short abstractive summaries of news articles. Reportik was trained on a novel corpus of over 25,000 articles, but struggles to derive context and meaning from them due to their long length.

Dataset

- 25,554-article dataset scraped from CNBC archives ranging from January 1, 2019 to July 22, 2019
- The data was tokenized, truncated/padded to fit a fixed length (1300 tokens for documents, and 150 tokens for summaries)
- Pretrained word embeddings applied to the data (GloVe Twitter 100)



Tokenized Document Example

<pad> (x897)
<start> president donald trump has canceled a planned trip to ireland in november a spokeswoman for the irish prime minister confirmed tuesday to cnbc <punct> <newline> <newline> the proposed visit of the us president is <unk> information services officer <unk> <unk> told cnbc adding the us side has cited <unk> reasons <punct> <newline> <newline> […] <eos>

Process + Results

Model Structure

- Encoder + decoder with 4 LSTM layers each to produce Seq2Seq model
- Attention layer used over input sequence to develop "context vector," which aids in selection of output token

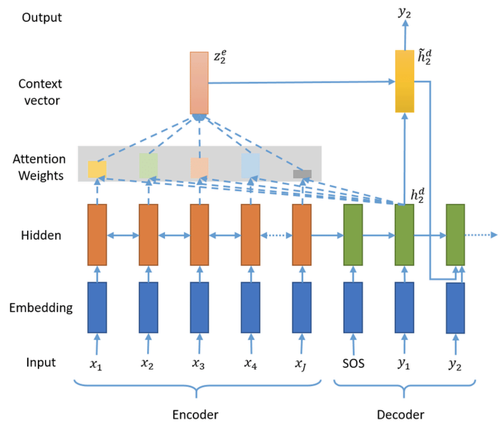


Image Source: Nallapati *et al.*[2]

Training Details

This model was trained over 2 epochs using the ADAM optimizer ($\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$). The loss metric was MSE. The final metrics are below:

Data Subset	Loss	Accuracy*
Training	**0.333**	**0.1878**
Test	**0.322**	**0.2032**
Validation	**0.322**	**0.2021**

*the importance of accuracy for NLP tasks is debated

Findings

The model struggled in identifying long-term dependencies in the source sequences and generated quite incoherent summaries of the input article. The output sequences also repeated many words. These results suggest the hidden state of the encoder LSTM and the token-level attention were insufficient context to produce an effective summary of the input sequence. However, this model significantly outperforms random guessing of words (the probability of guessing the correct word is approximately 1/2700)

Generated Summary Example

The input text discusses regulatory challenges for a company during its earnings call.

<start> regarding regarding from from details details details despite despite interest despite speculation resolutions success survival survival challenge in […] <eos>

Conclusion & Reference

Conclusion

The model's inability to identify context and dependencies in the source sequence suggest that a pure Sequence-to-Sequence model approach may be insufficient for abstractive summarization of longer texts with upwards of 500 tokens. More effective approaches to text summarization, as described by Shi *et al.*[1], include pointer-generator networks, where abstractive and extractive text summarization are combined to produce a more coherent and less repetitive summary. This type of model has achieved state-of-the-art results with input sequences of around 800 tokens. An alternative approach to achieve better results is to utilize a chunk-level attention layer[1], where chunks of tokens are also weighed into the context vector.

References

1. Shi T, Keneshloo Y, Ramakrishnan N, and Reddy CK. Neural abstractive text summarization with sequence-to-sequence models. arXiv preprint arXiv:1812.02303 2018; .
2. Nallapati R, Zhou B, Gulcehre C, and Xiang B. Abstractive text summarization using sequence-to-sequence rnns and beyond. arXiv preprint arXiv:1602.06023 2016; .
3. Sutskever I, Vinyals O, and Le QV. Sequence to sequence learning with neural networks. 2014; 3104-3112.

Acknowledgements/Contact

Aditya Rathod
aditya.rathod@utdallas.edu
https://adityar.me/

Acknowledgements

I would like to thank Dr. Anurag Nagar for his encouragement and helpful guidance throughout the Clark Summer Research Program. I would also like to thank Omeed Ashtiani and Seyyed Hosseini for their teaching and advice, and The University of Texas at Dallas for sponsoring this program. Finally, I would like to thank my family who supported my efforts and made participating in this program possible.