


Newsroom Article Summarization

Evan Calkins
Brian Dorsey

A dark blue diagonal gradient bar that starts from the bottom left and extends towards the top right, covering the lower half of the slide.

Agenda

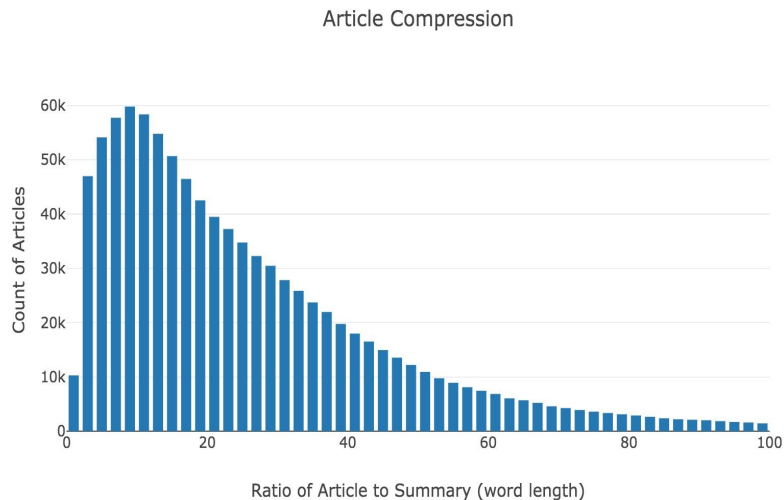
- Recap
- Data Prep
- Modeling & Experiments
- Results
- Conclusion

Recap

- Article summarization using seq2seq
- Cornell Newsroom dataset
- Abstractive vs extractive summarization
- Evaluation using custom ROUGE metric

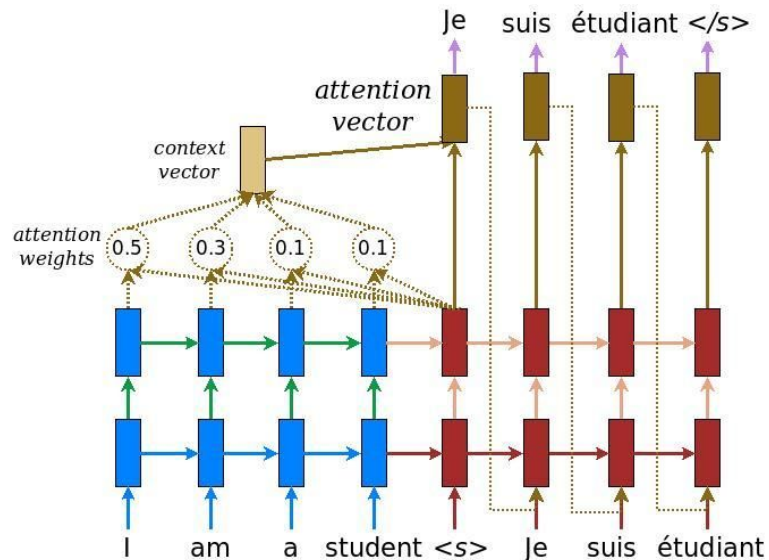
Data Preparation

- Data generated via web-scraping
 - Many extreme outliers
- Clipped data for modeling
 - Article length
 - Summary length
 - Compression
- Created vocab, encoded sentences as integers, and padded



Modeling & Experiments

- Word2Vec pre-trained embeddings - frozen
- Seq2Seq
 - Single GRU Layers in Encoder and Decoder
 - with Attention - **increased word extraction**
- Top-k Sampling - **decreased repetitive words**
 - Top 1000 words were roughly equal to 1%



Prediction

Actual: researchers have found dozens new genes that may play role in causing autism according two studies published in the medical journal nature

Predicted: researchers have found several new incurable that may play role in causing autism according two studies published tuesday in the medical journal underlying nature medical journal nature data kids likely stress reports cells that this kingdom p

Model Results

- Article length < 600 words
- Summary length > 16 words
- Vocabulary size: 10000 words
- Embedding sizes: 300
- Hidden size: 300
- Learning Rate: 0.05-0.001
- Teacher Forcing: 0.50-0.75

- Rouge Scores - **an extractive metric**
 - Higher Recall - predicted same words in actual summary
 - Lower Precision - predicted words not in actual summary
- Predicted summaries longer than actual summaries
- Language model undertrained - imperfect grammar and syntax

Conclusion

- Abstractive summarization is difficult - we accomplished more extractive summarization
- Future considerations
 - Reduce vocabulary size
 - Beam Search - Tries with probabilities
 - Train on more data and for more epochs
 - Train pre-trained embeddings further
 - Bi-directional encoder and decoder layers

Thank you

References

- <https://summari.es/>
- <https://aclweb.org/anthology/N18-1065>
- [https://en.wikipedia.org/wiki/ROUGE_\(metric\)](https://en.wikipedia.org/wiki/ROUGE_(metric))
- https://developingideas.me/wp-content/uploads/2017/09/Corbin.Albert_psxca1_4269843_Dissertation.pdf
- <https://danijar.com/tips-for-training-recurrent-neural-networks/>