COMP 551 - T1-05 Executive Summary

Justin Bell 260561261
Stuart Mashaal 260639962
Harley Wiltzer 260690006

April 20, 2018

The 2017 paper *A Deep Reinforced Model for Abstractive Summarization* by Paulus, Xiong, and Socher presents a novel attention method and Maximum-Likelihood/Reinforcement-Learning (RL) hybrid model for creating abstractive summarization of news articles. The authors also include some baselines in their analysis, some original and others borrowed from related works.

To obtain better baselines. we implemented and tuned Lead-N, SumBasic, and LSTM-RNN without Attention on the CNN/Daily Mail dataset used in the paper. We used ROUGE-1, and ROUGE-2 F1 scores to compare the results.

The simple Lead-N model, which extracts the first N sentences of an article, was significantly improved by finding the best value of N via model selection with the validation set. With a value of N=4, this naive model outperformed even the best results from the state-of-the-art models presented in the reference paper with a ROUGE-1 F1 score of 41.4%.

We also implemented and tuned an extended SumBasic model via model selection with the validation set. This model also showed surprisingly good results with a ROUGE-1 F1 score of 39.7%. Although this does not beat the score from the best Lead-N classifier, it is still considerably greater than many of the reported baseline scores from *A Deep Reinforced Model for Abstractive Summarization* Although this does not beat the score from the best Lead-N classifier, it is still considerably greater than many of the reported baseline scores from *A Deep Reinforced Model for Abstractive Summarization*.

The last baseline we included was an implementation of the Maximum Likelihood (ML) LSTM-RNN encoder-decoder without attention. We tried 4 experiments, each with a different set of hyperparameters, all of which were much worse than even the baselines shown in the original paper. We suspect this is due to limited training time and compromises we made to make the training faster. We support this claim with reference to the training-validation error charts.

The hyper-parameters that were considered for the encoder-decoder were vocabulary size and the teacher forcing ratio. We chose to compare a smaller vocabulary to check if we could get similar results with much shorter computational costs. We considered teacher forcing ratio in hopes that a higher ratio would correlate with faster convergence.

The article ultimately culminates in the observation that the simplest models reported the greatest scores for a minute fraction of the computational power, and discuss the implications of this result. Various suggestions for the reason behind these results are pondered and rationalized. The results are summarized in Table 1.

The full article is available for perusal at `http://cs.mcgill.ca/~hwiltz/literature/COMP551AbstractiveSummarization.pdf`.

Model	Rouge-1 (%)	Rouge-2 (%)
Lead-3 (Nallapati et al, 2017)	39.2	15.7
Lead-3	40.7	16.1
Lead-4	41.4	16.8
words-lvt2k-temp-att (Nallapati et al., 2016)	35.46	13.30
Extended SumBasic (sentences over tokens)	36.3	13.2
Extended SumBasic (tokens over sentences)	39.7	15.8
ML without intra-attention (Paulus et al., 2017)	37.86	14.69
ML without intra-attention (LSTM Encoder-Decoder)	17.4	1.12

Table 1: Results of various baseline models in comparison with the results reported in *A Deep Reinforced Model for Abstractive Summarization* (results highlighted in blue were yielded with the models proposed in this paper)