Summarization of News Articles using Pointer-Generator Networks

Anmol Sood
Indian Institute of Technology, Delhi
anmolsood1995@gmail.com

Mayank Sharan
Indian Institute of Technology, Delhi
myselfsharan@gmail.com

Abstract

Simple sequence-to-sequence models for abstractive summarization suffer from repetition and quite often produce factually incorrect details. To improve upon this, pointer-generator networks with coverage have been proposed. However, these models result in being mostly extractive and very rarely produce novel n-grams (phrases). A modification to the beam search scoring function has been proposed very recently to rectify this. We propose an improved beam search scoring function which results in better abstractive summaries.

1 Introduction

The task of summarization is to capture the gist of a large source text in a condensed form. There are to schools of thought on summarization : extractive and abstractive. Extractive methods focus efforts on putting together summaries using words, phrases and sentences from the original text. Abstractive methods are capable of generating fresh words and phrases not in the original text. Human generated summaries are more abstractive than extractive in nature. Abstractive summarization provides the power to paraphrase, change sentences, generalize etc.

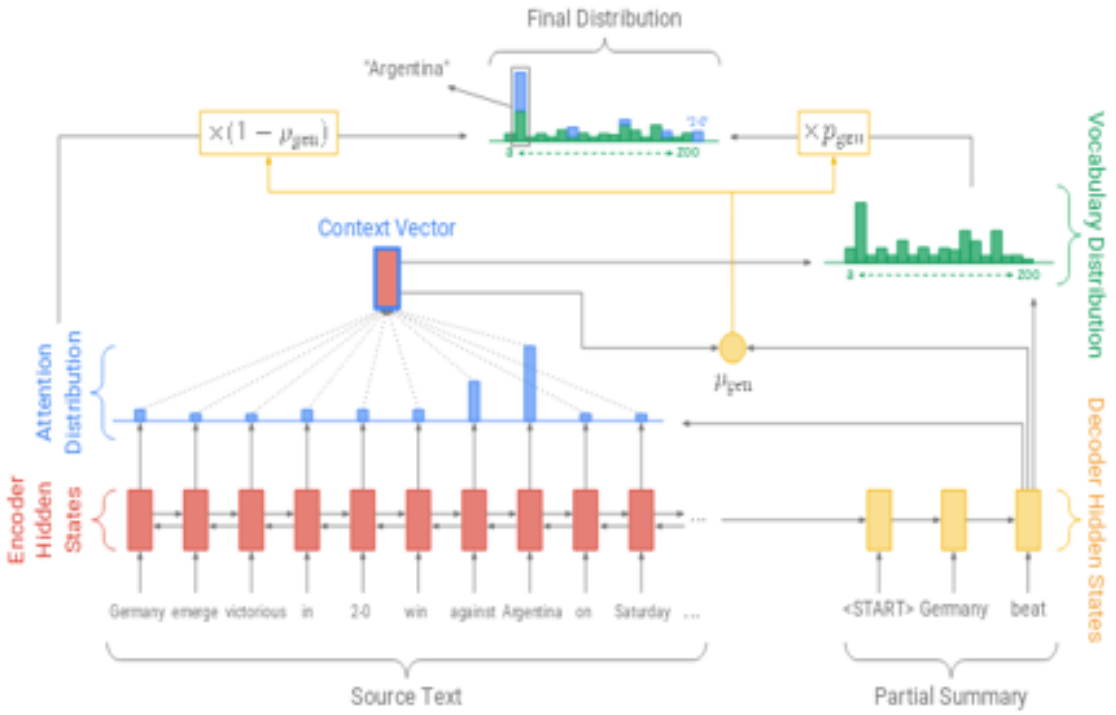
The comparative ease that comes with using the extractive approach has led to a majority of the previous work in summarization being extractive. Recently, with the advent and the success of sequence to sequence models there has been a lot of focus on abstractive summarization. Here we extend the work done on the state of the art to improve upon the abstractive nature of the summaries generated by the model.

2 Related Work

The work done heavily uses the work presented in (See et al., 2017). The approach they followed was to use a pointer generator network with a coverage model in an encoder decoder fashion to create a model that is both abstractive and extractive. The model here uses a probability to decide between generation and copying. Copying leading to extractive summary that builds on critical details that are present in the original text. The generation mechanism leads to fresh words and phrases in the summary contributing to it's abstractive nature. Even though the model is capable of generating abstractive summaries, during testing the model turns out to be heavily biased towards copying. The same is said in their own paper by (See et al., 2017).

The paper (Paulus et al., 2017) presented ideas regarding using reinforcement learning to train the encoder decoder model to try and tackle the problem of repetition. The attempt here is to try and tackle the problem of 'exposure bias' that happens in the situation of supervised learning like the one caused by teacher forcing that is used in the pointer generator network model proposed in (See et al., 2017). Even though this leads to improvements this model still faces the major shortcoming of being mostly extractive in the nature of it's generated summaries.

The paper (Weber et al., 2017) looks to fix the problem faced in the previous two papers. The idea here is to try and penalize when the algorithm copies words or phrases from the original text so that the the model learns to not do that. This will improve the abstractive nature of the summary generated by the model. Here this is done by modifying the beam search. The cost function used in the beam search is changed to penalize over usage of the copying mechanism which causes the sum-



/2/2

Figure 1: Pointer Generator Network with Coverage (See et al., 2017)

mary to be of extractive nature.

Across all three papers that are mentioned there is a very clear trend as to how to improve the abstractive nature of the summaries. This trend is to perform incremental improvements over the existing model showing good results. This is a similar direction in which the efforts we made were. Even subtle changes lead to much better performance.

3 Our Models

This section contains the details of the (1) baseline sequence to sequence model, (2) Pointer Generator with coverage (SoTA), (3) SoTA with CCD and (4) SoTA with CCD+

3.1 Sequence to Sequence model with Attention

The baseline model is a simple encoder decoder architecture with attention. The article is provided as tokenized input to the encoder, generating a series of hidden encoder states h_i. At each time step the decoder takes in the previous word as an input to generate the the decoder state of s_t. The attention is calculated as

$$e_i^t = v^T tanh(W_h h_i + W_s s_t + b_{attn})$$

$$a^t = softmax(e^t)$$

Here v, W_h,W_s and b_{attn} are parameters to be trained. Further use the attention to compute the context vector as a weighted sum of the encoder hidden states.

$$h_t^* = {}_i a_i^t h_i$$

3.2 Pointer Generator network with coverage

This model integrates a pointer network into the baseline model. This allows the additional option of allowing both the copying of words via pointing and the generation from the vocabulary. This is made possible by using the probability p_{gen} calculated as follows

$$p_{gen} = (w_{h^+}^T h_t^* + w_s^T s_t + w_x^T x_t + b_{ptr})$$

Further a coverage mechanism is put in place to make sure that repetition does not keep on happening in the generated summary.In coverage a coverage vector is maintained which is the sum of attention distributions over all previous decoder time steps. This coverage vector is used as an extra input to the attention calculations. This incorporates the ability to detect when the attention is being applied to the same point again and make sure that does not happen.

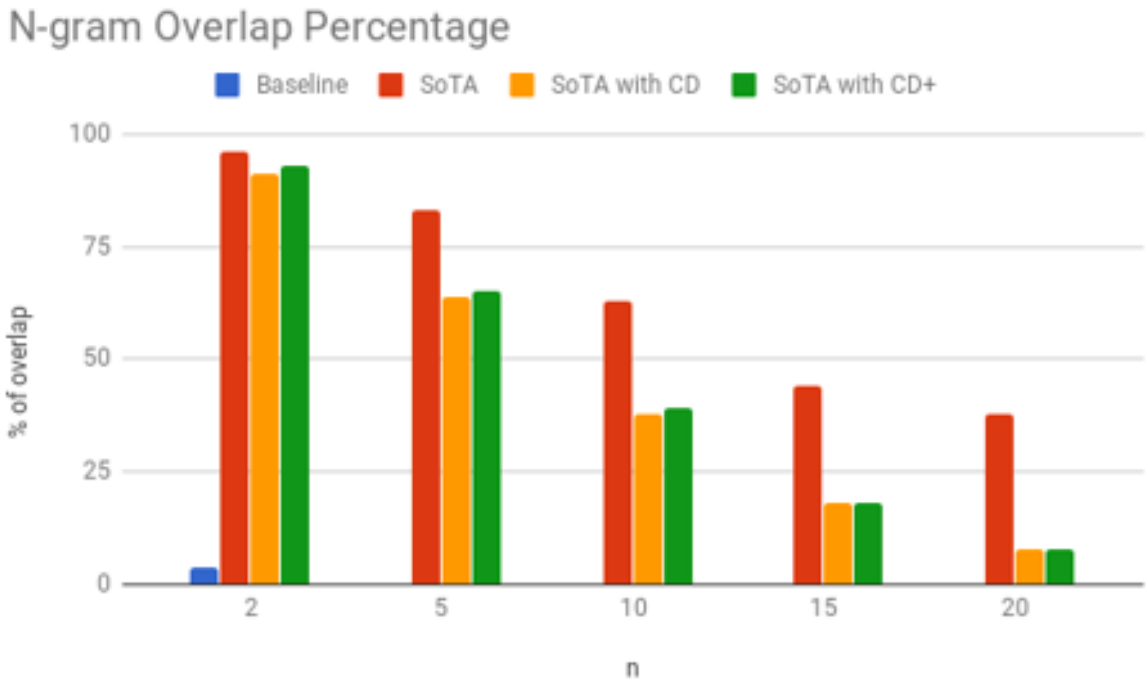


Figure 2: n gram overlap percentage for different models

3.3 State of the Art with CCD

The scoring function in the beam search scoring function is where the modification lies for this model. The modified beam search scoring function to be used here is

$$s(y_{\leq t}, X) = \sum_{t'=1}^{t'=t} log(p(y'_{\leq t'}, X)) - \eta_o * t' * max(0, p^*_{gen} - avg(p_{gen})) \quad (1)$$

3.4 State of the Art with CCD+

We propose a further modification in the beam search scoring function to incorporate the intuition that the first few words of the summary should not be forced to be abstractive. In fact it is better for the quality of the summary if they are extractive.The modified beam search scoring function to be used here is

$$s(y_{\leq t}, X) = \sum_{t'=1}^{t'=t} log(p(y'_{\leq t'}, X)) - max(\eta_o * t', 15) * max(0, p^*_{gen} - avg(p_{gen})) \quad (2)$$

4 Experiments

The CNN/Daily Mail news dataset with multi-sentence summaries was used in the experiments. The metrics used to evaluate the models were Rouge 1 F1 scores and n gram overlap percentage to compare novel phrase generation.

Model	ROUGE 1 F1 Score
Baseline	30.21
SoTA	38.76
SoTA with CCD	36.32
SoTA with CCD+	36.94

Table 1: Results comparison across the different models

5 Conclusion

In conclusion we see that the modification we do to the beam search scoring function not only leads to a 0.62 increase in the ROUGE 1 score metric but it also maintains comparability with the CCD model in terms of the n-gram overlap with the original text. So objectively this is a marginal improvement.

References

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304* .

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer generator networks. *arXiv preprint arXiv:1704.04368* .

Noah Weber, Leena Shekhar, Nilanjan Balasubramanian, and Kyunghyon Cho. 2017. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304* .