On Reducing Repetition in Abstractive Summarization

Pranav Ajit Nair
Indian Institute of Technology (BHU)
Varanasi
pranavajitnair.cse18@itbhu.ac.in

Anil Kumar Singh
Indian Institute of Technology (BHU)
Varanasi
aksingh.cse@iitbhu.ac.in

Abstract

Repetition in natural language generation reduces the informativeness of text and makes it less appealing. Various techniques have been proposed to alleviate it. In this work, we explore and propose techniques to reduce repetition in abstractive summarization. First, we explore the application of unlikelihood training and embedding matrix regularizers from previous work on language modeling to abstractive summarization. Next, we extend the coverage and temporal attention mechanisms to the token level to reduce repetition. In our experiments on the CNN/Daily Mail dataset, we observe that these techniques reduce the amount of repetition and increase the informativeness of the summaries, which we confirm via human evaluation.

1 Introduction

Abstractive summarization (Nallapati et al., 2016; See et al., 2017; Chopra et al., 2016; Rush et al., 2015) aims at capturing the essence of a document in a limited number of words. Despite great performance improvement over the years, many abstractive summarization models still suffer from the repetition problem wherein they tend to repeat the same phrase or ngram over and over again reducing the amount of information conveyed and also reducing the quality of the generated text. For example such a system may generate "He scored five goals. five goals. five goals ...". Such kind of repetition is highly uncommon in human written text.

A true understanding of the problem is still unknown. Some argue that it is because of model architecture (Vig, 2018; Holtzman et al., 2020), other suggest that it is because of the nature of sampling performed by the model which is very different from human natural language (Holtzman et al., 2020). Welleck et al. (2020) suggests that the cause of the problem is maximum likelihood training of language generation models. Holtzman et al. (2020) alleviates this problem to some extend via nucleus sampling and Fan et al. (2018) propose TopK sampling to solve the repetition problem.

To solve this repetition problem for abstractive summarization we adopt two approaches.

1. The unlikelihood training approach proposed by Welleck et al. (2020). Their approach has significantly reduced repetition on several language modeling datasets. We extend their approach to abstractive summarization. In our experiments we observe that the unlikelihood objective significantly reduces repetition and also boosts the model performance in terms of evaluation metric such as ROUGE. All our experiments are based on Transformers (Vaswani et al., 2017). Our model, on the CNN/DailyMail (Chen et al., 2016) dataset outperforms the baseline by a significant margin.

2. See et al. (2017) used the coverage mechanism (Tu et al., 2016) which penalizes the model if the decoder cross attention attends to the same source token multiples times. We propose to extend this idea at the token level where the model is penalized in a similar way if it gives high probability to the same token multiple times while generating the summary. In this way we ensure that model does not assign high probabilities to the same token over and over again. In our experiments we see that this additional regularization not only reduces repetition in the generated summaries but also improves the performance in terms of evaluation metric beating the baseline Transformer.

Our key contributions are:

- We qualitatively and quantitatively analyze repetition in abstractive summarization.

Proceedings of the Student Research Workshop associated with RANLP-2021, pages 126–134,
held online, Sep 1–3, 2021.
https://doi.org/10.26615/issn.2603-2821.2021_018

- We extend the unlikelihood objective for language modeling to abstractive summarization ans show that there is a significant reduction in repetition.

- We propose an extension of the coverage mechanism to the token level to reduce repetition ans show that our approach reduces repetition in abstractive summarization.

- We experiment with the embedding regularizer proposed by Gao et al. (2019) to increase the expressiveness of the token embeddings for transformers, which we suspect would allow the model to utilize rare tokens and help in repetition reduction.

- We also propose an extension of the temporal attention mechanism (Sankaran et al., 2016), which dampens the attention distribution by dividing it with the sum of the previous attention distributions to the token level. We conduct experiments on this approach to verify it's effectiveness on repetition reduction.

2 Related Work

Abstractive summarization aims at capturing the entire essence of a document within a relatively smaller number of words. There has been great progress made in the recent past where better and better architectures have been proposed to solve the problem. See et al. (2017) used pointer generator networks to copy words from the input document. Duan et al. (2019) augmented the Transformer architecture with a contrastive attention mechanism to ignore the irrelevant parts of the document. Zhang et al. (2020a) pretrained a Transformer model for summarization. Paulus et al. (2018) proposed to use rewards from policy gradient reinforcement learning to alleviate exposure bias. Gehrmann et al. (2018) developed a bottom-up copy attention mechanism to over-determine phrases in the document that should be included in the summary. There approach is build upon the pointer generator network (See et al., 2017). Our work unlike previous work tries to reduce the repetition in Transformer models via techniques that have been successfully applied to language modeling.

Several methods have been proposed to reduce repetition in language generation. Holtzman et al. (2020) use different sampling techniques to reduce repetition. Welleck et al. (2020) propose the unlikelihood objective to reduce repetition. Fu et al. (2021) propose to combine pairs of words that often follow each other to reduce repetition. Our work applies such techniques to abstractive summarization where repetition reduces the amount of information conveyed in the summary.

To reduce repetition in abstractive summarization Li et al. (2019) propose a reinforcement learning based approach. They utilize BERTScore (Zhang et al., 2020b) to reward the model for better generations and reduce repetition. Lin et al. (2018) propose a CNN based encoder to perform global encoding of the source side information. Their approach also reduces repetition. Chen and Bansal (2018) propose a reinforcement learning based sentence level neural abstractive summarization model. Unlike these approaches both our approaches explicitly try to reduce repetition without using any heuristic based decoding strategies.

3 Method

In this section we provide the training objectives used by us to reduce repetition. All our models and experiments are based on the standard Transformer encoder-decoder architecture. The input to the encoder is the input document x represented as $x = (x_1, x_2, ..., x_n)$, and the summary to be generated is $y = (y_1, y_2, ..., y_m)$. We can write the loss function obtained via maximum likelihood estimation as:

$$L_{MLE} = -\sum_{i=1}^{m} \log p_\theta(y_i|y_{1,...,i-1}, x) \quad (1)$$

where θ represents the model parameters.

3.1 Unlikelihood Objective

The idea is to decrease the probability of the tokens which have been already generated by the model in the summary to avoid frequent repetition. So for generation of i^{th} word the negative candidates would be $C_i = \{y_1, ..., y_{i-1}\} \setminus \{y_i\}$. The unlikelihood loss for the i^{th} token's generation can be written as:

$$L_{UN} = -\sum_{c \in C_i} \log(1 - p_\theta(c|y_{1,...,i-1}, x)) \quad (2)$$

Thus the total loss for the i^{th} token's generation

is:

$$L^i_{token} = -\log p_\theta(y_i|y_{1,...,i-1}, x) - \gamma \sum_{c \in C_i} \log(1 - p_\theta(c|y_{1,...,i-1}, x) \quad (3)$$

We call this objective UOMLE. In all our experiments we use $\gamma = 1$.

3.2 Embedding Regularizer

Gao et al. (2019) propose to minimize the cosine similarity between token embeddings for Transformer models to solve the representation degeneration problem where the word embeddings tend to degenerate over the course of training and form a narrow cone. Since all our experiments are based on the Transformer model we add this regularization term into the model to check for any improvement in the quality of the generated summaries. We suspect that increasing the expressiveness of the embeddings would allow the model to better generate rare words, which in turn would reduce repetition. The regularization loss added is:

$$L_W = \frac{1}{N^2} \sum_{i}^{N} \sum_{i \neq j}^{N} \hat{w_i}^T \hat{w_j} \quad (4)$$

where $\hat{w_i} = \frac{w_i}{\|w_i\|}$ $w_i \in \mathbb{R}^d$, and $W = [w_1, w_2, ..., w_N]^T$ is the embedding matrix. d is the embedding dimension and N is the vocabulary size. We call this objective EmbedReg.

3.3 Coverage Regularizer

Here the model is penalized if it gives high probability to the same token multiple times while generating the summary. The loss function added is an extension of the coverage mechanism used in See et al. (2017). Defining the context vector at time step t as:

$$c_t = \sum_{i=1}^{t-1} p_\theta(\cdot|y_{1,...,i-1}, x) \quad (5)$$

Let $p_t = p_\theta(\cdot|y_{1,...,t-1}, x)$. The coverage loss is defined as:

$$L_t = \sum_{j=1}^{N} \min(c_t^j, p_t^j) \quad (6)$$

This coverage loss can be added to the model loss function without any other change to the model making it's implementation fairly straightforward. Thus if the model gives a higher probability to a token which has been previously given a higher probability the loss incurred is higher. We refer to it as CovReg.

3.4 Temporal Mechanism

Sankaran et al. (2016) proposed the temporal attention mechanism. In this approach, each attention distribution is divided by the sum of the previous, which effectively dampens repeated attention. We extend this approach to the token level where the distribution over the vocabulary at every time step is divided by the sum of the vocabulary distributions from the previous time steps. We suspect that this would reduce the probability of generating the same token over and over again since a higher probability of the token in the previous time steps would incur a greater dampening effect. In our experiments we refer to it as TemporalMech.

4 Experiments

We use the CNN/Daily Mail dataset for all our experiments. All our models are Transformer based, we use the Transformer base model with 8 encoder layers, 8 decoder layers, 8 attention heads, 256 is the hidden and embedding dimension and 512 dimensional feed forward layer. We use the linear warm-up followed by square root decay schedule proposed in Vaswani et al. (2017). We train our baseline for 200K iterations with a batch size of 32 and we employ early stopping based on validation loss. We use beam search decoding in all our experiments with a beam size of 4, we also employ a length penalty of 1.0. We use label smoothing with $\alpha = 0.1$. We use Sentencepiece tokenizer in all our experiments. All the experiments are performed using PyTorch on a single NVIDIA Tesla V100 GPU.

4.1 Evaluation metrics

As a token level metric we use fraction of the net token predicted that has previously occurred in a window of l tokens :

$$rep/l = \frac{1}{|S|} \sum_{i=1}^{|S|} \sum_{t=1}^{|y^{(i)}|} \mathbb{I}[\text{argmax} p_\theta(y_t|y_{1,...,t-1}, x) \in y_{t-l-1:t-1}] \quad (7)$$

Where $\mathbb{I}$ is the indicator function S in the validation/test data and y in the generated summary. For a single token repeat $\mathbb{I}$ is 1.

Model variant	ROUGE-1	ROUGE-2	ROUGE-L	rep/4	seq-rep-4
Baseline	21.55	7.27	23.74	0.3751	0.2336
CovReg	**21.94**	**7.50**	**24.15**	**0.3618**	**0.2204**
TemporalMech	20.50	7.53	22.92	0.3978	0.2413
EmbedReg	20.97	7.31	24.80	0.3757	0.2212
UOMLE	**25.47**	**8.34**	**25.70**	**0.2241**	**0.0966**

Table 1: Results of our final models on the validation set.

Model variant	ROUGE-1	ROUGE-2	ROUGE-L	rep/4	seq-rep-4
Baseline	21.25	7.04	23.55	0.3785	0.2420
CovReg	**21.75**	**7.44**	**24.02**	**0.3634**	**0.2208**
TemporalMech	20.36	7.08	21.34	0.3983	0.2431
EmbedReg	20.81	7.11	23.78	0.3779	0.2214
UOMLE	**25.39**	**8.13**	**25.43**	**0.2255**	**0.0942**

Table 2: Results of our final models on the test set.

A similar sequence level metric to measure repetition will be:

$$\text{seq-rep-n} = 1.0 - \frac{|\text{unique n-grams in y}|}{|\text{n-grams}|} \quad (8)$$

This metric calculates the fraction of unique n-grams in the generated summary. Both the above metrics have been proposed in Welleck et al. (2020).

4.2 Results

The results on the validation set can be found in Table 1 for all the four approaches. From the results we can see that all approaches except the temporal mechanism reduce repetition and increase the the ROUGE scores. An increase in ROUGE scores can be attributed to the fact that reducing repetition allows the model to generate more informative summaries. For the repetition metrics we report rep/4 and seq-rep-4.

The Unlikelihood objective significantly reduces repetition and also increase ROUGE scores, since UOLME explicitly decreases the probability of previously generated words. A reduction in repetition allows the model to generate better summaries which in turn increases the ROUGE scores. The Embedding regularizer method also reduces seq-rep-4. There is no decrease in rep/4 probably because EmbedReg objective does not try to explicitly or implicitly penalize the model for repetition.

From Table 1 one can see that CovReg beats the baseline on the ROUGE metric and reduces repetition as well. Since we penalized the model for giving a high probability to the same token over and over again the model tends to repeat less generating better summaries, which is evident from the ROUGE scores. With TempMech, we see that repetition in the model has increased, also the performance in terms of ROUGE metric drops slightly below that of the baseline. This drop in performance and increase in repetition may be because of the distortion of the probability distribution over the vocabulary as pointed out by See et al. (2017).

We also test a combination of these objectives wherein we combine the EmbedReg method with the UOMLE and CovReg objectives. We find no significant decrease in repetition. Results on the test set for all our models are presented in Table 2.

4.3 Qualitative Analysis

To analyze the generated summaries qualitatively we present summaries generated by UOMLE objective in Table 3. In the first example the baseline model unlike the model trained with UOMLE objective tends to repeat 'second-half' multiple times. In the second example the phrases 'jaw' and '29-year-old' are repeated several times in the baseline, whereas the model trained with UOMLE shows no such repetition, and in the third example the word 'middle' is repeated several times by the baseline. This clearly shows that UOMLE reduces repetition. We also present samples generated with the CovReg objective in Table 4. Here, in the first example the baseline repeats the phrase 'best footballer' several times in the summary, whereas the model trained with CovReg shows no such repetition. Similarly in the second example the baseline repeats 'year' multiple times and in the third ex-

Article: when defenders who have not scored for two years find the net, watford might begin to think this can be the year they return to the premier league. matthew connolly has only been at the club for a month but struck a goal that proved key at a stadium that has humbled all three of their nearest promotion rivals. watford were made to sweat for the three points during a late surge by 10-man nottingham forest, but left having leapfrogged middlesbrough into third, behind norwich in second on goal difference. watford forward odion ighalo celebrates having given the hornets the lead after just four minutes . defender matthew connolly celebrates after putting watford 2-0 up against nottingham forest on wednesday . (Truncated) **UOMLE**: matthew connolly' s first-half at the premier league cup of 2014 world cup in his last 16 clash with watford on wednesday . it was a late first-minute ten-minute treatment to be the premier league two-half premier league season after half- macsbury . matthew connolly' s first-half minutes against nottingham forest to make a late on wednesday . **Baseline**: matthew connolly' s first-half-rated to four- cent-minute goal . don' first-half minutes of his **second-half from second-half** from a point to four minutes from **second half** half-ry lead for the **second-half** from the **second-half** .
Article: a four-year-old boy has returned home after being mauled by their family' s newly adopted pit bull. gavin tobeck of lacey, washington suffered from a broken cheekbone, jaw and nose bridge, and needed sections of a rib to repair his eye sockets after the attack by their dog smash on april 1. the boy also lost five baby teeth and three permanent tooth buds. in an effort to save her son, 29-year-old alissa evans' friend threw hot coffee on the dog which led to the animal releasing his grip on gavin . (Truncated) **UOMLE**: a four-year-old boy has returned to his eye sockets after being mauled by their family' s newly adopted pit bull . the young boy also lost five-year old to repair his eye sockets after being mauled by their dog in a 11th permanent tooth bud . **Baseline**: gavin to repair his eye socket, **jaw, jaw, jaw, jaw, jaw, jaw** and nose bridge, **29-year-old** a normal life and nose bridge, **29-year-old** a drug at harborview medical center in a broken cheekbone' s newly adopted pit bull smash on the animal releasing his eye sockets after spending six days at harborview medical center, **29-year-old** a broken cheekbone, **29-year-old** a broken cheekbone, **29-year-old** a normal life .
Article: a pensioner has broke down in tears in court as he was spared jailed after admitting he stabbed his sick wife to death after she begged him to ' bring her suffering to an end' . anthony mann, 78, sat quietly sobbing in the dock as he listened to the details of the mercy killing of his wife janet at their home in nuneaton, warwickshire. he was given a two year suspended sentence after he appeared at leamington justice centre today. anthony mann was spared jail for killing his wife janet after she asked him to end his life . (Truncated) **UOMLE**: anthony mann was given a three-month suspended jail term for killing his sick wife . he was given a two-year community community service by the suicide of his life in nunea . **Baseline**: anthony mann, 78, sat quietly sobbing in nuneaton, sat quietly sobbing in a kitchen knife, sat quietly sobbing in the **middle of the middle of the middle of the middle of the middle of the middle of the middle of the middle of the middle of the middle of the middle of the middle of the middle of the middle** of the year .

Table 3: Summaries generated by UOMLE from test set. Repetitions made by the baseline are highlighted.

Article: real madrid striker karim benzema has boldly stated he can win the ballon d' or award during his career. the france international is confident his club success at the bernabeu can force his way into contention for the prestigious award. however benzema' s los blancos team-mate cristiano ronaldo has picked up the last two accolades for the world' s best footballer whereas barcelona talisman lionel messi prevailed between 2009-12. french and real madrid forward karim benzema believes he can win the ballon d' or award in his career . despite the duo' s dominance, benzema is adamant he can compete for the title if he continues to perform for real madrid and at international level . (Truncated) **CovReg**: ma karim benzema has boldlyly stated he can win the ballon d' or award for real madrid and real' s best footballer whereas barcelona talisman lionel messi . cristiano ronaldo' s team-mate cristiano ronaldo . **Baseline**: karim benzema has picked up the ballon d' or award for the ballon d' or award during the ballon d' or award during the last two accolades team-mate cristiano ronaldo' s **best footballer** whereas barcelona talisman lionel messi' s **best footballer** whereas barcelona' s **best footballer** whereas barcelona talisman lionel lee whereas' s **best footballer' s best footballer' s** first-time .
Article: louis van gaal will continue with his big-spending ' galacticos' transfer policy this summer as manchester united prepare to go big in their search of title glory. and the dutchman will focus his attentions on landing proven international talent, rather than high-risk, big-money youngsters, that have become a feature of the club' s transfer policy in recent seasons. united spent close to £150million on the likes of radamel falcao and angel di maria last summer - deals that have not gone as well as expected. manchester united boss louis van gaal is planning to spend big money on proven talent this summer . united spent close to £150m on the likes of radamel falcao (above) - one deal that has not gone as expected . (Truncated) **CovReg**: louis van gaal' s side have not been a feature since on landing proven international-spending ' transfer policy in recent seasons as well as manchester united boss louis van gaal and angel di maria last summer' s transfer policy this summer . **Baseline**: louis van gaal' s new manchester united boss louis van gaal will focus on landing proven international talent on landing proven international talent on landing proven international talent this summer . the four-and-risk-plus transfer policy this summer . the four-at-a-half-week-half-**year-year-year-year-year-year–year-year**-old will also play paul recommendations on a free-**year**-old to make a new club .
Article: when the first astronauts land on mars, they will not use a conventional parachute or heat shield that has been used before. instead, upon impacting the upper martian atmosphere, a large inflatable saucer-shaped structure will slow their progress. this is the low density supersonic decelerator (ldsd) and, in june, nasa will perform the latest test of this groundbreaking technology - a vital next step in the journey to mars. scroll down for a video from last year' s test . in june, the vehicle will be sent into near-space from the navy' s pacific missile range facility on kauai, hawaii, to test its re-entry capabilities into earth' s atmosphere . (Truncated) **CovReg**: a large inflatable saucer-shaped structure will be sent into earth' s pacific missile range on june-space saucer-space saucer-entry capabilitiessd on januaryam-entry capabilities to a vital next-shaped structure will be sent into earth' s pacific missile range . **Baseline**: a large inflatable **saucer**-shaped inflatable **saucer**-shaped structure will perform a large inflatable **saucer**-entry capabilities into near-entry capabilities into near-shaped inflatable **saucer**-shaped inflatable **saucer**-shaped inflatable **saucer**-entry capabilities into near-shaped inflatable **saucer**-shaped inflatable **saucer**-shaped inflatable **saucer**-shaped inflatable **saucer**-entry capabilities in june .

Table 4: Summaries generated by CovReg from validation set. Repetitions made by the baseline are highlighted.

ample it repeats 'saucer' multiple times. These examples clearly show that the CovReg objective reduces repetition.

4.4 Human Evaluation Results

We randomly sampled 50 examples from the test set for human evaluation. All evaluators were engineering undergraduates with professional working proficiency in English. Each evaluator was presented with the original document, the summary generated by the baseline and the summary generated by the UOMLE objective model, and each summary was evaluated by exactly one evaluator. 40 evaluators were involved in the process. We asked the evaluators to rate the summaries based on three factors: 1) Essence of the document captured by the summary 2) Repetition in the summary and 3) overall quality (i.e linguistic quality) and informativeness of the summary (i.e how well were the more informative parts of the document captured). The evaluators were asked to pick the better summary with respect to each evaluation criteria. In 74% of the samples human evaluators find the summaries generated by the UOMLE objective to better capture the essence of the document. In 90% of the cases human evaluators find the summaries generated by the UOMLE objective to be less repetitive. In 76% of the cases human evaluators find the overall quality and informativeness to better in the summaries generated by UOMLE objective. Thus as per human evaluation UOMLE objective not only reduces repetition but also increases informativeness of the summaries.

We repeated the same experiment with our CovReg objective model. In 44% of the samples human evaluators find the summaries generated by the CovReg objective to better capture the essence of the document, whereas only in 40% of the samples they find the base model to be better. In 36% of the cases human evaluators find the summaries generated by the CovReg objective to be less repetitive , whereas only in 32% of the samples they find the base model to be less repetitive. In 46% of the cases human evaluators find the overall quality and informativeness to better in the summaries generated by CovReg objective, whereas only in 40% of the cases they find that the base model has generated better summaries. Thus we can conclude that the CovReg objective indeed reduces repetition and increases the overall quality and informativeness of the generated summaries.

5 Conclusion

In this work we present various techniques to reduce repetitiveness in abstractive summarization. We utilize the Unlikelihood objective and the Embedding regularizer from previous work to reduce repetition in abstractive summarization. We also propose extensions of the coverage mechanism and the temporal attention mechanism to reduce repetition. Our results show that these techniques not only reduce repetition in summaries but also increase performance in terms of ROUGE scores. Our human evaluation results confirm that the summaries generated via our techniques are more informative and have less repetition. The application of our methods on state-of-the-art summarization systems should further improve the results. In future work we would devise newer and better methods to reduce repetition and extend our methods to other sequence to sequence tasks such as machine translation.

Acknowledgements

We would like to thank the anonymous reviewers for their insightful comments and suggestions. The support and the resources provided by the PARAM Shivay Facility under the National Supercomputing Mission, Government of India at the Indian Institute of Technology (BHU), Varanasi, are gratefully acknowledged. We also thank Manav Jain for helping with human evaluation and Samyak Jain for comments on early drafts of the paper.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A thorough examination of the CNN/Daily Mail reading comprehension task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2358–2367, Berlin, Germany. Association for Computational Linguistics.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–686, Mel-

bourne, Australia. Association for Computational Linguistics.

Sumit Chopra, Michael Auli, and Alexander M Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98.

Xiangyu Duan, Hongfei Yu, Mingming Yin, Min Zhang, Weihua Luo, and Yue Zhang. 2019. Contrastive attention mechanism for abstractive sentence summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Zihao Fu, Wai Lam, Anthony Man-Cho So, and Bei Shi. 2021. A theoretical analysis of the repetition problem in text generation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2019. Representation degeneration problem in training natural language generation models. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Siyao Li, Deren Lei, Pengda Qin, and William Yang Wang. 2019. Deep reinforcement learning with distributional semantic rewards for abstractive summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6038–6044, Hong Kong, China. Association for Computational Linguistics.

Junyang Lin, Xu Sun, Shuming Ma, and Qi Su. 2018. Global encoding for abstractive summarization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 163–169, Melbourne, Australia. Association for Computational Linguistics.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *International Conference on Learning Representations*.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Baskaran Sankaran, Haitao Mi, Yaser Al-Onaizan, and Abe Ittycheriah. 2016. Temporal attention model for neural machine translation. *CoRR*, abs/1608.02927.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–85, Berlin, Germany. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Jesse Vig. 2018. Deconstructing bert: Distilling 6 patterns from 100 million parameters. *Medium, December*.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *International Conference on Learning Representations*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020a. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.