Improving Neural Abstractive Summarization via Reinforcement Learning with BERTScore

Yuhui Zhang, Ruocheng Wang, Zhengping Zhou*
Department of Computer Science
Stanford University
[yuhuiz, rcwang, zpzhou]@stanford.edu

1 Introduction

Abstractive summarization aims to paraphrase long text with a short summary. While it is a common practice to train encoder-decoder models in an end-to-end supervised learning fashion to maximize the log-likelihood objective, Paulus et al. (2018) introduce a method which utilizes reinforcement learning to directly maximize the non-differentiable ROUGE score, enhancing the model's performance.

ROUGE score (Lin, 2004) is the most widely-used evaluation metric for summarization, which simply evaluates the summarization quality by computing n-gram overlap between generated summaries and reference summaries. While ROUGE score is simple and easy to compute, there is a gap between the benchmark and the actual summarization performance. As ROUGE score only measures token hard-match, in some cases the ROUGE score will penalize two sentences conveying exactly the same semantic information, but highly reward sentences with completely different semantics yet in similar surface forms.

BERTScore (Zhang et al., 2019) is a recently proposed evaluation metric. Similar to ROUGE score, it computes a similarity score for each token in the generated summaries with each token in the reference summaries. However, by computing token similarity using contextualized word embeddings provided by BERT (Devlin et al., 2019), BERTScore successfully incorporates semantic information behind sentences, thus can provide better evaluations for cases where ROUGE score fails to account for meaning-preserving lexical and semantic diversity.

As BERTScore is demonstrated to have better correlations with human judgments for natural language generation (Table 1), can we use reinforcement learning with BERTScore to improve neural abstractive summarization? In this work, we answer this question by comparing different training schemas, such as supervised learning, reinforcement learning with ROUGE score (**RL-ROUGE**) and reinforcement learning with BERTScore (**RL-BERTScore**). We choose the strongly-performed pointer-generator networks with coverage loss (See et al., 2017) as our baseline model, and use **CNN/Daily Mail** corpus (Hermann et al., 2015) for training and evaluations.

Our experiments demonstrate that RL-BERTScore improves both ROUGE score and BERTScore, whereas RL-ROUGE only improves ROUGE score. We manually compare 50 summaries generated by RL-BERTScore and RL-ROUGE with regard to their fluency, redundancy, and overall quality. We find RL-BERTScore achieves much better quality than RL-ROUGE for human evaluation. Our analysis and case studies further explain the success of BERTScore as a reward function. We hope our work could cast light on the design of better RL reward for natural language generation.

Metric	Spearman ρ	Pearson ρ
ROUGE-L	14.54	14.51
BERTScore	21.26	22.75

Table 1: Correlations of ROUGE-L and BERTScore with human evaluations. Scores are computed on the human-annotated summarization dataset (Chaganty et al., 2018).

*Code available at *https://drive.google.com/open?id=1VJBbjKSh3Hd4AgDT1HMqDrj1HoVomU3e*

CS229: Machine Learning, Fall 2019, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

2 Method

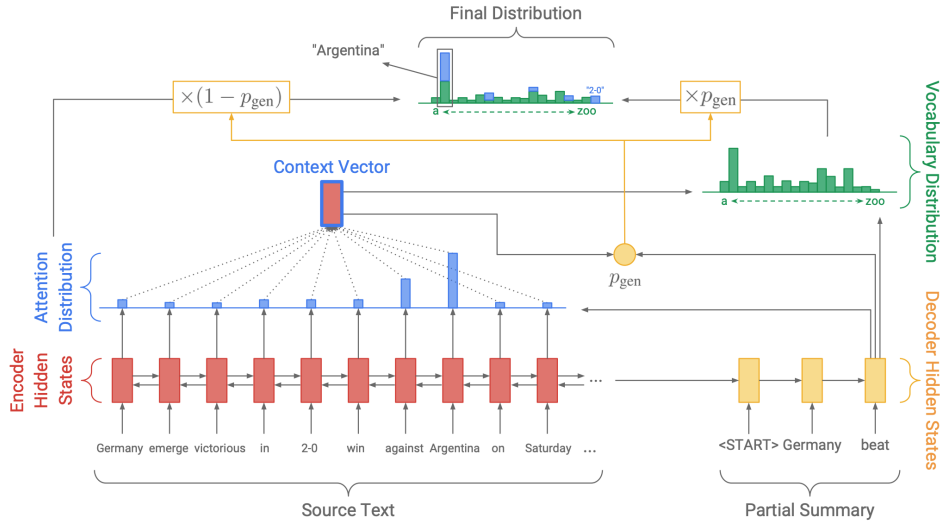

Figure 1: Pointer-Generator Networks (Baseline).

2.1 Baseline

Summarization can be formulated as generating sequence of tokens $\mathbf{y} = (y_1, y_2, ..., y_d)$ from the source sequence $\mathbf{x} = (x_1, x_2, ..., x_n)$. The generated sequence should have length $d \ll n$ while preserves most of the information in the source sequence. To probe into whether reinforcement learning with BERTScore can boost the performance, we choose the state-of-the-art **Pointer-Generator Networks (PGN)** (See et al., 2017) as our baseline (Figure 1).

The basic idea of PGN is a sequence-to-sequence conditional language model with copy mechanism and coverage penalty. In the training phase, we try to minimize the negative log-likelihood loss:

$$\mathcal{L}_{\text{nll}}(\theta) = -\sum_{t=1}^{d^*} \log p(y_t^* | y_{t-1}^*, \ldots, y_1^*, x_n, \ldots, x_1; \theta)$$

where y^* is the reference summary (ground truth).

2.2 Reinforcement Learning

Modeling NLL loss assumes that reference summary is given during the training phase, which is unknown in the inference stage, leading to the issue of so-called "exposure bias". Paulus et al. (2018) takes a reinforcement learning approach which tries to minimize the negative expected reward:

$$\mathcal{L}(\theta) = -\mathbb{E}_{y^s \sim p_\theta}[r(y^s)] \approx -r(y^s), \quad y^s \sim p_\theta$$

where the reward function $r(\cdot)$ is the ROUGE score between the randomly sampled generated summary y^s and the reference summary y^*. We can compute the corresponding gradient using REINFORCE algorithm (Williams, 1992):

$$\nabla_\theta \mathcal{L}(\theta) = -\mathbb{E}_{y^s \sim p_\theta}[r(y^s) \nabla_\theta \log p_\theta(y^s)] \approx -r(y^s) \nabla_\theta \log p_\theta(y^s), \quad y^s \sim p_\theta$$

To reduce the variance, we can subtract a baseline to the gradient estimation:

$$\nabla_\theta \mathcal{L}(\theta) \approx -(r(y^s) - r(\hat{y})) \nabla_\theta \log p_\theta(y^s)$$

where $\hat{y}$ is the summary generated by the model under the greedy decoding strategy. Finally, we optimize the reinforcement learning objective as follows:

$$\mathcal{L}_{\text{rl}}(\theta) = (r(\hat{y}) - r(y^s)) \sum_{t=1}^{d^s} \log p(y_t^s | y_{t-1}^s, \ldots, y_1^s, x_n, \ldots, x_1; \theta)$$

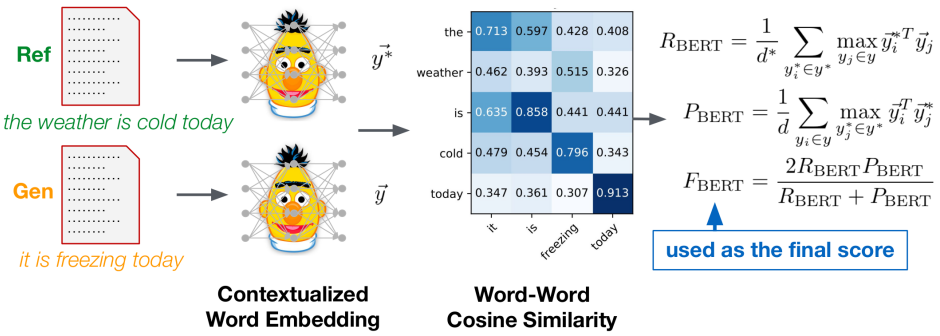

Figure 2: BERTScore computational process.

2.3 From ROUGE to BERTScore

For the generated summary y and reference summary y^*, BERTScore (Zhang et al., 2019) firstly utilizes the BERT (Devlin et al., 2019) model to generate contextualized word embeddings, then normalize these embeddings to compute the cosine similarity between tokens. Finally the BERTScore can be computed as (Figure 2):

$$R_{\text{BERT}} = \frac{1}{d^*} \sum_{y_i^* \in y^*} \max_{y_j \in y} \vec{y}_i^{*T} \vec{y}_j, \quad P_{\text{BERT}} = \frac{1}{d} \sum_{y_i \in y} \max_{y_j^* \in y^*} \vec{y}_i^{T} \vec{y}_j^*, \quad F_{\text{BERT}} = \frac{2R_{\text{BERT}} P_{\text{BERT}}}{R_{\text{BERT}} + P_{\text{BERT}}}$$

As BERTScore successfully incorporates semantic information behind sentences, it can provide better evaluations for cases. We will use F_{BERT} as the reward function for the reinforcement learning and compare the performance with the ROUGE case.

2.4 Mixed Training

We try to combine supervised learning and reinforcement learning objectives in our training schema:

$$\mathcal{L}(\theta) = (1 - \gamma)\mathcal{L}_{\text{nll}}(\theta) + \gamma \mathcal{L}_{\text{rl}}(\theta)$$

where γ is a hyperparameter to control the strength of reinforcement learning. In our experiments, we use ROUGE and BERTScore as the reward function, and we expect BERTScore to provide better reward, thus improving the performance of reinforcement learning.

3 Experiments

3.1 Dataset

CNN/Daily Mail (CNN/DM) is the most commonly-used corpora for neural abstractive summarization. We preprocess the dataset following the settings of Paulus et al. (2018). Data statistics are listed in Table 2.

Data Split			Average #Tokens	
Train	Dev	Test	Source	Reference
287K	13K	11K	384.0	61.3

Table 2: CNN/DM dataset statistics. Data split and average number of tokens are reported.

3.2 Experimental Setup

For training, we use 1-layer BiLSTM (hidden size 512) as the encoder and 1-layer LSTM as the decoder (hidden size 512). We pretrain the model using only NLL loss for 200K iterations with batch size 32, and use the Adagrad optimizer with initial learning rate 0.15. We select the best pretrained model based on validation perplextiy, and then perform the mixed training with both NLL and RL loss with $\gamma = 0.9984$ for another 20K iterations with batch size 12. We select the best final model based on validation reward. For inference, we use beam search with beam size 10, minimum length 35, and maximum length 100. For BERTScore computation, we use BERT-base to generate contextualized word embedding.

3.3 Quantitative Evaluation

We show the model performances in Table 3: **Baseline** refers to the Pointer-Generator Network (PGN) without RL (setting the γ to 0), **RL-ROUGE** refers to PGN with RL on ROUGE-L reward, and **RL-BERTScore** refers to PGN with RL on BERTScore reward. Note that although RL-ROUGE improves ROUGE scores, BERTScore actually decreases; In contrast, RL-BERTScore boosts both ROUGE scores and BERTScore.

Model	ROUGE Score			BERTScore
	R-1	R-2	R-L	
Baseline	39.37	17.15	34.67	60.77
RL-ROUGE	**43.28** ↑	**19.11** ↑	**38.55** ↑	59.26 ↓
RL-BERTScore	42.60 ↑	18.69 ↑	36.58 ↑	**62.77** ↑

Table 3: Model performances on the CNN/DM dataset.

3.4 Human Evaluation

As neither ROUGE nor BERTScore is a gold metric for the actual summarization quality, we can not determine which model is better according to Table 3. Therefore, we randomly sample 50 cases from the test set, and manually compare the fluency (*are words in the generated summary grammatically and accurately*), redundancy (*how much information is redundant in the generated summary*) and overall quality of summaries generated by RL-BERTScore and RL-ROUGE. Results in Table 4 indicates that our model, RL-BERTScore, achieves much better performance under human judgment, demonstrating BERTScore is a better reward for reinforcement learning of the abstractive summarization task.

Metric	Win	Tie	Loss
Fluency	**32%**	56%	12%
Redundancy	**62%**	18%	20%
Overall	**46%**	40%	14%

Table 4: Human evaluations for RL-BERTScore vs RL-ROUGE.

4 Analysis and Case Study

4.1 Correlation between BERTScore and ROUGE

We evaluate the correlation between ROUGE-L and BERTScore. For 2000 cases randomly sampled from the test set, we compute their ROUGE-L and BERTScore, which indicates a strong correlation between these two metrics (Figure 3). This helps explain why reinforcement learning with BERTScore also improves ROUGE score.

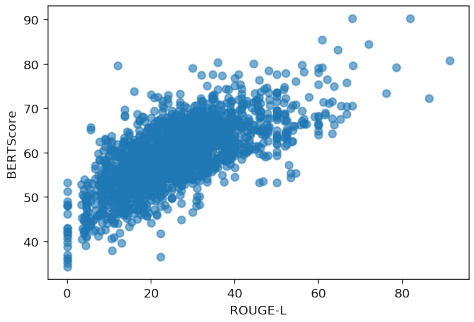

Figure 3: Correlation between BERTScore and ROUGE.

4.2 Case Study

BERTScore vs ROUGE In spite of the strong correlation between ROUGE-L and BERTScore, we find that in general, BERTScore better captures semantics between generated summaries and reference summaries, while ROUGE-L is easier to be "fooled" by the similarities of the surface forms. Table 5 lists some examples whose rankings by these

two metrics differ greatly. In contrast to ROUGE-L, BERTScore highly rewards the first two pairs sharing similar semantics but in diverse surface forms, and penalizes the last two cases that look similar yet with different meanings.

Reference	Generated	r_{BERT}	r_{ROUGE}
floyd mayweather will fight manny pacquiao in las vegas on may 2 . the bout is expected to generate $ 300 million in revenue . iyanna mayweather has been in training camp with her father floyd .	mayweather vs pacquiao will generate revenue upwards of $ 300 million in what is being billed as the most lucrative bout in boxing history . iyanna mayweather has been spending time in her father floyd 's training camp .	311	2585
renault have promised red bull that they will resolve power-unit problems . red bull owner dietrich mateschitz threatened to pull out of f1 . daniil kvyat 's red bull and max verstappen ' toro rosso both suffered a power-unit failure at chinese grand prix in shanghai .	renault managing director cyril abiteboul warned renault to solve their problems otherwise he would consider pulling his teams out of formula one . renault endured a chinese grand prix to forget as daniil kvyat 's red bull and the toro rosso of max verstappen both suffered a power-unit failure .	282	2841
helen dunn opened sunday mercury newspaper , saw raunchy pics . the 78-year-old was on cover of vintage lads mag , span , in 1955 . went on to marry manchester united pro footballer , alan dunn .	helen dunn , 78 , was flicking through the newspaper when she spotted herself on the cover of a raunchy lads ' mag from 1955 . the grandma , from stourbridge , west midlands , was only 17 years old at the time , perched seductively on a stool with her legs slightly crossed .	9529	3429
andy murray and kim sears married at dumblane cathedral on saturday . his mother judy murray has revealed that she ' ca n't wait to be a granny ' tennis coach already has plans to introduce her grandchildren to the sport .	judy murray has already wasted no time in revealing she simply ' ca n't wait to be a granny ' the tennis coach helped propel andy , 27 , and his brother , jamie , to sporting success , has a couple of caveats – he has to get the french open and wimbledon out of the way first .	7309	1509

Table 5: Examples of summaries ranked by ROUGE and BERTScore. r_{ROUGE} and r_{BERT} denote the ranking of the example in test set according to ROUGE and BERTScore, respectively.

RL-BERTScore vs RL-ROUGE We present two summaries generated by RL-BERTScore and RL-ROUGE in Table 6. Compared with the RL-BERTScore model, the RL-ROUGE model tends to output more repetitive sentences (e.g. the highlighted part in the first example) . At the same time, as shown in the second example, RL-ROUGE can generate incomplete sentences, which hampers the readability of the summary. And regarding the reference summary, RL-BERTScore is better at catching the key information of the source text.

Reference	RL-ROUGE	RL-BERTScore
kenyans gather in nairobi to remember victims of a terrorist attack that stunned a nation . the attack at a garissa university last week killed 147 people , mostly students .	the attack at a university in garissa on thursday killed 147 people . the al-shabaab militant group claimed responsibility . kenyan authorities have not released a list of the names of the victims . using the hashtag #147notjustanumber , kenyans used social media to talk about the lives of the victims . kenyan authorities have not released a list of the names of the victims.	kenyans remembered them as innocent victims of a terrorist attack that stunned a nation and left communities heartbroken . mourners read names of some of the victims as candles flickered in the dark . the attack at a university in garissa on thursday killed 147 people , mostly students . kenyan authorities have not released a list of the names of the victims .
the total eclipse lasted 4 minutes and 43 seconds . people west of the mississippi river had the best view in the u.s. parts of south america , india , china and russia were able to see the eclipse .	the third blood moon in a four-part series was the shortest eclipse of the bunch . a total lunar eclipse for nearly five minutes . nasa says will be the shortest such eclipse of the century . the " blood moon " will be the shortest such eclipse of the century . watchers in the eastern half of	the third blood moon in a four-part series was the shortest eclipse of the bunch , but still a sweet treat for early risers in north america . watchers in the eastern half of north america caught only a partial eclipse . parts of south america , india , china and russia were able to catch the total eclipse .

Table 6: Examples of summaries generated by RL-ROUGE and RL-BERTScore and corresponding reference summaries.

5 Conclusion and Future Work

In this work, we investigate the effectiveness of BERTScore, a recently proposed evaluation metric based on n-gram soft-match, as a novel reward function for reinforcement learning on the abstractive summarization task. We demonstrate its advantage over the most widely-used metric, ROUGE score, via both quantitative evaluation and human evaluation.

This work can be extended in the following directions: **1)** Hyperparameter tuning for better performance, which has not been exhaustively performed due to limited time and computing resources; **2)** An ensembled reward function as a linear combination of ROUGE, BLEU, BERTScore, etc. that better approximates human evaluation; **3)** Learning to learn: use linear regression on annotated data to learn the linear weights for 2).

Contributions Y.Z., R.W. and Z.Z. discussed the project and finished the poster and paper. Y.Z. preprocessed the dataset, implemented baseline codes, implemented and encapsulated scorers, trained the model, and performed hyperparameter search. R.W. implemented the BERTScore scorer, proposed correlation analysis, conducted human evaluations, and performed case studies. Z.Z. implemented the reinforcement learning algorithm on top of PGN with ROUGE-L and BERTScore.

References

Arun Chaganty, Stephen Mussmann, and Percy Liang. 2018. The price of debiasing automatic metrics in natural language evalaution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in neural information processing systems*, pages 1693–1701.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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

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)*, pages 1073–1083.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.