Deep Reinforcement Learning with Distributional Semantic Rewards for Abstractive Summarization

Siyao Li[1][*], **Deren Lei**[1][*], **Pengda Qin**[2], **William Yang Wang**[1]
[1]University of California, Santa Barbara
[2]Beijing University of Posts and Telecommunications
{siyaoli, derenlei}@ucsb.edu, qinpengda@bupt.edu.cn, william@cs.ucsb.edu

Abstract

Deep reinforcement learning (RL) has been a commonly-used strategy for the abstractive summarization task to address both the exposure bias and non-differentiable task issues. However, the conventional reward ROUGE-L simply looks for exact n-grams matches between candidates and annotated references, which inevitably makes the generated sentences repetitive and incoherent. In this paper, instead of ROUGE-L, we explore the practicability of utilizing the distributional semantics to measure the matching degrees. With distributional semantics, sentence-level evaluation can be obtained, and semantically-correct phrases can also be generated without being limited to the surface form of the reference sentences. Human judgments on Gigaword and CNN/Daily Mail datasets show that our proposed distributional semantics reward (DSR) has distinct superiority in capturing the lexical and compositional diversity of natural language.

1 Introduction

Abstractive summarization is a task of paraphrasing a long article with fewer words. Unlike extractive summarization, abstractive summaries can include tokens out of the article's vocabulary. There exists several encoder-decoder approaches such as attention-based architecture (Bahdanau et al., 2015; Rush et al., 2015; Nallapati et al., 2016; Chopra et al., 2016), incorporating graph techniques (Moawad and Aref, 2012; Ganesan et al., 2010), and adding pointer generator (Vinyals et al., 2015; Nallapati et al., 2016; See et al., 2017; Bello et al., 2017). However, long sentence generation suffers from exposure bias (Bahdanau et al., 2017) as the error accumulates during the decoding process.

Many innovative deep RL methods (Ranzato et al., 2016; Wu et al., 2016; Paulus et al., 2018; Lamb et al., 2016) are developed to alleviate this issue by providing sentence-level feedback after generating a complete sentence, in addition to optimal transport usage (Napoles et al., 2012). However, commonly used automatic evaluation metrics for generating sentence-level rewards count exact n-grams matches and are not robust to different words that share similar meanings since the semantic level reward is deficient.

Currently, many studies on contextualized word representations (Peters et al., 2018; Devlin et al., 2019) prove that they have a powerful capacity of reflecting distributional semantic. In this paper, we propose to use the distributional semantic reward to boost the RL-based abstractive summarization system. Moreover, we design several novel objective functions.

Experiment results show that they outperform the conventional objectives while increasing the sentence fluency. Our main contributions are three-fold:
• We are the first to introduce DSR to abstractive summarization and achieve better results than conventional rewards.
• Unlike ROUGE, our DSR does not rely on cross-entropy loss (XENT) to produce readable phrases. Thus, no exposure bias is introduced.
• DSR improves generated tokens' diversity and fluency while avoiding unnecessary repetitions.

2 Methodology

Background While sequence models are usually trained using XENT, they are typically evaluated at test time using discrete NLP metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005). Therefore, they suffer from both the exposure bias

---

*Equal contributions.

and non-differentiable task metric issues. To solve these problems, many resorts to deep RL with sequence to sequence model (Paulus et al., 2018; Ranzato et al., 2016; Ryang and Abekawa, 2012), where the learning agent interacts with a given environment. However, RL models have poor sample efficiency and lead to very slow convergence rate. Therefore, RL methods usually start from a pretrained policy, which is established by optimizing XENT at each word generation step.

$$L_{\text{XENT}} = -\sum_{t=1}^{n'} \log P(y_t | y_1, \ldots, y_{t-1}, x). \quad (1)$$

Then, during RL stage, the conventional way is to adopt self-critical strategy to fine-tune based on the target evaluation metric,

$$L_{\text{RL}} = \sum_{t=1}^{n'} \log P(\hat{y}_t | \hat{y}_1, \ldots, \hat{y}_{t-1}, x) \quad (2)$$
$$\times (r_{metric}(y^b) - r_{metric}(\hat{y}))$$

Distributional Semantic Reward During evaluating the quality of the generated sentences, ROUGE looks for exact matches between references and generations, which naturally overlooks the expression diversity of the natural language. In other words, it fails to capture the semantic relation between similar words. To solve this problem, distributional semantic representations are a practical way. Recent works on contextualized word representations, including ELMO (Peters et al., 2018), GPT (Radford et al., 2018), BERT (Devlin et al., 2019), prove that distributional semantics can be captured effectively. Based on that, a recent study, called BERTSCORE (Zhang et al., 2019), focuses on sentence-level generation evaluation by using pre-trained BERT contextualized embeddings to compute the similarity between two sentences as a weighted aggregation of cosine similarities between their tokens. It has a higher correlation with human evaluation on text generation tasks comparing to existing evaluation metrics.

In this paper, we introduce it as a DSR for deep RL. The BERTSCORE is defined as:

$$R_{\text{BERT}} = \frac{\sum_{y_i \in y} \text{idf}(y_i) \max_{\hat{y}_j \in \hat{y}} \mathbf{y_i}^\top \mathbf{\hat{y}_j}}{\sum_{y_i \in y} \text{idf}(y_i)} \quad (3)$$

$$P_{\text{BERT}} = \frac{\sum_{y_j \in y} \text{idf}(y_j) \max_{\hat{y}_i \in \hat{y}} \mathbf{y_j}^\top \mathbf{\hat{y}_i}}{\sum_{y_j \in y} \text{idf}(y_j)} \quad (4)$$

$$F_{\text{BERT}} = 2 \frac{R_{\text{BERT}} \cdot P_{\text{BERT}}}{R_{\text{BERT}} + P_{\text{BERT}}} \quad (5)$$

where $\mathbf{y}$ and $\mathbf{\hat{y}}$ represent BERT contextual embeddings of reference word y and candidate word $\hat{y}$, respectively. The function idf(·) calculates inverse document frequency (idf). In our DSR, we do not use the idf since Zhang et al. (2019) requires to use the entire dataset including test set for calculation. Besides, ROUGE do not use similar weight, so we do not include idf for consistency.

3 Experimental Setup

3.1 Datasets

Gigaword corpus It is an English sentence summarization dataset based on annotated Gigaword (Napoles et al., 2012). A single sentence summarization is paired with a short article. We use the OpenNMT provided version It contains 3.8M training, 189k development instances. We randomly sample 15k instances as our test data.

CNN/Daily Mail dataset It consists of online news articles and their corresponding multi-sentence abstracts (Hermann et al., 2015; Nallapati et al., 2016). We use the non-anonymized version provided by See et al. (2017), which contains 287k training, 13k validation, and 11k testing examples. We truncate the articles to 400 tokens and limit the summary lengths to 100 tokens.

3.2 Pretrain

We first pretrain a sequence-to-sequence model with attention using XENT and then select the best parameters to initialize models for RL. Our models have 256-dimensional hidden states and 128-dimensional word embeddings and also incorporates the pointer mechanism (See et al., 2017) for handling out of vocabulary words.

3.3 Baseline

In abstractive summarization, ROUGE (Lin, 2004) is a common evaluation metric to provide a sentence-level reward for RL. However, using ROUGE as a pure RL objective may cause too many repetitions and reduced fluency in outputs. Paulus et al. (2018) propose a hybrid learning objective that combines XENT and self-critical ROUGE reward (Paulus et al., 2018).

$$L_{\text{baseline}} = \gamma L_{\text{rouge}} + (1 - \gamma) L_{\text{XENT}} \quad (6)$$

where γ is a scaling factor and the F score of ROUGE-L is used as the reward to calculate L_{Rouge}. In our experiment, we select $\gamma = 0.998$ for Gigaword Corpus and $\gamma = 0.9984$ for CNN/Daily Mail dataset. [1] Note that we do not avoid repetition during the test time as Paulus et al. (2018) do, because we want to examine the repetition of sentence directly produced after training.

3.4 Proposed Objective Functions

Inspired by the above objective function (Paulus et al., 2018), we optimize RL models with a similar loss function as equation 2. Instead of ROUGE-L, we incorporate BERTSCORE, a DSR to provide sentence-level feedback.

In our experiment, L_{DSR} is the self-critical RL loss (equation 2) with F_{BERT} as the reward. We introduce the following objective functions:

DSR+ROUGE: A combined reward function of ROUGE and F_{BERT}

$$L_1 = \gamma L_{\text{DSR}} + (1 - \gamma) L_{\text{rouge}} \quad (7)$$

In our experiment, we select $\gamma = 0.5$ for both datasets to balance the influence of two reward functions.

DSR+XENT: F_{BERT} reward with XENT to make the generated phrases more readable.

$$L_2 = \gamma' L_{\text{DSR}} + (1 - \gamma') L_{\text{XENT}} \quad (8)$$

In our experiment, we select $\gamma' = 0.998$ for Gigaword Corpus and $\gamma' = 0.9984$ for CNN/Daily Mail dataset.

DSR: Pure F_{BERT} objective function without any teacher forcing.

$$L_3 = L_{\text{DSR}} \quad (9)$$

4 Results

For the abstractive summarization task, we test our models with different objectives on the Gigaword and CNN/Daily Mail datasets. We choose F_{BERT} and ROUGE-L as our automatic evaluation metrics. For multi-sentence summaries in
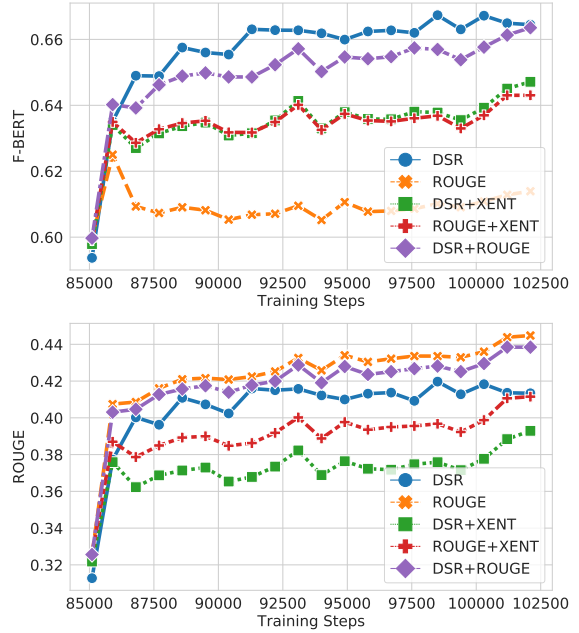

Figure 1: F_{BERT} (left) and ROUGE (right) on Gigaword Corpus during training time after pretraining with XENT. Note that ROUGE is not our target evaluation metrics.

CNN/Daily Mail dataset, we calculate ROUGE-L and F_{BERT} after concatenating the original abstract sentences or their BERT contextualized embeddings. The results are shown in Table 1. The model DSR+ROUGE (equation 7) achieves the best ROUGE performance on Gigaword and similar performance comparing to ROUGE model on CNN/Daily Mail. Training with DSR (equation 9) and DSR+XENT (equation 8) can obtain the best BERTScores as expected. It is also expected that ROUGE model will obtain worse BERTScores as simply optimizing ROUGE will generate less readable sentences (Paulus et al., 2018); however, DSR model without XENT as a teacher forcing can improve the performance of pretrained model in both F_{BERT} and ROUGE-L scale.

Note that DSR model's ROUGE-L is high in training time but does not have a good generalization on test set, and ROUGE-L is not our target evaluation metrics. In the next section, we will do human evaluation to analyze the summarization performance of different reward systems.

5 Human Evaluation

We perform human evaluation on the Amazon Mechanical Turk to assure the benefits of DSR on output sentences' coherence and fluency. We randomly sample 500 items as an evaluation set us-

Model	Gigawords				CNN/Daily Mail			
	F_{BERT}	P_{BERT}	R_{BERT}	ROUGE	F_{BERT}	P_{BERT}	R_{BERT}	ROUGE
XENT	65.78	67.61	64.53	40.77	62.77	62.18	63.79	29.46
ROUGE	61.46	60.89	62.70	42.73	60.11	59.31	61.33	**33.89**
ROUGE+XENT	66.50	67.24	66.28	42.72	61.38	61.07	62.17	33.07
DSR+ROUGE	66.48	66.79	66.65	**42.95**	65.01	65.92	64.56	33.61
DSR+XENT	67.02	**67.69**	66.85	42.29	66.64	66.06	67.63	31.28
DSR	**67.06**	67.34	**67.28**	41.73	**66.93**	**66.27**	**67.98**	30.96

Table 1: Results on Gigawords and CNN/Daily Mail for abstractive summarization. Upper rows show the results of baselines. Rouge stands for the F score of Rouge-L.

Task	Model v. base	Gigawords			CNN/Daily Mail		
		Win	Lose	Tie	Win	Lose	Tie
Relevance	DSR + XENT	31.0%	25.2%	43.8%	51.1%	34.1%	14.8%
	DSR	45.4%	27.2%	27.4%	48.7%	38.5%	12.8%
Fluency	DSR + XENT	40.2%	19.6%	40.2%	55.8%	28.5%	15.6%
	DSR	45.4%	28.8%	25.8%	54.0%	31.7%	14.3%

Table 2: Human evaluation results on Gigaword and CNN/Daily Mail for abstractive summarization. ROUGE+XENT is the selected baseline model.

ing uniform distribution. During the evaluation, given the context without knowing the ground truth summary, judges are asked to decide two things between generations of ROUGE+XENT and DSR+XENT: 1. which summary is more relevant to the article. 2. which summary is more fluent. Ties are permitted. A similar test is also done between ROUGE+XENT and DSR model. As shown in Table 2, DSR and DSR+XENT models improve the relevance and fluency of generated summary significantly. In addition, using pure DSR achieve better performances on Gigaword Corpus and comparable results as DSR+XENT on CNN/Daily Mail. While Paulus et al.'s (2018) objective function requires XENT to make generated sentence readable, our proposed DSR does not require XENT and can limit the exposure bias originated from it.

6 Analysis

Diversity Other than extractive summarization, abstractive summarization allows more degrees of freedom in the choice of words. While simply selecting words from the article made the task easier to train, higher action space can provide more paths to potentially better results (Nema et al., 2017). Using the DSR as deep RL reward will support models to choose actions that are not n-grams of the articles. In Table 4, we list a few generated samples on Gigaword Corpus. In our first example in Table 4, the word "sino-german" provides

Model	Gigawords		CNN/Daily Mail	
	Rep(%)	Div(%)	Rep(%)	Div(%)
ROUGE+XENT	11.57	16.33	66.82	21.32
DSR+ROUGE	18.42	15.97	50.41	34.70
DSR+XENT	20.15	16.24	25.98	31.80
DSR	**7.20**	**19.17**	**22.03**	**41.35**

Table 3: Qualitative analysis on repetition(Rep) / diversity(Div). They are calculated by the percentage of repeat/out-of-article n-grams (unigrams for Gigaword and 5-grams for CNN/Daily Mail) in generated sentences.

an interesting and efficient way to express the relation between China and Germany. F_{BERT} is also improved by making this change. In addition, the second example in Table 4 shows that RL model with DSR corrects the sentence' grammar and significantly improves the F_{BERT} score by switching "down" to an unseen word "drops". On the other hand, when optimizing DSR to improve the diversity of generation, some semantically similar words may also be generated and harm the summarization quality as shown in the third example in Table 4. The new token "wins" reduces the scores of both metrics. We also evaluate the diversity of a model quantitively by averaging the percentage of out-of-article n-grams in generated sentences. Results can be found in Table 3. The DSR model achieves the highest diversity.

1) Context	economic links between china and germany will be at the center of talks here next month between li peng and chancellor helmut kohl when the chinese premier makes a week-long visit to germany , sources on both sides said .
Groundtruth	li peng 's visit to germany to focus on economic ties
ROUGE+XENT	chinese premier to visit germany next month $\{ F_{\text{BERT}}: 49.48, \text{ROUGE}:\textbf{18.64}\}$
DSR+XENT	chinese premier to visit germany next month $\{ F_{\text{BERT}}: 49.48, \text{ROUGE}:\textbf{18.64}\}$
DSR	**sino-german** economic **links** to be at center of talks **with** germany $\{F_{\text{BERT}}: \textbf{50.77}, \text{ROUGE}:17.33\}$
2) Context	the sensitive index dropped by #.## percent friday on the bombay stock exchange -lrb- bse -rrb- following brisk inquiries from operators that generally pressed sales to square up positions on the last day of the current settlement .
Groundtruth	sensex falls in bombay stock exchange
ROUGE+XENT	sensitive index down on bombay stock exchange $\{F_{\text{BERT}}: 70.03, \text{ROUGE}:\textbf{45.62}\}$
DSR+XENT	sensitive index **drops** on bombay stock exchange $\{F_{\text{BERT}}: \textbf{73.92}, \text{ROUGE}:\textbf{45.62}\}$
DSR	sensitive index **drops** on bombay stock exchange $\{F_{\text{BERT}}: \textbf{73.92}, \text{ROUGE}:\textbf{45.62}\}$
3) Context	belgium 's rik verbrugghe took victory in tuesday 's prologue stage of the tour de romandie , with a confident ride through the streets of geneva 's old town .
Groundtruth	verbrugghe takes prologue victory in tour de romandie
ROUGE+XENT	verbrugghe **takes** victory in prologue stage of tour de romandie $\{F_{\text{BERT}}: 91.41, \text{ROUGE}:75.93\}$
DSR+XENT	verbrugghe **takes** victory in tour de romandie prologue stage $\{F_{\text{BERT}}: \textbf{91.56}, \text{ROUGE}:\textbf{81.79}\}$
DSR	verbrugghe **wins** victory in tour de romandie prologue stage $\{F_{\text{BERT}}: 89.45, \text{ROUGE}:70.10\}$
4) Context	european finance ministers on saturday urged swedes to vote " yes " to adopting the euro , saying entry into the currency bloc would be good for sweden and the rest of europe .
Groundtruth	european finance ministers urge swedes to vote yes to euro
ROUGE+XENT	**eu** finance ministers urge swedes to vote yes to euro $\{F_{\text{BERT}}: \textbf{96.63}, \text{ROUGE}:\textbf{93.47}\}$
DSR+XENT	**eu** finance ministers urge swedes to vote yes to adopting euro $\{F_{\text{BERT}}: 90.38, \text{ROUGE}:88.89\}$
DSR	**eu** finance ministers urge swedes to vote yes to euro euro $\{F_{\text{BERT}}: 95.05, \text{ROUGE}:\textbf{93.47}\}$

Table 4: Qualitative analysis of generated samples on Gigaword corpus. Generated words that do not appear in the context are marked blue. Repeated words are marked red. The first two examples represent DSR's generated tokens are more diverse. However, it may suffer from problems as shown in example 3 and 4.

Repetition Repetition will lead to lower F_{BERT} as shown in the last example in Table 4. Using DSR reduces the probability of producing repetitions. The average percentage of repeated n-grams in generated sentences are presented in the Table 3. As shown in this table, unlike ROUGE, the DSR model can achieve high fluency without XENT; moreover, it produces the fewest repetitions among all the rewards. Table 4 gives an example that DSR produces a repeated word (from example 4), but it does not reflect the overall distribution of repeated word generation for all evaluated models.

7 Conclusion

This paper demonstrates the effectiveness of applying the distributional semantic reward to reinforcement learning in abstractive summarization, and specifically, we choose BERTSCORE. Our experimental results demonstrate that we achieve better performance on Gigaword and CNN/Daily Mail datasets. Besides, the generated sentences have fewer repetitions, and the fluency is also improved. Our finding is aligned to a contemporaneous study (Wieting et al., 2019) on leveraging semantic similarity for machine translation.

References

Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. 2017. An actor-critic algorithm for sequence prediction. In *5th International Conference on Learning Representations, Conference Track Proceedings*.

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, Conference Track Proceedings*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Irwan Bello, Hieu Pham, Quoc V. Le, Mohammad Norouzi, and Samy Bengio. 2017. Neural combinatorial optimization with reinforcement learning. In *5th International Conference on Learning Representations, Workshop Track Proceedings*.

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.

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.

Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. 2010. Opinosis: A graph based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 340–348.

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.

Alex M Lamb, Anirudh Goyal Alias Parth Goyal, Ying Zhang, Saizheng Zhang, Aaron C Courville, and Yoshua Bengio. 2016. Professor forcing: A new algorithm for training recurrent networks. In *Advances In Neural Information Processing Systems*, pages 4601–4609.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Ibrahim F Moawad and Mostafa Aref. 2012. Semantic graph reduction approach for abstractive text summarization. In *2012 Seventh International Conference on Computer Engineering & Systems (ICCES)*, pages 132–138. IEEE.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Ça glar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *CoNLL 2016*, page 280.

Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. 2012. Annotated gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, pages 95–100. Association for Computational Linguistics.

Preksha Nema, Mitesh M Khapra, Anirban Laha, and Balaraman Ravindran. 2017. Diversity driven attention model for query-based abstractive summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1063–1072.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *6th International Conference on Learning Representations, Conference Track Proceedings*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*, pages 2227–2237.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *4th International Conference on Learning Representations, Conference Track Proceedings*.

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.

Seonggi Ryang and Takeshi Abekawa. 2012. Framework of automatic text summarization using reinforcement learning. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 256–265. Association for Computational Linguistics.

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.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Advances in Neural Information Processing Systems*, pages 2692–2700.

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond bleu: Training neural machine translation with semantic similarity. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 4344–4355.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

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