SimCLS: A Simple Framework for Contrastive Learning of Abstractive Summarization

Yixin Liu
Carnegie Mellon University
yixinl2@cs.cmu.edu

Pengfei Liu *
Carnegie Mellon University
pliu3@cs.cmu.edu

Abstract

In this paper, we present a conceptually simple while empirically powerful framework for abstractive summarization, SIMCLS, which can bridge the gap between the *learning objective* and *evaluation metrics* resulting from the currently dominated sequence-to-sequence learning framework by **formulating text generation as a reference-free evaluation problem** (i.e., quality estimation) assisted by *contrastive learning*. Experimental results show that, with minor modification over existing top-scoring systems, SimCLS can improve the performance of existing top-performing models by a large margin. Particularly, 2.51 absolute improvement against BART (Lewis et al., 2020) and 2.50 over PEGASUS (Zhang et al., 2020a) w.r.t ROUGE-1 on the CNN/DailyMail dataset, driving the state-of-the-art performance to a new level. We have open-sourced our codes and results: https://github.com/yixinL7/SimCLS. Results of our proposed models have been deployed into EXPLAINABOARD (Liu et al., 2021a) platform, which allows researchers to understand our systems in a more fine-grained way.

1 Introduction

Sequence-to-sequence (Seq2Seq) neural models (Sutskever et al., 2014) have been widely used for language generation tasks, such as abstractive summarization (Nallapati et al., 2016) and neural machine translation (Wu et al., 2016). While abstractive models (Lewis et al., 2020; Zhang et al., 2020a) have shown promising potentials in the summarization task, they share the widely acknowledged challenges of Seq2Seq model training. Specifically, Seq2Seq models are usually trained under the framework of Maximum Likelihood Estimation (MLE) and in practice they are commonly trained with the *teacher-forcing* (Williams and Zipser, 1989) algorithm. This introduces a gap between the *objective function* and the *evaluation metrics*, as the objective function is based on local, token-level predictions while the evaluation metrics (e.g. ROUGE (Lin, 2004)) would compare the holistic similarity between the gold references and system outputs. Furthermore, during the test stage the model needs to generate outputs autoregressivelly, which means the errors made in the previous steps will accumulate. This gap between the *training* and *test* has been referred to as the *exposure bias* in the previous work (Bengio et al., 2015; Ranzato et al., 2016).

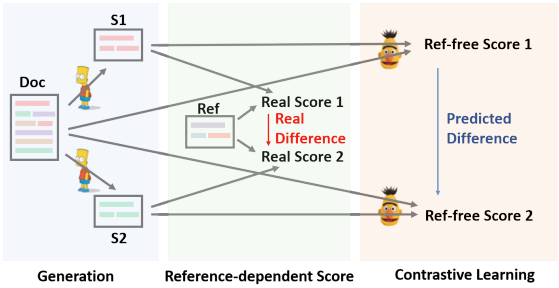

Figure 1: SimCLS framework for two-stage abstractive summarization, where Doc, S, Ref represent the document, generated summary and reference respectively. At the first stage, a Seq2Seq generator (BART) is used to generate candidate summaries. At the second stage, a scoring model (RoBERTa) is used to predict the performance of the candidate summaries based on the source document. The scoring model is trained with contrastive learning, where the training examples are provided by the Seq2Seq model.

A main line of approaches (Paulus et al., 2018; Li et al., 2019) proposes to use the paradigm of Reinforcement Learning (RL) to mitigate the aforementioned gaps. While RL training makes it possible to train the model with rewards based on global predictions and closely related to the evaluation metrics, it introduces the common challenges of deep RL. Specifically, RL-based training suffers from the noise gradient estimation (Greensmith et al., 2004) problem, which often makes the training un-

*Corresponding author.

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Short Papers), pages 1065–1072
August 1–6, 2021.

stable and sensitive to hyper-parameters. Minimum risk training, as an alternative, has also been used in the language generation tasks (Shen et al., 2016; Wieting et al., 2019). However, the accuracy of the estimated loss is restricted by the number of sampled outputs. Other methods (Wiseman and Rush, 2016; Norouzi et al., 2016; Edunov et al., 2018) aim to extend the framework of MLE to incorporate sentence-level scores into the objective functions. While these methods can mitigate the limitations of MLE training, the relation between the evaluation metrics and the objective functions used in their methods can be indirect and implicit.

Among this background, in this work we generalize the paradigm of contrastive learning (Chopra et al., 2005) to introduce an approach for abstractive summarization which achieves the goal of directly optimizing the model with the corresponding evaluation metrics, thereby mitigating the gaps between training and test stages in MLE training. While some related work (Lee et al., 2021; Pan et al., 2021) have proposed to introduce a contrastive loss as an augmentation of MLE training for conditional text generation tasks, we instead choose to disentangle the functions of contrastive loss and MLE loss by introducing them at different stages in our proposed framework.

Specifically, inspired by the recent work of Zhong et al. (2020); Liu et al. (2021b) on text summarization, we propose to use a two-stage model for abstractive summarization, where a Seq2Seq model is first trained to generate candidate summaries with MLE loss, and then a parameterized evaluation model is trained to rank the generated candidates with contrastive learning. By optimizing the generation model and evaluation model at separate stages, we are able to train these two modules with supervised learning, bypassing the challenging and intricate optimization process of the RL-based methods.

Our main contribution in this work is to approach metric-oriented training for abstractive summarization by proposing a generate-then-evaluate two-stage framework with contrastive learning, which not only put the state-of-the-art performance on CNN/DailyMail to a new level (2.2 ROUGE-1 improvement against the baseline model), also demonstrates the great potentials of this two-stage framework, calling for future efforts on optimizing Seq2Seq models using methods beyond maximum likelihood estimation.

2 Contrastive Learning Framework for Abstractive Summarization

Given a source document D and a reference summary $\hat{S}$, the goal of an abstractive summarization model f is to generate the candidate summary $S = f(D)$ such that it receives the highest score $m = M(S, \hat{S})$ assigned by an evaluation metric M. In this work, we break down the holistic generation process into two stages which consist of a *generation model* g for generating candidate summaries and a *evaluation model* h for scoring and selecting the best candidate. Fig 1 illustrates the general framework.

Stage I: Candidate Generation The generation model $g(\cdot)$ is a Seq2Seq model trained to maximize the likelihood of reference summary $\hat{S}$ given the source document D. The pre-trained $g(\cdot)$ is then used to produce multiple candidate summaries $S_1, \cdots, S_n$ with a sampling strategy such as Beam Search, where n is the number of sampled candidates.

Stage II: Reference-free Evaluation The high-level idea is that a better candidate summary S_i should obtain a higher quality score w.r.t the source document D. We approach the above idea by contrastive learning and define an *evaluation function* $h(\cdot)$ that aims to assign different scores $r_1, \cdots, r_n$ to the generated candidates solely based on the similarity between the source document and the candidate S_i, i.e., $r_i = h(S_i, D)$. The final output summary S is the candidate with the highest score:

$$S = \underset{S_i}{\operatorname{argmax}}\, h(S_i, D). \tag{1}$$

Here, we instantiate $h(\cdot)$ as a large pre-trained self-attention model, RoBERTa (Liu et al., 2019). It is used to encode S_i and D separately, and the cosine similarity between the encoding of the first tokens is used as the similarity score r_i.

Contrastive Training Instead of explicitly constructing a positive or negative example as most existing work with contrastive learning have adopted (Chen et al., 2020; Wu et al., 2020), here the "*contrastiveness*" is reflect in the diverse qualities of naturally generated summaries evaluated by a parameterized model $h(\cdot)$. Specifically, we introduce a ranking loss to $h(\cdot)$:

$$\begin{aligned} L = &\sum_i \max(0, h(D, \tilde{S}_i) - h(D, \hat{S})) \\ &+ \sum_i \sum_{j>i} \max(0, h(D, \tilde{S}_j) - h(D, \tilde{S}_i) + \lambda_{ij}), \end{aligned} \tag{2}$$

where $\tilde{S}_1, \cdots, \tilde{S}_n$ is descendingly sorted by $M(\tilde{S}_i, \hat{S})$. Here, $\lambda_{ij} = (j-i)*\lambda$ is the corresponding margin that we defined following Zhong et al. (2020), and λ is a hyper-parameter.[1] M can be any automated evaluation metrics or human judgments and here we use ROUGE (Lin, 2004).

3 Experiments

3.1 Datasets

We use two datasets for our experiments. The dataset statistics are listed in Appendix A.

`CNNDM` CNN/DailyMail[2] (Hermann et al., 2015; Nallapati et al., 2016) dataset is a large scale news articles dataset.

`XSum` XSum[3] (Narayan et al., 2018) dataset is a highly abstractive dataset containing online articles from the British Broadcasting Corporation (BBC).

3.2 Evaluation Metrics

We use ROUGE-1/2/L (R-1/2/L) as the main evaluation metrics for our experiments. We also evaluate our model on the recently developed semantic similarity metrics, namely, BERTScore (Zhang et al., 2020b) and MoverScore (Zhao et al., 2019).

3.3 Base Systems

As the generation model and the evaluation model in our two-stage framework are trained separately, we use pre-trained state-of-the-art abstractive summarization systems as our generation model. Specifically, we use **BART** (Lewis et al., 2020) and **Pegasus** (Zhang et al., 2020a) as they are popular and have been comprehensively evaluated.

3.4 Training Details

For baseline systems, we use the checkpoints provided by the *Transformers*[4] (Wolf et al., 2020) library. We use diverse beam search (Vijayakumar et al., 2016) as the sampling strategy to generate candidate summaries. We use 16 groups for diversity sampling, which results in 16 candidates. To train the evaluation model, we use Adam optimizer (Kingma and Ba, 2015) with learning rate scheduling. The model performance on the validation set is used to select the checkpoint. More details are described in Appendix B.

[1] As it is insensitive, we fix it to 0.01 in our experiments.

[2] https://cs.nyu.edu/~kcho/DMQA/

[3] https://github.com/EdinburghNLP/XSum

[4] https://github.com/huggingface/transformers

System	R-1	R-2	R-L	BS	MS
BART*	44.16	21.28	40.90	-	-
Pegasus*	44.17	21.47	41.11	-	-
Prophet*	44.20	21.17	41.30	-	-
GSum*	45.94	**22.32**	42.48	-	-
Origin	44.39	21.21	41.28	64.67	58.67
Min	33.17	11.67	30.77	58.09	55.75
Max	54.36	28.73	50.77	70.77	61.67
Random	43.98	20.06	40.94	64.65	58.60
SimCLS	**46.67**†	22.15†	**43.54**†	**66.14**†	**59.31**†

Table 1: Results on `CNNDM`. **BS** denotes BERTScore, **MS** denotes MoverScore. **Origin** denotes the original performance of the baseline model. **Min**, **Max**, **Random** are the oracles that select candidates based on their ROUGE scores. †: significantly better than the baseline model (Origin) ($p < 0.01$). *: results reported in the original papers.

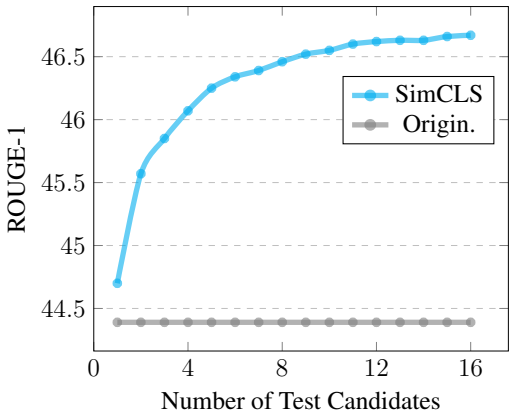

Figure 2: Test performance with different numbers of candidate summaries on `CNNDM`. **Origin** denotes the original performance of the baseline model.

3.5 Results on CNNDM dataset

The results on `CNNDM` dataset are shown in Tab. 1. We use the pretrained BART[5] as the base generation model (**Origin**). We use **BART**, **Pegasus**, **GSum** (Dou et al., 2021) and **ProphetNet** (Qi et al., 2020) for comparison. Notably, the Max oracle which always selects the best candidate has much better performance than the original outputs, suggesting that using a diverse sampling strategy can further exploit the potential power of the pretrained abstractive system. Apart from ROUGE, we also present the evaluation results on semantic similarity metrics. Our method is able to outperform the baseline model on all metrics, demonstrating its improvement is beyond exploiting the potential artifacts of ROUGE. While the scale of improvement is harder to interpret with these metrics, we note that the improvement is able to pass the significance test.

[5] 'facebook/bart-large-cnn'

System	Summary	Article
Ref.	chris ramsey says he has no problem shaking hands with john terry . queens park rangers host chelsea in the premier league on sunday . terry was once banned and fined for racist comments at loftus road . rio ferdinand , brother of anton , will not be fit to play against chelsea .	**queens park rangers manager chris ramsey has revealed he will have no problem shaking john terry's hand in light of the racist comments the former england captain directed at former rs defender anton ferdinand four years ago .** ***terry , who will line up against ramsey's side , was banned for four games and fined # 220,000 for the remarks made in october 2011 during chelsea's 1-0 defeat at loftus road .*** but ramsey , the premier league's only black manager , thinks the issue has been dealt with . ... ' i don't know what his feelings are towards me . as long as there wasn't anything on the field that was unprofessional by him , i would shake his hand . . **queens park rangers manager chris ramsey speaks to the media on friday ahead of the chelsea match .** chelsea captain john terry controls the ball during last weekend's premier league match against stoke . ramsey arrives for friday's pre-match press conference as qpr prepare to host chelsea at loftus road . ' the whole episode for british society sat uncomfortably . it's not something we want to highlight in football . it happened and it's being dealt with . we have to move on . and hopefully everyone has learned something from it . ' . *ramsey revealed that rio ferdinand , who labelled terry an idiot for the abuse aimed at his brother , won't be fit in time for a reunion with the chelsea skipper this weekend .* but the 52-year-old suspects his player's one-time england colleague will be on the receiving end of a hostile welcome from the home fans on his return the scene of the unsavoury incident . ... ferdinand and terry argue during qpr's 1-0 victory against chelsea at loftus road in october 2011 . **rio ferdinand , brother of anton , will not be fit for sunday's match against chelsea .**
SimCLS	queens park rangers host chelsea in the premier league on sunday . qpr boss chris ramsey says he will have no problem shaking john terry's hand . terry was banned for four games and fined # 220,000 for racist comments . rio ferdinand , brother of anton , will not be fit for the match at loftus road .	
Origin.	john terry was banned for four games and fined # 220,000 for the remarks made in october 2011 during chelsea's 1-0 defeat at loftus road . terry will line up against chris ramsey's side on sunday . rio ferdinand , who labelled terry an idiot for the abuse aimed at his brother , won't be fit in time for a reunion with the chelsea skipper this weekend .	

Table 2: Sentence alignments between source articles and summaries on CNNDM dataset. The aligned sentences for reference and our summaries are **bolded** (they are the same in this example). The aligned sentences for baseline summaries are *italicized*. **Origin** denotes the original performance of the baseline model.

Level	System	Precision	Recall	F-Score
Entity	Origin	40.70	59.13	48.22
	SimCLS	**43.36**	**59.79**	**50.27**
Sentence	Origin	38.11	38.65	37.18
	SimCLS	**42.58**	**40.22**	**40.12**

Table 3: Performance analysis on CNNDM dataset. **Origin** denotes the original performance of the baseline model.

With the constraints of computation power, we try to use as many candidates as possible for the evaluation model training. However, we also notice that our method is robust to the specific number of candidates, as during test we found that our model is still able to outperform the baseline model with fewer candidates, which is illustrated in Fig. 2.

3.6 Fine-grained Analysis

To demonstrate that our method is able to make meaningful improvement w.r.t the summary quality, here we compare our method with the baseline model at different semantic levels on CNNDM.

3.6.1 Entity-level

Inspired by the work of Gekhman et al. (2020) and Jain et al. (2020), we compare the model performance w.r.t the *salient entities*, which are entities in source documents that appear in the reference summaries. Specifically, (1) we extract the entities from the source documents,[6] (2) select the *salient entities* based on the entities in reference summaries, (3) compare the *salient entities* with entities in candidate summaries. Results in Tab. 3 demonstrate that our method can better capture the important semantic information of the source documents.

3.6.2 Sentence-level

Sentence Alignments Here we investigate if our method makes sentence-level differences compared to the baseline model. Specifically, (1) we match each sentence in the summaries to a sentence in the source documents based on their similarity (indicated by ROUGE scores),[7] (2) compute the sentence-level similarity between the reference and system-generated summaries based on the overlaps of their matched sentences in the source documents. The results in Tab. 3 demonstrate that the generated summaries of our method is more similar to the reference summaries at the sentence level.

Positional Bias In Tab. 2, we present a case study of the sentence alignment. We use the same matching approach to map the summary sentences to the sentences in source articles. In this example, the output of our method focuses on the same sentences as the reference summary does, while the baseline summary focuses on some different sentences.

Interestingly, the reference summary focuses on the very last sentence in the article, and our method can follow this pattern. Upon examining this pattern, we notice a positional bias of abstractive models when handling long source articles (more than

[6] We use a pre-trained NER model provided by spaCy to extract the entities: https://spacy.io/

[7] Notably, this matching approach formulates an extractive oracle when reference summaries are used for matching, which achieves 54.54/30.73/50.35 ROUGE-1/2/L scores.

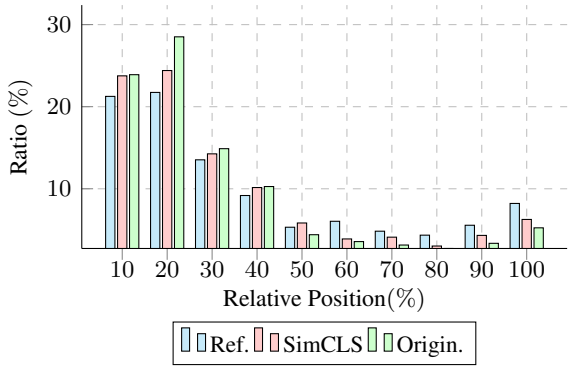

Figure 3: Positional Bias. X-asis: the relative position of the matched sentence in source documents. Y-axis: the ratio of the matched sentences. For fair comparison, articles are first truncated to the generator's maximum input length. **Origin** denotes the original performance of the baseline model.

System	R-1	R-2	R-L	BS	MS
BART*	45.14	22.27	37.25	-	-
Pegasus*	47.21	24.56	39.25	-	-
GSum*	45.40	21.89	36.67	-	-
Origin	47.10	24.53	39.23	69.48	61.34
Min	40.97	19.18	33.68	66.01	59.58
Max	52.45	28.28	43.36	72.56	62.98
Random	46.72	23.64	38.55	69.30	61.23
SimCLS	**47.61**†	**24.57**	**39.44**†	**69.81**†	**61.48**†

Table 4: Results on XSum dataset. **BS** denotes BERTScore, **MS** denotes MoverScore. **Origin** denotes the original performance of the baseline model. **Min**, **Max**, **Random** are the oracles that select candidates based on their ROUGE scores. †: significantly better than the baseline model (Origin) ($p < 0.05$). *: results reported in the original papers.

30 sentences). Fig. 3 shows that the baseline summaries are more likely to focus on the head sentences compared to the references, which may result from the autoregressive generation process of the Seq2Seq models. Our method is able to mitigate this bias, as the candidate sampling process (diverse beam search) generates candidates different from the original outputs, and our evaluation model can assess the holistic quality of the candidates.

3.7 Results on XSum dataset

To evaluate our method's performance beyond CNNDM dataset, we also test our method on XSum dataset, and the results are shown in Tab. 4. Here, we use Pegasus[8] as the base system since it achieves better performance than BART on XSum. We follow the same sampling strategy to generate the training data. However, as this strategy generally results in lower ROUGE-2 score on XSum dataset, we use a different strategy to generate the validation and test data (4 candidates generated by 4 diverse groups). Our method is still able to outperform the baseline, but with a smaller margin compared to CNNDM. Summaries in XSum are shorter (one-sentence) and more abstractive, which restricts the semantic diversity of candidates and makes it harder to make meaningful improvement.

4 Conclusion

In this work, we present a contrastive summarization framework that aims to optimize the quality of generated summaries at summary-level, which mitigates the discrepancy between the training and test stages in the MLE framework. Apart from the significant improvement over the baseline model on CNNDM dataset, we present a comprehensive evaluation at different semantic levels, explaining the sources of the improvement made by our method. Notably, our experimental results also indicate that the existing abstractive systems have the potential of generating candidate summaries much better than the original outputs. Therefore, our work opens up the possibility for future directions including (1) extending this two-stage strategy to other datasets for abstractive models; (2) improving the training algorithms for abstractive models towards a more holistic optimization process.

[8] 'google/pegasus-xsum'

Acknowledgements

We thank Professor Graham Neubig and anonymous reviewers for valuable feedback and helpful suggestions. This work was supported in part by a grant under the Northrop Grumman SOTERIA project and the Air Force Research Laboratory under agreement number FA8750-19-2-0200. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Research Laboratory or the U.S. Government.

References

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks.

In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 1171–1179, Cambridge, MA, USA. MIT Press.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE.

Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2021. GSum: A general framework for guided neural abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4830–4842, Online. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Classical structured prediction losses for sequence to sequence learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 355–364, New Orleans, Louisiana. Association for Computational Linguistics.

Zorik Gekhman, Roee Aharoni, Genady Beryozkin, Markus Freitag, and Wolfgang Macherey. 2020. KoBE: Knowledge-based machine translation evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3200–3207, Online. Association for Computational Linguistics.

Evan Greensmith, Peter L Bartlett, and Jonathan Baxter. 2004. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*, 5(9).

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 1693–1701, Cambridge, MA, USA. MIT Press.

Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. 2020. SciREX: A challenge dataset for document-level information extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7506–7516, Online. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Seanie Lee, Dong Bok Lee, and Sung Ju Hwang. 2021. Contrastive learning with adversarial perturbations for conditional text generation. In *International Conference on Learning Representations*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Pengfei Liu, Jinlan Fu, Yang Xiao, Weizhe Yuan, Shuaicheng Chang, Junqi Dai, Yixin Liu, Zihuiwen Ye, and Graham Neubig. 2021a. Explainaboard: An explainable leaderboard for nlp. *arXiv preprint arXiv:2104.06387*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Yixin Liu, Zi-Yi Dou, and Pengfei Liu. 2021b. RefSum: Refactoring neural summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1437–1448, Online. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, 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.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Mohammad Norouzi, Samy Bengio, zhifeng Chen, Navdeep Jaitly, Mike Schuster, Yonghui Wu, and Dale Schuurmans. 2016. Reward augmented maximum likelihood for neural structured prediction. In *Advances in Neural Information Processing Systems*, volume 29, pages 1723–1731. Curran Associates, Inc.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation.

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

Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. ProphetNet: Predicting future n-gram for sequence-to-SequencePre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2401–2410, Online. Association for Computational Linguistics.

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, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Minimum risk training for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1692, Berlin, Germany. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA. MIT Press.

Ashwin K. Vijayakumar, Michael Cogswell, Ramprasaath R. Selvaraju, Qing Sun, Stefan Lee, David J. Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *CoRR*, abs/1610.02424.

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 Annual Meeting of the Association for Computational Linguistics*, pages 4344–4355, Florence, Italy. Association for Computational Linguistics.

Ronald J. Williams and David Zipser. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural Comput.*, 1(2):270–280.

Sam Wiseman and Alexander M. Rush. 2016. Sequence-to-sequence learning as beam-search optimization. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1296–1306, Austin, Texas. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Hanlu Wu, Tengfei Ma, Lingfei Wu, Tariro Manyumwa, and Shouling Ji. 2020. Unsupervised reference-free summary quality evaluation via contrastive learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3612–3621, Online. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

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 *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. 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 563–578, Hong Kong, China. Association for Computational Linguistics.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208, Online. Association for Computational Linguistics.

A Dataset Statistics

Datasets	# Num			Avg. Len	
	Train	Valid	Test	Doc.	Sum.
CNNDM	287K	13K	11K	768.6	55.7
XSum	203K	11K	11K	429.2	23.3

Table 5: Datasets Statistics. Len is the length of tokens.

The source documents and reference summaries are lower-cased. Due to the input length limitation, some source documents are truncated during training.

B Experiment Details

Candidate Generation We use diverse beam search to generate the candidate summaries. We use the same beam search configuration as the original work except those related to diverse beam search. In particular, the diversity penalty is set to 1, and we use 16 diversity groups with 16 beams, which results in 16 candidates.

Model We use the pretrained RoBERTa with 'roberta-base' version provided by the *Transformers* library as our evaluation model, which contains 125M parameters.

Optimizer We use Adam optimizer with learning rate scheduling:

$$lr = 0.002 \cdot \min(\text{step\_num}^{-0.5}, \text{step\_num} \cdot \text{warmup\_steps}^{-1.5}), \quad (3)$$

where the warmup_steps is 10000.

Training details The batch size in our experiments is 32. We evaluate the model performance on the validation set at every 1000 steps, using the averaged ROUGE-1/2/L score as the selecting criteria. The training is converged in 5 epochs, which takes around 40 hours on 4 GTX-1080-Ti GPUs on CNN/DailyMail dataset and 20 hours on XSum dataset.