Balancing Lexical and Semantic Quality in Abstractive Summarization

Jeewoo Sul and **Yong Suk Choi***
Department of Computer Science
Hanyang University, Seoul, Korea
{jeewoo25, cys}@hanyang.ac.kr

Abstract

An important problem of the sequence-to-sequence neural models widely used in abstractive summarization is *exposure bias*. To alleviate this problem, re-ranking systems have been applied in recent years. Despite some performance improvements, this approach remains underexplored. Previous works have mostly specified the rank through the ROUGE score and aligned candidate summaries, but there can be quite a large gap between the lexical overlap metric and semantic similarity. In this paper, we propose a novel training method in which a re-ranker balances the lexical and semantic quality. We further newly define false positives in ranking and present a strategy to reduce their influence. Experiments on the CNN/DailyMail and XSum datasets show that our method can estimate the meaning of summaries without seriously degrading the lexical aspect. More specifically, it achieves an 89.67 BERTScore on the CNN/DailyMail dataset, reaching new state-of-the-art performance. Our code is publicly available at https://github.com/jeewoo1025/BalSum.

1 Introduction

The performance of sequence-to-sequence (Seq2Seq) neural models for abstractive summarization (Lewis et al., 2020; Nallapati et al., 2016; See et al., 2017; Zhang et al., 2020) has improved significantly. The dominant training paradigm of Seq2Seq models is that of Maximum Likelihood Estimation (MLE), maximizing the likelihood of each output given the gold history of target sequences during training. However, since the models generate the sequence in an auto-regressive manner at inference, the errors made in the previous steps accumulate in the next step thereby affecting the entire sequence. This phenomenon is known as *exposure bias* (Bengio et al., 2015; Ranzato et al., 2016). To mitigate this problem, re-ranking systems (Liu et al., 2021; Liu and Liu, 2021; Liu et al., 2022; Ravaut et al., 2022) have recently been introduced to generate a more appropriate summary.

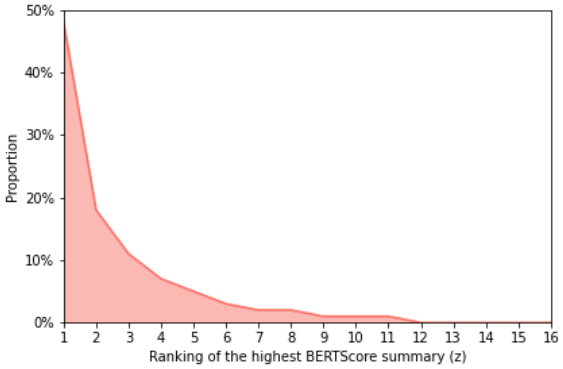

Figure 1: Distribution of z (%) for a base BART model on CNN/DM. Since a BART model generates a pool of 16 diverse beam search candidates, the X-axis ranges from 1 to 16. If $z = 1$, it means that both ROUGE and BERTscore are high. As z increases, the gap between ROUGE and BERTScore tends to increase. The Y-axis represents the proportion of z in the test set. The distribution for XSum is in Appendix A.

There are two training objectives for applying re-ranking to abstractive summarization: *contrastive learning* and *multi-task learning*. The contrastive learning-based approaches deploy margin-based losses. SimCLS (Liu and Liu, 2021) and BRIO-Ctr (Liu et al., 2022) train a large pre-trained model, such as RoBERTa (Liu et al., 2019) and BART (Lewis et al., 2020), to align the candidate summaries according to the quality. The authors use the ROUGE (Lin, 2004) score as a quality measurement. The multi-task learning-based approaches combine at least two losses that perform different roles. SummaReranker (Ravaut et al., 2022) minimizes the average over the binary cross-entropy losses optimized for each evaluation metric. In addition, BRIO-Mul (Liu et al., 2022) demonstrates that the combination of the contrastive and cross-

*Corresponding author

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 637–647

entropy loss works complementarily and has better performance.

In this paper, we analyze the three main drawbacks of existing re-ranking approaches. First, we argue that current methods focus excessively on ranking summaries in terms of lexical overlap. Inspired by Zhong et al. (2020), we conduct a preliminary study, by sorting candidate summaries in descending order based on the ROUGE score and then defining z as the rank index of the highest BERTScore summary. As demonstrated in Fig. 1, we can observe that there is a large gap between lexical overlap and semantic similarity. In a majority (52%) of cases $z > 1$. Second, despite more than half of the candidates with the same ROUGE score, previous studies do not accurately reflect quality measurements as they are trained with different ranks even if they have equal scores (Appendix F). Lastly, for the first time, we find summaries with high lexical overlap but low semantic similarity as false positives (Appendix G). They can be noises during training phrase, which are not considered substantially in the prior works.

To address these issues, we propose a novel training method in which a re-ranker balances lexical and semantic quality. Based on a two-stage framework, our model, named *BalSum*, is trained on multi-task learning. We directly reflect the ROUGE score difference on a ranking loss to preserve the lexical quality as much as possible. Then, we use a contrastive loss with instance weighting to identify summaries whose meanings are close to the document. Specifically, we define novel false positives (semantic mistakes) and present a strategy to reduce their influence in ranking. Experiments on CNN/DM and XSum datasets demonstrate the effectiveness of our method. Notably, BalSum achieves an 89.67 BERTScore on CNN/DM, reaching a new state-of-the-art performance.

2 Method

Our method follows the two-stage framework. Given a source document D, a function g is to generate a pool of candidate summaries $\mathbb{C} = \{C_1, C_2, ..., C_m\}$ at the first stage:

$$\mathbb{C} \leftarrow g(D) \tag{1}$$

Then, a function f is to assign scores to each candidate and select the best summary C^* with the highest score at the second stage:

$$C^* = \underset{C_i \in \mathbb{C}}{\text{argmax}}\{f(C_i, D)\} \tag{2}$$

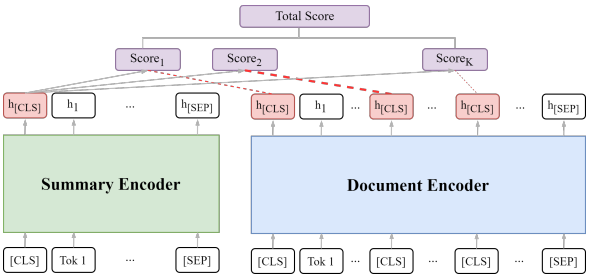

Figure 2: BalSum model architecture. The model predicts scores for candidate summaries based on the document. The thickness of the red dashed line indicates the magnitude of each score's weight.

Our goal is to train the ranking model f that identifies the correct summary from the outputs of the generation model g.

2.1 Model Architecture

We start with a bi-encoder using RoBERTa-base (Liu et al., 2019) as a back-bone neural network. Inspired by Khattab and Zaharia (2020), we aim to capture rich semantic units at the sentence level. As shown in Fig. 2, we insert the *[CLS]* tokens in front of K sentences in the document D to let them encode into multi-vector representations. Then, we compute the individual score $Score_k$ which is modeled as an inner-product:

$$Score_k = sim(E_1(C_i), E_k(D)) \tag{3}$$

where $E_1(C_i)$ and $E_k(D)(k = 1, 2, ..., K)$ mean the representations of *[CLS]* tokens for candidate summary C_i and document D, respectively. We calculate the similarity score $f(C_i, D)$:

$$f(C_i, D) = \sum_{k=1}^{K} \frac{Score_k}{\sum_{j=1}^{K} Score_j} Score_k = \sum_{k=1}^{K} w_k \cdot Score_k \tag{4}$$

In Appendix E, we show that our model can capture more information from documents at the sentence level.

2.2 Training objective

Ranking Loss The core idea is that the higher the quality of the candidate summary, the closer to the document. We introduce a ranking loss to $f(\cdot)$:

$$\begin{aligned}\mathcal{L}_{rank} = \sum_i \sum_{j>i} max(0, f(C_j, D) - f(C_i, D) \\ +(-cost(C_i, S) + cost(C_j, S)) * \lambda)\end{aligned} \tag{5}$$

where S is the reference summary and λ is the hyper-parameter.[1] Here, $cost(C_i, S) = 1 -$

[1]We set λ to 1.0 on CNN/DM and 0.1 on XSum.

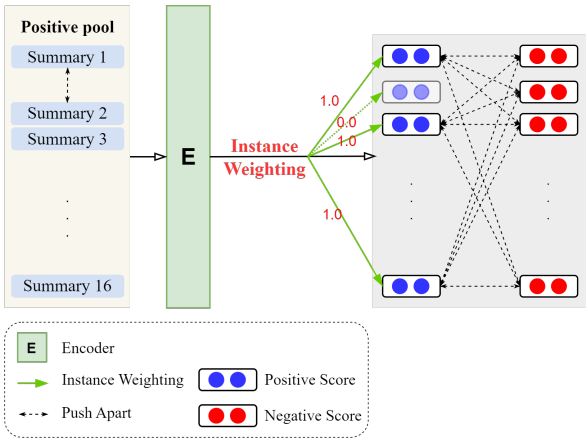

Figure 3: Overview of our proposed training objective.

$M(C_i, S)$ is the margin, and M is the automatic evaluation metric. We define it as ROUGE. We use the same metric in previous work (Liu and Liu, 2021; Liu et al., 2022), but the difference is that our loss directly reflects the quality measure during training. In other words, the quality was not properly reflected before because different margin $((j - i) * \lambda)$ was assigned even if the candidate summaries had the same ROUGE score.

Contrastive Loss with Instance Weighting The construction of positive and negative pairs is the critical point in constrative learning. Therefore, we consider generated summaries from the same document as *positive samples* and irrelevant summaries from other documents as *negative samples*. Thus, we design a set of candidate summaries $\mathbb{C}$ in Eq. 1 as *positive* and a set of randomly sampled summaries N as *negative*.[2] To identify summaries whose meanings are close to the document, we introduce a contrastive learning objective with instance weighting:

$$\mathcal{L}_{ctr} = \frac{1}{|\mathbb{C}|} \sum_{C_i \in \mathbb{C}} -log \frac{\alpha_{C_i} \times e^{f(C_i, D)}}{e^{f(C_i, D)} + \sum_{s_i \in N} e^{f(s_i, D)}} \tag{6}$$

We newly define summaries that have a high lexical matching but a low semantic similarity as *false positives*. Inspired by Zhou et al. (2022), we design an instance weighting method to reduce the influence of false positives. We produce the weights for positives using the SimCSE (Gao et al., 2021) which is the state-of-the-art model for the sentence representation task:

$$\alpha_{C_i} = \begin{cases} 0, & sim(C_i, S) < \phi \\ 1, & sim(C_i, S) \geq \phi \end{cases} \tag{7}$$

where ϕ is a hyper-parameter of the instance weighting threshold, and $sim(\cdot)$ is the cosine similarity score evaluated by the SimCSE model.

Finally, as shown in Fig. 3, we combine the ranking (Eq. 5) and contrastive (Eq. 6) losses:

$$\mathcal{L} = \gamma_1 \mathcal{L}_{rank} + \gamma_2 \mathcal{L}_{ctr} \tag{8}$$

where γ is the scale factor of each loss and we find the optimal values ($\gamma_1 = 10, \gamma_2 = 0.1$) in Appendix H.

3 Experiments

3.1 Datasets

We experiment on two datasets, whose statistics are shown in Appendix C.

CNN/DailyMail (Hermann et al., 2015) is the most commonly used summarization dataset which contains articles from the CNN and DailyMail newspapers.

XSum (Narayan et al., 2018) is a one-sentence summary dataset from the British Broadcasting Corporation (BBC) for the years 2010 - 2017.

3.2 Training Details

We use diverse beam search (Vijayakumar et al., 2016) to generate 16 candidate summaries. We start from pre-trained checkpoints of RoBERTa-base (Liu et al., 2019). We train BalSum for five epochs. It takes 33 hours on CNN/DM and 22 hours on XSum on a single RTX 3090 GPU. More details are described in Appendix D.

3.3 Main Results

In terms of the two-stage framework, we compare our results with SimCLS (Liu and Liu, 2021), SummaReranker (Ravaut et al., 2022), and BRIO (Liu et al., 2022). We apply BalSum on top of each base model which is BART or PEGASUS.

The results on CNN/DM are described in Table 1. BalSum outperforms a base BART model, according to gains of 2.54/1.27/2.63 R-1/2/L. Notably, while it has comparable performances on ROUGE to previous models, it achieves an 89.67 BERTScore, reaching a new state-of-the-art performance. When ranking the candidate summaries, our model can estimate the meaning of summaries

[2] As it is insensitive, we fix a negative strategy to random sampling in our experiments.

Model	R-1	R-2	R-L	BS
BART*	44.16	21.28	40.90	-
BART‡	44.04	21.06	40.86	88.12
Pegasus*	44.16	21.56	41.30	-
BRIO-Mul*	47.78	23.55	44.57	-
BRIO-Mul‡	**47.50**	**23.48**	44.01	89.08
BRIO-Ctr*	47.28	22.93	44.15	-
BRIO-Ctr‡	47.08	23.03	**44.06**	89.03
SummaReranker*	47.16	22.55	43.87	87.74
SimCLS*	46.67	22.15	43.54	-
SimCLS‡	46.34	22.07	43.30	88.92
BalSum	46.58†	22.33†	43.49†	**89.67**†

Table 1: **Results on CNN/DM**. R-1/2/L are the ROUGE-1/2/L F_1 scores. **BS** denotes BERTScore. *: results reported in the original papers. ‡: results from our own evaluation script. †: significantly better than the baseline model (BART).

Model	R-1	R-2	R-L	BS
BART*	45.14	22.27	37.25	-
Pegasus*	47.21	24.56	39.25	-
Pegasus‡	46.82	24.44	39.07	91.93
BRIO-Mul*	49.07	25.59	40.40	-
BRIO-Mul‡	**48.74**	**25.38**	**40.16**	**92.60**
BRIO-Ctr*	48.13	25.13	39.84	-
BRIO-Ctr‡	48.12	25.24	39.96	91.72
SummaReranker*	48.12	24.95	40.00	92.14
SimCLS*	47.61	24.57	39.44	-
SimCLS‡	47.37	24.49	39.31	91.48
BalSum	47.17†	24.23	39.09	91.48

Table 2: **Results on XSum**. R-1/2/L are the ROUGE-1/2/L F_1 scores. **BS** denotes BERTScore. *: results reported in the original papers. ‡: results from our own evaluation script. †: significantly better than the baseline model (PEGASUS).

without seriously degrading the lexical aspect. We argue that this is because BalSum decreases more false positives than other ranking models. We provide fine-grained analyses for this result and present a case study in Sec.3.4.

In addition, we apply our method on XSum, as shown in Table 2. Though we use a different strategy to generate the validation and test data [3], our method improves a base PEGASUS with a small margin. We believe the one of reasons is that XSum is restricted to capturing diverse semantic units because it consists of much shorter summaries (one-sentence) than CNN/DM.

[3]We use 4 groups for diversity sampling, which results in 4 candidates. This is the same as SimCLS.

ϕ	N/A	0.7	0.75	0.8	0.85	0.9
BS	89.37	89.35	89.36	89.63	89.37	**89.67**

Table 3: BERTScore (noted **BS**) results with different weighting threshold ϕ on CNN/DM. "N/A": no instance weighting.

Model	BS@1	BS@3	BS@5	R@1	R@3	R@5
Oracle (R)	90.77	90.42	90.18	44.85	42.68	41.16
Oracle (BS)	91.06	90.66	90.38	43.32	41.46	40.18
SimCLS	88.92	88.87	88.82	37.24	36.95	36.65
BRIO-Ctr	89.03	88.93	88.85	**38.06**	**37.55**	**37.14**
BalSum	**89.67**	**89.60**	**89.54**	37.46	37.08	36.78

Table 4: Analysis of re-ranking performance on CNN/DM. **BS** and **R** denote BERTScore and the mean ROUGE F_1 score, respectively. Oracle (R) is ordered by ROUGE scores, while Oracle (BS) is ordered by BERTScore.

3.4 Analysis

Weighting Threshold ϕ Intuitively, the larger the weighting threshold, the lower false positives. We train our model with different instance weighting thresholds from 0.7 to 0.9. In Table 3, the highest threshold ($\phi = 0.9$) shows the best performance and it rises largely to 0.3 BERTScore compared to when not applied. We also find that increasing the threshold leads to performance improvement. Therefore, we demonstrate that false positives can be considered noise in training.

Ranking Evaluation Regardless of the number of candidates, an ideal ranking model should yield oracle results considering diverse aspects of summarization. We conduct an experiment to measure the qualities by selecting the top-k summaries after aligning the candidates through different models. As shown in Table 4, we can see that our model shows consistent performance in both evaluation metrics depending on the k (about ± 0.06 BERTScore, ± 0.34 ROUGE average score). Compared to SimCLS and BRIO-Ctr, the second block in Table 4 demonstrates that BalSum captures semantic similarity best while maintaining the intermediate level from the perspective of lexical overlap quality. Moreover, we find that BalSum has the lowest drop ratio of BERTScore (-1.52%) from the perfect ranking "oracle" scores.

We also investigate whether all ranked summaries by models satisfy both lexical and semantic quality. We evaluate models using F_1 which measures the cases where the higher-ranked summary

	CNNDM		XSum	
Model	F_1	**FP**(%)	F_1	**FP**(%)
BRIO-Ctr	78.50	10.96	**76.95**	10.01
BalSum	**78.84**	10.73	76.32	10.49

Table 5: F_1 score and percentage of false positives on all two datasets. The high F_1 score indicates how well the ranking model estimates both lexical and semantic quality of all candidate summaries in the pool. **FP** stands for false positives.

has both larger ROUGE and BERTScore than the lower-ranked summary. In addition, we calculate the percentage of false positives. Following Table 5, while BalSum has worse ($+0.48\%$ FP, -0.63 F_1) than BRIO-Ctr on XSum, it has better ranking performance (-0.23% FP, $+0.34$ F_1) on CNN/DM. We observe that the decrease of false positives leads to an improvement in F_1 score, demonstrating that the result of Table 1 can be interpreted as reducing semantic mistakes in ranking. As a result, we find that (1) our model is able to learn how to score each summary by balancing the lexical and semantic quality, and (2) the other reason of weak performance on XSum is related to small decline of false positives compared to CNN/DM.

Case Study on CNN/DM Table 10 presents an intriguing pattern we observed when comparing the results of BRIO-Ctr and BalSum, which demonstrate that our model helps to capture precise details from documents. While BRIO-Ctr contains some irrelevant information in the summaries (shown as highlighted text in blue), BalSum selects the summaries where the last sentence is more consistent with the reference (shown as highlighted text in yellow). Furthermore, despite the comparable ROUGE scores of both models, we note that BalSum's selected summaries consistently have higher BERTScore than those of BRIO-Ctr.

4 Conclusion

In this work, we propose BalSum which aims to evaluate summaries by considering the balance between lexical and semantic quality. To achieve this, we perform a multi-task learning, which aligns summaries according to their lexical overlap qualities and identifies whether they are similar to the document. In addition, to our best knowledge, our method is the first attempt to present a new perspective of false positives (semantic mistakes) in ranking and creating the model to reduce their influence. Our experimental results and fine-grained analyses validate that our model achieves consistent improvements over competitive baselines.

Limitations

Candidate Summaries Dependency While we mainly investigate a training objective to select the best summary among a set of candidates, we find that our model has been dependent on those obtained from the generation model. Recently, several works have been presented to improve language generation. For example, Narayan et al. (2022) and Xu et al. (2022) improve decoding methods to generate diverse outputs. It will be beneficial when applying our method to these approaches.

One-sentence Summary Our approach can fail to capture the information from an extremely short summary. Since Table 2 shows that our approach has a smaller improvement than CNN/DM, we plan to investigate that our model aims to capture more detailed features from an input text.

Acknowledgements

We thank Soohyeong Kim and anonymous reviewers for valuable feedback and helpful suggestions. This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(*MSIT) (No.2018R1A5A7059549 , No.2020R1A2C1014037) and supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(*MSIT) (No.2020-0-01373). *Ministry of Science and ICT

References

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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, volume 28. Curran Associates, Inc.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. SIGIR '20, page 39–48, New York, NY, USA. Association for Computing Machinery.

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.

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.

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. 2021. 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.

Yixin Liu and Pengfei Liu. 2021. SimCLS: A simple framework for contrastive learning of abstractive summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1065–1072, Online. Association for Computational Linguistics.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. BRIO: Bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903, Dublin, Ireland. Association for Computational Linguistics.

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

Shashi Narayan, Gonçalo Simões, Yao Zhao, Joshua Maynez, Dipanjan Das, Michael Collins, and Mirella Lapata. 2022. A well-composed text is half done! composition sampling for diverse conditional generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1319–1339, Dublin, Ireland. 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*.

Mathieu Ravaut, Shafiq Joty, and Nancy Chen. 2022. SummaReranker: A multi-task mixture-of-experts re-ranking framework for abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4504–4524, Dublin, Ireland. 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, Vancouver, Canada. Association for Computational Linguistics.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4596–4604. PMLR.

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.

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.

Jiacheng Xu, Siddhartha Jonnalagadda, and Greg Durrett. 2022. Massive-scale decoding for text generation using lattices. In *Proceedings of the 2022*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4659–4676, Seattle, United States. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. 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. 2019. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.

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.

Kun Zhou, Beichen Zhang, Xin Zhao, and Ji-Rong Wen. 2022. Debiased contrastive learning of unsupervised sentence representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6120–6130, Dublin, Ireland. Association for Computational Linguistics.

A Distribution of z on XSum

The result in Fig. 4 shows that there is a majority (53%) of cases where $z > 1$.

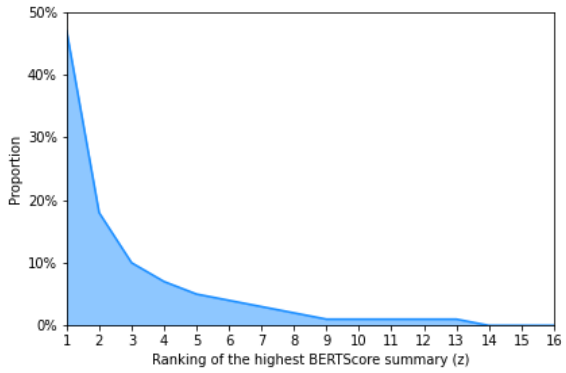

Figure 4: Distribution of z(%) for a base PEGASUS model on XSum. Because a PEGASUS model generates a pool of 16 diverse beam search candidates, the X-axis ranges from 1 to 16. The Y-axis represents the proportion of z in the test set.

B Evaluation Metrics

We examine our model with two evaluation metrics.

- **ROUGE** (Lin, 2004) is a widely used metric for summarization evaluation. We use the standard ROUGE Perl package[4] for evluation.
- **BERTScore** (Zhang et al., 2019) is a semantic similarity metric for multiple tasks. We use the public *bert-score* package[5] shared by the authors.

C Datasets Statistics

Dataset	Train	Valid	Test
CNN/DM	287,227	13,368	11,490
XSum	204,045	11,332	11,334

Table 6: Statistics of two datasets

D Implementation Details

Model We implement our model based on Huggingface Transformers library (Wolf et al., 2020). We use the pre-trained RoBERTa with 'roberta-base' version, containing around 125M parameters. Our experiments are conducted on a single NVIDIA RTX 3090 GPU with 24GB memory.

Decoding Setttings We use the diverse beam search algorithm (Vijayakumar et al., 2016) to decode summaries. We generate candidate summaries from 16 diversity groups with 16 beams. On CNN/DM and XSum, we use the pre-trained BART[6] and PEGASUS[7] models as the generation model.

Training Settings We train our models for 5 epochs using an Adafactor optimizer (Shazeer and Stern, 2018). The batch size is 4 and the learning rate is 2e-3. During training, we randomly select 4 negative samples for each input document. We evaluate the model every 1000 steps on the validation set.

[4] https://github.com/summanlp/evaluation/tree/master/ROUGE-RELEASE-1.5.5

[5] https://github.com/Tiiiger/bert_score

[6] The checkpoint is "facebook/bart-large-cnn", containing around 400M parameters.

[7] The checkpoint is "google/pegasus-xsum", containing around 568M parameters.

E Effect of Model Architecture

We train BalSum with different model architectures and evaluate them on CNN/DM test set. For a fair comparison, we use only ranking loss in Eq. 5. Table 7 shows that taking the weighted sum of scores in Eq. 4 leads to better performance than others.

Model	R-1	R-2	R-L
[CLS]	45.40	21.18	42.36
Avg.	46.59	**22.40**	43.47
Ours	**46.64**	22.38	**43.52**

Table 7: Ablation studies of different model architectures on CNN/DM. **R-1/2/L** denotes ROUGE-1/2/L. *[CLS]*: using the first [CLS] token. *Avg.*: averaging all scores in Eq. 3.

F Identical Candidates Scores

As shown in Table 8, we note cases that have at least two identical R-avg on CNN/DM and XSum are a majority. Since we count after removing the same summaries in the pool, we ensure that it is the number of summaries with different content but the same R-avg score.

Dataset	Decoding methods	# Summary candidates	# of pools with at least two same R-avg (%)
CNN/DM	Diverse beam search	16	46.09
Xsum	Diverse beam search	16	73.01

Table 8: Number of pools with at least two same R-avg (%). A pool consists of 16 diverse beam search candidates generated on different datasets (CNN/DM, XSum) with different base models (PEGASUS, BART). R-avg is the average of ROUGE-1/2/L scores.

G Examples for False Positive

Table. 9 shows that #2 has 2.33 R-avg lower than #1, but 3.67 BERTScore higher. Also, when evaluated qualitatively, it can be seen that #2 is closer to the gold summary. While the sentence in green is discarded, the sentence in red is included in the reference summary.

H Negative Size and Scale Factors

We have tuned the scale factor γ_1 of ranking loss and γ_2 of contrastive loss in Eq. 8 with different sizes of negative samples. As shown in Fig. 5, suitable scale factors ($\gamma_1 = 10, \gamma_2 = 0.1$) can improve more than others. Though $size = 4$ and $size = 12$ showed similar performance, we set the negative size to 4 due to memory efficiency.

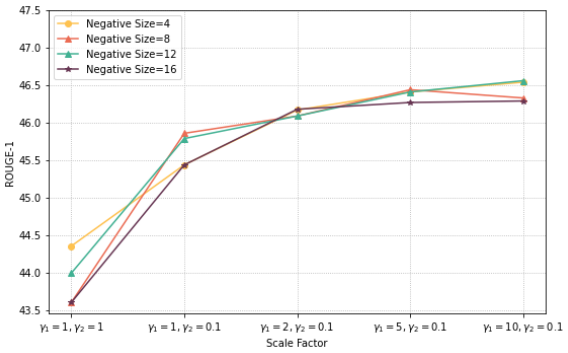

Figure 5: ROUGE-1 on CNN/DM w.r.t scale factors and N negative samples at inference time, with $N \in \{4, 8, 12, 16\}$.

I Number of Candidate Summaries

We set the size of the candidate summary pool to 16, as it is close to the maximum which could fit in a standard 24GB RAM GPU. Fig. 6 reports that our method is robust to the number of candidates.

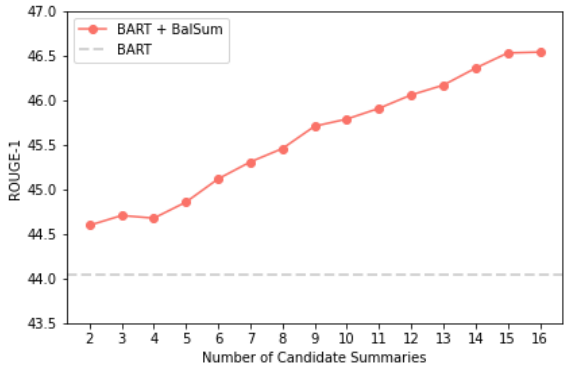

Figure 6: ROUGE-1 with different numbers of candidate summaries on CNN/DM. The gray dashed line denotes the performance of a base model (BART).

System	R-avg	BS	Summary
Reference	—	—	Didier Drogba played first Chelsea game after joining on free from Galatasaray. Ivory Coast striker was second half substitute for Diego Costa in 3-0 defeat by Werder Bremen. John Terry handed him captaincy later in game, **but 36-year-old failed to trouble German side in front of goal.**
Diverse beam #1	30.72	87.50	Ivory Coast striker made his second return to the club. Drogba was a half-time substitute in the 3-0 defeat at the Weserstadion. The 36-year-old was replaced by Diego Costa at half- time. Dobar was the first player on the pitch after John Terry left.
Diverse beam #2	28.39	91.17	Didier Drogba made his second Chelsea debut in pre-season friendly at Werder Bremen. The 36-year-old was a half-time substitute as Chelsea lost 3-0. Drogbba was captain after John Terry left the pitch in the second half. The Ivorian striker missed a penalty and failed to make an impact on the game.

Table 9: False positive examples from fine-tuned BART model on CNN/DM. **R-avg** is the average of ROUGE-1/2/L scores. **BS** denotes BERTScore. The related sentences in the reference are in **bold**.

System	R-1	R-2	R-L	BS	Summary
Reference	-	-	-	-	arsene wenger will have chat with theo walcott ahead of arsenal clash. walcott was substituted after 55 minutes of england's draw with italy. arsenal boss is wenger is concerned by the winger's confidence. **the gunners take on liverpool at the emirates stadium on saturday.**
BRIO-Ctr	60.61	41.24	46.46	89.93	theo walcott played just 55 minutes of england's 1-1 draw with italy. arsene wenger says he is concerned by the winger's confidence. the arsenal manager will speak with walcott ahead of liverpool clash. walcott could start against liverpool on saturday with alex oxlade-chamberlain out and danny welbeck a doubt.
BalSum	61.54	38.20	41.76	92.36	arsenal winger theo walcott struggled for england against italy. arsene wenger says he is concerned by the winger's confidence. walcott was replaced after 55 minutes of england's 1-1 draw in turin. the gunners face liverpool on saturday in a top-four clash.
Reference	-	-	-	-	experts have voiced concerns over diy brain stimulation kits for children. for a few hundred dollars, one can be purchased online from various sites. it promises to help children with math homework and claims to help adhd. professor colleen loo from the black dog institute strongly believes that the equipment poses a danger to amateurs and children. **the equipment is currently being used to treat people with speech impediments but is still very much in trial stages.**
BRIO-Ctr	40.0	16.26	19.20	87.11	for a few hundred dollars, you can purchase a brain stimulation kit online. experts have voiced concerns over the potential side effects. the kits are being sold online for as little as $ 55 us. one site even advertises how to make your own electrodes using a household sponge.
BalSum	36.92	17.19	27.69	89.90	parents are buying diy brain stimulation kits for their children. the kits are being sold online for as little as $ 55 us. experts are concerned about the potential side effects of the equipment. the devices are used to improve speaking in those with speech problems. the equipment is still relatively new and experimental.
Reference	-	-	-	-	ross barkley has been repeatedly linked with a move to manchester city. former city star gareth barry says his everton team-mate is too young. **the toffees face manchester united in the premier league on sunday.**
BRIO-Ctr	47.19	27.59	29.21	88.85	everton team-mate gareth barry has advised ross barkley against moving to manchester city. the 21-year-old has been linked with a move away from goodison park. barry believes it is too early for the youngster to decide on his future. the veteran midfielder spent four seasons at the etihad before joining everton.
BalSum	46.34	25.0	34.15	91.16	gareth barry has advised ross barkley against moving to manchester city. the everton midfielder believes it is too early for the 21-year-old to decide on his future. barry spent four seasons at the etihad before arriving on merseyside. the toffees face manchester united on sunday.
Reference	-	-	-	-	local councils are urged to draw up maps of the residents who are at risk. **essex and gloucestershire have already made 'loneliness maps' experts warn that being lonely can lead to serious health problems.**
BRIO-Ctr	50.57	28.24	29.89	90.30	two county councils have already implemented 'loneliness maps' to target 'danger zones' being lonely can lead to health problems including dementia and high blood pressure. campaigners say councils should draw up maps of the places where pensioners are most at risk. study by university of kent and campaign to end loneliness recommends maps.
BalSum	50.0	27.91	43.18	91.28	campaigners say councils should draw up maps of places where pensioners and others are most likely to suffer from social isolation. two county councils, essex and gloucestershire, have already implemented the maps. they allow them to target 'danger zones' of loneliness. being lonely can lead to health problems including dementia and high blood pressure.
Reference	-	-	-	-	the gruesome vision was captured in australia and uploaded last week. the lizard swings its neck back and forth in a bid to swallow the rabbit. **goannas can unhinge their lower jaws allowing them to swallow large prey.**
BRIO-Ctr	51.16	23.81	27.91	88.75	two-metre long reptile is filmed balancing on top of a power pole to swallow rabbit. the lizard swings its neck back and forth as it battles to swallow its catch. it finishes the feat in under a minute, and the video was uploaded to youtube last week.
BalSum	46.91	20.25	34.57	90.72	two-metre long lizard filmed battling to swallow rabbit in under one minute. video shows lizard balance at the top of a power pole while swallowing its prey. goannas can unhinge their lower jaws when feeding, allowing them to eat over-sized prey.

Table 10: **Case Study** on CNN/DM. R-1/2/L are the ROUGE-1/2/L F_1 scores. **BS** denotes BERTScore. The related sentences in the reference are in **bold**.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations

☑ A2. Did you discuss any potential risks of your work?
Limitations

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract, 1. Introduction

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

3.1 Datasets, 3.2 Training Details, Appendix B. Evaluation Metrics, Appendix D. Implementation Details

☑ B1. Did you cite the creators of artifacts you used?
3.1 Datasets, 3.2 Training Details, Appendix B. Evaluation Metrics, Appendix D. Implementation Details

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
3.1 Datasets, 3.2 Training Details, Appendix B. Evaluation Metrics, Appendix D. Implementation Details

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
3.1 Datasets, 3.2 Training Details, Appendix D. Implementation Details

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
3.1 Datasets

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Appendix C. Datasets Statistics

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ Did you run computational experiments?

3.2 Training Details, Appendix D. Implementation Details

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix H. Negative Size and Scale Factors

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
3.2 Training Details, Appendix D. Implementation Details, Appendix H. Negative Size and Scale Factors

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
3.3 Main Results, 3.4 Analysis

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix B. Evaluation Metrics, Appendix D. Implementation Details

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.