SummaReranker: A Multi-Task Mixture-of-Experts Re-ranking Framework for Abstractive Summarization

Mathieu Ravaut[♣♢], **Shafiq Joty**[*♣♠], **Nancy F. Chen**[*♢]
[♣] Nanyang Technological University, Singapore
[♢] Institute of Infocomm Research (I2R), Singapore
[♠] Salesforce Research Asia, Singapore
{mathieuj001@e.ntu, srjoty@ntu}.edu.sg
nfychen@i2r.a-star.edu.sg

Abstract

Sequence-to-sequence neural networks have recently achieved great success in abstractive summarization, especially through fine-tuning large pre-trained language models on the downstream dataset. These models are typically decoded with beam search to generate a unique summary. However, the search space is very large, and with the exposure bias, such decoding is not optimal. In this paper, we show that it is possible to directly train a second-stage model performing *re-ranking* on a set of summary candidates. Our mixture-of-experts SummaReranker learns to select a better candidate and consistently improves the performance of the base model. With a base PEGASUS, we push ROUGE scores by 5.44% on CNN-DailyMail (47.16 ROUGE-1), 1.31% on XSum (48.12 ROUGE-1) and 9.34% on Reddit TIFU (29.83 ROUGE-1), reaching a new state-of-the-art. Our code and checkpoints will be available at https://github.com/ntunlp/SummaReranker.

1 Introduction

In recent years, sequence-to-sequence neural models have enabled great progress in abstractive summarization (See et al., 2017; Lin et al., 2021). In the news domain, they have surpassed the strong LEAD-3 extractive baseline. With the rise of transfer learning since BERT (Devlin et al., 2019), leading approaches typically fine-tune a base pre-trained model that either follows a general text generation training objective like T5 (Raffel et al., 2019), BART (Lewis et al., 2020), ERNIE (Zhang et al., 2019b) and ProphetNet (Qi et al., 2021), or an objective specifically tailored for summarization like in PEGASUS (Zhang et al., 2020).

Decoding methods	**# Summary candidates**	**R-1**	**R-2**	**R-L**	**BS**	**BaS**
Beam search (top beam)	1	44.23	21.48	41.21	87.39	-2.78
Beam search	15	51.06	27.74	48.05	88.50	-2.48
Diverse beam search	15	**54.30**	**30.02**	**51.33**	**88.97**	**-2.40**
Top-k sampling	15	52.31	27.41	49.17	88.64	-2.56
Top-p sampling	15	53.52	28.88	50.46	88.87	-2.46
Adding all four methods above	60	**57.70**	**33.77**	**54.72**	**89.58**	**-2.25**

Table 1: **Oracle scores** (maximum scores over all generated candidates) for four popular decoding methods and five summarization evaluation measures for a base PEGASUS model on **CNN/DM**. **R-1/2/L** denotes ROUGE-1/2/L, **BS** and **BaS** denote BERTScore and BARTScore, respectively.

Most of these sequence-to-sequence models are history-based, where an output sequence is represented as a sequence of decisions and the probability of the sequence is computed as a product of decision probabilities. This is also known as the auto-regressive factorization. To transform the sequence of probabilities into summaries, beam search is commonly used. While auto-regressive decoding with beam search is simple and has many advantages, it can be difficult to encode global constraints such as grammaticality, coherence and factual consistency within this framework, properties that are believed to be useful in discriminating among candidate outputs. If the model starts decoding in a bad direction, mistakes might propagate, carry over the mistake of previous tokens to the generation of new ones, and the model has no way to know that it should adjust the decoding. Furthermore, these models are typically trained with teacher forcing (Williams and Zipser, 1989), which leads to an inherent discrepancy between training time and inference time known as the exposure bias problem (Bengio et al., 2015; Sun and Li, 2021).

Decoding methods such as beam search maintain a list of top-k best candidates, and output a single best one. In the case of beam search, candidates are sorted by decreasing log-probability, and the last $(k-1)$ hypotheses are discarded. However, these $(k-1)$ other hypotheses often contain considerably better sequences in terms of different evaluation measures. This observation holds over other decoding methods: diverse beam search (Vi-

[*] Equal contribution.

jayakumar et al., 2016), top-k sampling (Fan et al., 2018) and top-p sampling (Holtzman et al., 2019). In Table 1, we illustrate this phenomenon with the *oracle* scores (maximum scores over the pool of candidates) for four popular decoding methods and five metrics on the CNN-DailyMail (Hermann et al., 2015) dataset with a PEGASUS model. The oracle ROUGE-1 scores are up to 10 points higher (+22.8%) than the top beam baseline. Moreover, oracle gains significantly increase when mixing several generation methods together, reaching an improvement of more than 13 ROUGE-1 points (+30.5%). Such a gap is larger than the progress made by research in the whole field of neural abstractive summarization in the last five years (Nallapati et al., 2016; Dou et al., 2021). This suggests that current abstractive models are not exploited to their full capacity, calling for better methods to identify the best summary candidate.

Given this assessment, we investigate whether it is possible to train a *second-stage* summarization model which learns to select the best summary among a set of candidates obtained from a base model and with a decoding process, which itself can potentially involve a set of decoding methods (e.g., beam search variants). This way, the model would recover the gap that separates it with the oracle. This raises the question of what makes a summary candidate the *optimal* one? Admittedly, summarization has been an underconstrained task and its evaluation is complex and remains an active research area (Kryscinski et al., 2019; Fabbri et al., 2021; Koto et al., 2021). To build a flexible approach, we use a multi-task learning framework based on a mixture-of-experts architecture in order to optimize *jointly* over several measures.

To design a robust re-ranker, we systematically explore the dimensions of summary re-ranking: base model, decoding process, and evaluation measure. Our system, named *SummaReranker*, is flexible and multi-task: it can be trained with any set of evaluation metrics. It is considerably less computationnally expensive to train than the single-stage summarization models that it is plugged on. We apply our system across three different datasets {CNN-DailyMail, XSum, Reddit TIFU} and two base models {PEGASUS, BART}. Optimizing ROUGE metrics leads to relative performance improvements from 1.31% to 9.34% depending on the dataset. It outperforms recently proposed second-stage summarization approaches RefSum (Liu et al., 2021) and SimCLS (Liu and Liu, 2021) and sets a new state-of-the-art on CNN-DailyMail and XSum (Narayan et al., 2018). We present extensive quantitative results coupled with a qualitative human evaluation.

2 Related Work

Re-ranking has been adopted in several branches of NLP for long. In syntactic parsing, Collins and Koo (2005) were the first to employ a re-ranker on the outputs of a base parser, followed by Charniak and Johnson (2005), who used a Maximum Entropy re-ranker. Passage re-ranking is used as the first stage of question-answering systems, to retrieve relevant passages where the answer might lay (Kratzwald and Feuerriegel, 2018; Nogueira and Cho, 2019). Some recent question-answering models also propose to perform answer re-ranking, to refine the answer selection (Kratzwald et al., 2019; Iyer et al., 2021). Re-ranking has also been used in neural machine translation. Checkpoint reranking (Pandramish and Sharma, 2020) generates several translation candidates with multiple model checkpoints, based on the observation (similar to the one we made in §1) that the oracle across checkpoints is of higher quality than just the last checkpoint. Bhattacharyya et al. (2021) use an energy-based model on top of BERT to select translation candidates with higher BLEU score.

In abstractive summarization, second-stage approaches such as re-ranking remain underexplored. Recently, RefSum (Liu et al., 2021) defined a second-stage summarization framework which helps address the problem of the train-test distribution mismatch in second-stage models. With a base GSum model (Dou et al., 2021), the authors reach a 46.18 state-of-the-art ROUGE-1 on CNN-DailyMail. In SimCLS (Liu and Liu, 2021), the authors train a second-stage model with contrastive learning, using a ranking loss to select the best summary candidate from a pool of 16 diverse beam search candidates, reaching 46.67 ROUGE-1 on CNN-DailyMail. Our approach differs from RefSum and SimCLS in terms of model architecture and loss function, as well as summary candidate generation process. In contrast with RefSum, we use a single base model, but mix several decoding methods, as our goal is single-model improvement. Unlike SimCLS, we do not use a ranking loss, but directly model the probability that a summary candidate is the best one. To the best of our knowledge, we are the first ones to propose a *multi-task* re-ranking system for abstractive summarization. This enables practitioners to leverage the recent rich literature in automatic abstractive summarization evaluation (Lin, 2004; Zhang et al., 2019a; Zhao et al., 2019a; Yuan et al., 2021).

3 Model

3.1 Re-ranking Framework

Our approach follows the paradigm of second-stage models. Specifically, given a source document S, a base model B, and a set of decoding methods $\mathbb{D}$, we get a pool of m summary candidates $\mathbb{C} = \{C_1, \dots, C_m\}$. Given an evaluation metric μ in a set of metrics $\mathbb{M}$, we get associated scores for each candidates $\mathbb{S}_\mu = \{\mu(C_1), \dots, \mu(C_m)\}$. Our goal is to train a model f_θ parameterized by θ to explicitly identify the best summary candidate C^*_μ according to the metric, which is given by:

$$C^*_\mu = \underset{C_i \in \mathbb{C}}{\arg\max}\ \{\mu(C_1), \dots, \mu(C_m)\} \quad (1)$$

We frame this problem as a binary classification. C^*_μ is the positive candidate, while other candidates are treated as negative. For a metric μ, the re-ranker f_θ is trained with a binary cross-entropy loss:

$$\mathcal{L}_\mu = -y_i \log p^\mu_\theta(C_i) - (1 - y_i) \log(1 - p^\mu_\theta(C_i)) \quad (2)$$

where $y_i = 1$ if $C_i = C^*_\mu$, otherwise $y_i = 0$.

Binary classification has been successfully employed for re-ranking in prior work (Nallapati, 2004; Nogueira and Cho, 2019). While multi-way classification could be an alternative, we noticed that for each generation method, a significant fraction of candidates share the same score for one or several metrics, while it is rare that *all* candidates share the same score (Appendix C-D). Thus, there is not enough signal to distinguish m candidates into m different classes, but enough for two classes.

To optimize for N different metrics $\mathbb{M} = \{\mu_1, \dots, \mu_N\}$ simultaneously, we use a separate prediction head (tower) for each and we minimize the average over metric losses defined as:

$$\mathcal{L} = \frac{1}{N} \sum_{\mu \in \mathbb{M}} \mathcal{L}_\mu \quad (3)$$

3.2 Model Architecture

We first need to get a good representation of the summary candidate. To use contextual information, we concatenate the source with the candidate, separating the two with a special token: `[CLS] Source [SEP] Candidate`, and feed it to a pre-trained language model. In all experiments, we use RoBERTa-large (Liu et al., 2019) as encoder. Concatenating the source with the candidate enables RoBERTa to perform cross-attention between the two, which finds parts of the source relevant to the summary candidate. We take the `[CLS]` representation from RoBERTa's last layer, and feed it to a multi-layer perceptron (MLP).

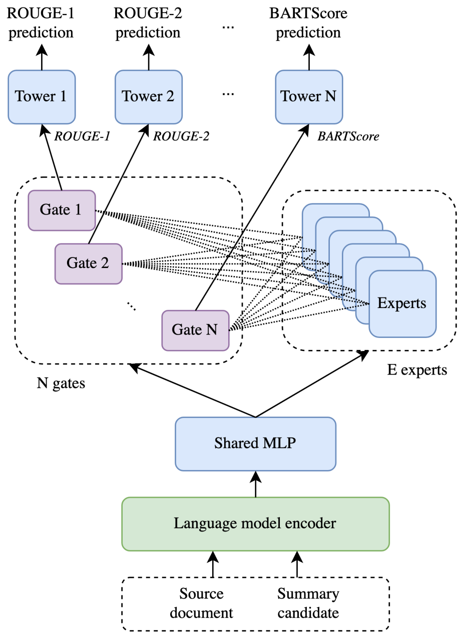


Figure 1: **SummaReranker model architecture**, optimizing N metrics. The summarization metrics here (*ROUGE-1*, *ROUGE-2*, ..., *BARTScore*) are displayed as examples.

Once we have a joint representation of the source with the candidate (noted $\boldsymbol{x}$), we perform multi-task learning in order to optimize for the desired metrics. Since metrics are different, yet may be strongly correlated (e.g., ROUGE variants), we adopt a mixture-of-experts (MoE) architecture. In particular, we follow the sparse MoE approach (Shazeer et al., 2017), which introduces experts dropout. To adapt it to multi-task training, we use the multi-gate approach proposed in Zhao et al. (2019b). Given E experts $\mathcal{E}_1, \dots, \mathcal{E}_E$ and N prediction towers $\mathcal{T}_1, \dots, \mathcal{T}_N$, the prediction for an input summary representation $\boldsymbol{x}$ for a metric μ indexed by $k \in \{1, \dots, N\}$ is:

$$f^k_\theta(\boldsymbol{x}) = \mathcal{T}_k\Big(\sum_{i=1}^{E} \text{softmax}(\boldsymbol{W}_k \boldsymbol{x})_{(i)} \mathcal{E}_i(\boldsymbol{x})\Big) \quad (4)$$

where $\boldsymbol{W}_k$ is the weight matrix associated with gate

k. The corresponding prediction probability is:

$$p_{\theta}^{\mu} = \text{sigmoid}(f_{\theta}^{k}(\boldsymbol{x})) \tag{5}$$

Experts are shared across all tasks, and through the softmax gates the model learns how much weight to assign to each expert for each task.

Our SummaReranker model architecture is shown in Fig. 1. In practice, the shared bottom MLP consists in two fully-connected layers with ReLU activation (Glorot et al., 2011). Each expert $\mathcal{E}_i$ is also a two-layer MLP with ReLU, and each prediction tower $\mathcal{T}_k$ is a single-layer MLP. We set the number E of experts to be equal to twice the number of tasks (N), and the experts dropout to 50%, so that the effective number of experts being used during training matches N. Our model has 370.09 million trainable parameters, representing a slight 4.14% increase due to the mixture-of-experts compared to the off-the-shelf RoBERTa-large.

3.3 Tackling Training and Inference Gap

Second-stage learning approaches may suffer from an inherent distribution bias. Indeed, the base model has a different output distribution on the training set than on the validation and test sets. Thus, it is ineffective to train a second-stage model on the training set outputs of the base model.

To resolve this distribution shift, we shuffle the training set and randomly split it into equal parts, then fine-tune a pre-trained model on each half. Then, to build a training set for the re-ranker, we infer with each model on the half that it was not trained on. At testing time, we face two options:

- **Base setup**: in this setup, we infer on the test set with *one of the two base models trained on half the training set*, then apply the re-ranker. Since the base models are trained on less data, their performance on the test set worsens. However, we will show that SummaReranker brings improvements which more than compensate this performance drop.

- **Transfer setup**: this setup consists in applying SummaReranker on top of a base model *trained on the whole training set*. Note that SummaReranker is still trained in the same fashion as before. There could be a distribution mismatch in this setting too, since SummaReranker needs to rank summary candidates of a potentially higher quality (generated by a model trained on the full data) than the summaries that it was trained on (generated by a model trained on half the data). Nevertheless, SummaReranker still transfers well and considerably improves the performance of the base model in this transfer setup.

	R-1	**R-2**	**R-L**	**BS**	**BaS**
R-1	1.000	0.884	0.977	0.858	0.662
R-2	0.884	1.000	0.910	0.833	0.665
R-L	0.977	0.910	1.000	0.855	0.669
BS	0.858	0.833	0.855	1.000	0.682
BaS	0.662	0.665	0.669	0.682	1.000

Table 2: **Pearson correlation coefficient** between the five evaluation metrics {R-1, R-2, R-L, BS, BaS} for a base PEGASUS with beam search on **CNN/DM**. **R-1/2/L** denotes ROUGE-1/2/L, **BS** and **BaS** denote BERTScore and BARTScore.

If $\mathbb{D}$ is made of multiple decoding methods $\{\delta_1, ..., \delta_j\}$, each producing several candidates, the overall candidate set may be large, slowing down inference. Thus, to explore lower-resource inference setups, we separate the sets of decoding methods $\mathbb{D}_{\text{train}}$ and $\mathbb{D}_{\text{test}}$ used for training and inference, respectively, and enforce that $\mathbb{D}_{\text{test}} \subset \mathbb{D}_{\text{train}}$.

4 Experiments

4.1 Scope & Datasets

Throughout our experiments, we vary all the three dimensions of our re-ranking framework: the base model B, the set of decoding methods $\mathbb{D}$ and the set of scoring metrics $\mathbb{M}$.

As base models, we use PEGASUS (Zhang et al., 2020) and BART (Lewis et al., 2020), each one in their large version, as they are leading summarization models with publicly available checkpoints. We obtain pre-trained and fine-tuned checkpoints from the HuggingFace transformers library (Wolf et al., 2020).

For decoding methods ($\mathbb{D}$), we experiment with beam search (referred to as 1), diverse beam search (2), top-k sampling (3) and top-p sampling (4). For each decoding method, we set the number of candidates to 15, as it is close to the maximum which could fit in a standard 11GB RAM GPU when doing generation with PEGASUS-large.

As set of metrics, we first use ROUGE (Lin and Hovy, 2003), in its commonly used three flavours of ROUGE-1 (noted *R-1*), ROUGE-2 (noted *R-2*)

Dataset	**Domain**	**# Data points**			**# Words**	
		Train	Val	Test	Doc.	Summ.
CNN/DM	News	287,113	13,368	11,490	766.56	54.78
XSum	News	204,045	11,332	11,334	414.51	22.96
Reddit TIFU	Social media	33,704	4,213	4,222	385.59	20.59

Table 3: **Statistics** of the three datasets.

Model	**Model stage**	**Decoding methods ($\mathbb{D}$)**	**R-1**	**R-2**	**R-L**	**Gain (%)**
PEGASUS - 1st half	1	{1}	42.23	19.62	38.90	_
PEGASUS - 1st half	1	{2}	42.50	19.75	39.55	_
PEGASUS - 2nd half	1	{1}	42.46	19.95	39.19	_
PEGASUS - 2nd half	1	{2}	42.75	19.93	**39.86**	_
BART - 1st half	1	{1}	42.79	20.25	39.66	_
BART - 1st half	1	{2}	40.70	18.99	37.88	_
BART - 2nd half	1	{1}	**42.93**	**20.36**	39.73	_
BART - 2nd half	1	{2}	41.93	19.79	39.06	_
PEGASUS - 1st half **+ SR**	2	{1}	44.02	20.97	40.68	5.23
PEGASUS - 1st half **+ SR**	2	{2}	**45.66**	21.31	42.51	7.61
PEGASUS - 2nd half **+ SR**	2	{1}	44.11	21.08	40.82	4.57
PEGASUS - 2nd half **+ SR**	2	{2}	45.73	21.31	**42.62**	6.94
BART - 1st half **+ SR**	2	{1}	44.23	21.23	41.09	3.94
BART - 1st half **+ SR**	2	{2}	45.05	21.47	42.12	11.65
BART - 2nd half **+ SR**	2	{1}	44.51	21.52	41.29	4.44
BART - 2nd half **+ SR**	2	{2}	45.61	**21.78**	**42.62**	9.32
PEGASUS - 1st half **+ SR**	2	{1, 2}	46.12	21.97	42.84	9.36
PEGASUS - 2nd half **+ SR**	2	{1, 2}	**46.19**	22.02	**42.92**	8.70
BART - 1st half **+ SR**	2	{1, 2}	45.76	22.14	42.71	7.99
BART - 2nd half **+ SR**	2	{1, 2}	45.96	**22.18**	42.88	7.98

Table 4: **Base setup results** for SummaReranker applied to PEGASUS and BART on the **CNN/DM** dataset. **SR** refers to SummaReranker. **Decoding method {1}** is beam search, **{2}** is diverse beam search. Best scores for each type of model are in bold. **Gain** represents the mean relative gain over {R-1, R-2, R-L} compared to the best decoding method.

and ROUGE-L (noted *R-L*) for summarization evaluation. We also leverage recently introduced model based evaluation methods BERTScore (noted *BS*) (Zhang et al., 2019a) and BARTScore (noted *BaS*) (Yuan et al., 2021), which both rely on contextual word embeddings from pre-trained language models. Thus, our total set of metrics is $\mathbb{M}$ = {R-1, R-2, R-L, BS, BaS}. As seen in Table 2, R-1 and R-L are strongly correlated (Pearson correlation score of 0.977). BARTScore is the least correlated to other metrics, suggesting that it captures aspects complementary to the other four.

We train SummaReranker on the following datasets, covering multiple domains:

- **CNN-DailyMail** (Hermann et al., 2015) contains 93k and 220k articles from the CNN and DailyMail newspapers, respectively. We use the non anonymized version from (See et al., 2017).
- **XSum** (Narayan et al., 2018) contains 227k articles from the BBC for years 2010 - 2017. While also in the news domain, XSum is by design significantly more abstractive than CNN/DM and is made of single-sentence summaries.
- **Reddit TIFU** (Kim et al., 2019) contains 120k posts from the popular online Reddit forum. As in other summarization works (Zhang et al., 2020), we use the TIFU-long subset, containing 37k posts. As there is no official split, we build a random 80:10:10 split for training:validation:test.

We refer to Table 3 for statistics on each dataset.

4.2 Training & Inference Details

To help the model better discriminate between candidates, we found that sampling was useful. Specifically, during training, we rank candidates by decreasing sum of normalized scores for the evaluation metrics and keep the top m_{top} and bottom m_{bottom} candidates. Thus, training time varies in $\mathcal{O}(m_{\text{top}} + m_{\text{bottom}})$, while inference is in $\mathcal{O}(m)$ as we need to score each candidate. In practice, we found that taking $m_{\text{top}} = 1$ and $m_{\text{bottom}} = 1$ performed well, on top of decreasing the training time. This means that at training time, the model only sees two candidates per data point. We scale the pool of candidates that these two are sampled from to *four* decoding methods, totalling 60 summary candidates per source document.

We train SummaReranker for five epochs. We use the Adafactor optimizer (Shazeer and Stern, 2018), with maximum learning rate 1e-5, warming up the learning rate linearly over the first 5% training steps. Training on CNN/DM takes four days on a single RTX 2080 Ti GPU.

For inference, we need to output a single candidate. After getting predicted probabilities across each metric $\mu \in \mathbb{M}$, we output the candidate maximizing the sum of predicted probabilities. Note that relaxing inference to allow for a different best candidate for each metric would improve performance, but is not practical. We perform inference with the model checkpoint maximizing the sum of the scores for the metrics on the validation set.

4.3 Base Setup Results

First, we investigate how our model performs in the base setup described in §3. We apply SummaReranker on top of PEGASUS and BART models fine-tuned on each half. For each model, we decode using beam search (1) and diverse beam search (2). The latter performs better for PEGASUS, while the former is better for BART. We then apply SummaReranker optimized jointly for R-1, R-2, and R-L on

		Decoding methods				Evaluation metrics					
Model	**Model stage**	$\mathbb{D}_{\text{train}}$	$\mathbb{D}_{\text{test}}$	m	**Optimized Metrics** ($\mathbb{M}$)	**R-1**	**R-2**	**R-L**	**BS**	**BaS**	**Gain (%)**
PEGASUS (Zhang et al., 2020)	1	{1}	{1}	8	_	44.16	21.56	41.30	_	_	_
PEGASUS - *our setup*	1	{1}	{1}	15	_	44.23	21.48	41.21	87.39	-2.78	_
PEGASUS - *our setup*	1	{2}	{2}	15	_	44.56	20.90	41.58	87.36	-2.81	_
BART (Lewis et al., 2020)	1	{1}	{1}	5	_	44.16	21.28	40.90	_	_	_
BART - *our setup*	1	{1}	{1}	15	_	43.28	20.44	40.06	87.78	-2.48	_
BART - *our setup*	1	{2}	{2}	15	_	44.48	21.21	41.60	**88.11**	**-2.33**	_
BART + R3F (Aghajanyan et al., 2020)	1	{1}	{1}	5	_	44.38	21.53	41.17	_	_	_
GSum (Dou et al., 2021)	1	{1}	{1}	4	_	**45.94**	**22.32**	**42.48**	_	_	_
GSum + RefSum (Liu et al., 2021)	2	{1}	{1}	4	_	46.18	22.36	42.91	_	_	_
BART + SimCLS (Liu and Liu, 2021)	2	{2}	{2}	16	_	46.67	22.15	43.54	66.14	_	_
PEGASUS + **SR**	2	{1}	{1}	15	{R-1, R-2, R-L}	45.56†	22.23†	42.46†	87.60†	-2.74†	3.18
PEGASUS + **SR**	2	{2}	{2}	15	{R-1, R-2, R-L}	**46.86**†	22.01†	**43.59**†	87.66†	-2.73†	5.10
PEGASUS + **SR**	2	{1, 2}	{1}	15	{R-1, R-2, R-L}	46.13†	**22.61**†	42.94†	87.67†	-2.72†	4.59
PEGASUS + **SR**	2	{1, 2}	{2}	15	{R-1, R-2, R-L}	46.83†	21.88†	43.55†	87.63†	-2.74†	4.84
BART + **SR**	2	{1}	{1}	15	{R-1, R-2, R-L}	44.60†	21.38†	41.36†	88.03†	-2.40†	3.63
BART + **SR**	2	{2}	{2}	15	{R-1, R-2, R-L}	46.47†	22.17†	43.45†	88.43†	-2.19†	4.48
BART + **SR**	2	{1, 2}	{1}	15	{R-1, R-2, R-L}	45.08†	21.79†	41.85†	88.13†	-2.37†	5.08
BART + **SR**	2	{1, 2}	{2}	15	{R-1, R-2, R-L}	46.50†	22.15†	43.50†	**88.45**†	**-2.18**†	4.51
PEGASUS + **SR** (**new SOTA**)	2	{1, 2}	{1, 2}	30	{R-1, R-2, R-L}	**47.16**†	**22.55**†	**43.87**†	87.74†	-2.71†	**5.44**
PEGASUS + **SR**	2	{1, 2}	{1, 2}	30	{BS, BaS}	45.00†	20.90	41.93†	87.56†	-2.55†	4.23
PEGASUS + **SR**	2	{1, 2}	{1, 2}	30	{R-1, R-2, R-L, BS, BaS}	46.59†	22.41†	43.45†	87.77†	-2.58†	4.39
BART + **SR**	2	{1, 2}	{1, 2}	30	{R-1, R-2, R-L}	46.62†	22.39†	43.59†	**88.47**†	-2.18†	5.05
BART + **SR**	2	{1, 2}	{1, 2}	30	{BS, BaS}	44.90†	20.85	42.03†	88.28†	**-2.05**†	6.11
BART + **SR**	2	{1, 2}	{1, 2}	30	{R-1, R-2, R-L, BS, BaS}	45.96†	21.79†	43.01†	88.44†	-2.09†	4.03
PEGASUS + **SR**	2	{1, 2, 3, 4}	{1, 2, 3, 4}	60	{R-1, R-2, R-L}	**47.04**†	**22.32**†	**43.72**†	**87.69**†	**-2.74**†	_

Table 5: **Transfer setup results on CNN/DM**. **SR** refers to SummaReranker, m refers to the number of summary candidates, **BS** and **BaS** to BERTScore and BARTScore, respectively. Best scores for each type of model (single stage, second-stage) are in bold. † marks are results significantly better than the base model counterpart among metrics that SummaReranker was optimized for. Results for optimized metrics are shaded. **Gain** represents the mean relative gain over optimized metrics.

top of each of the two base models, for each decoding method, and finally when using both decoding methods. Results are shown in Table 4.

SummaReranker improves a base PEGASUS by 4.57% to 7.21% with 15 candidates, and 8.70% to 9.36% with 30 candidates. With BART, SummaReranker improves by 3.94% to 11.65% with 15 candidates, and 7.98% with 30 candidates. When using several decoding methods, we compare the reranker performance with the best baseline among decoding methods. Notably, with SummaReranker, PEGASUS and BART models trained on 50% of the training set now surpass their counterparts trained on *the whole* training set, achieving 46.19 R-1 with PEGASUS and 45.96 R-1 with BART. This is better than GSum (Dou et al., 2021), the best reported summarization model on CNN/DM.

4.4 Transfer Setup Results

Next, we look at how SummaReranker performs in the transfer setup. That means, we apply it on top of PEGASUS and BART models fine-tuned on the entire dataset, using public checkpoints. We also include R3F (Aghajanyan et al., 2020) and GSum (Dou et al., 2021) in our single-stage model comparison. In terms of second-stage approaches, we compare SummaReranker with RefSum (Liu et al., 2021) and SimCLS (Liu and Liu, 2021). Note that SummaReranker is trained as usual, on the outputs of two base models each trained on 50%.

We first optimize for ROUGE metric {R-1, R-2, R-L} with multi-task training on CNN/DM (Table 5). With two decoding methods, PEGASUS + SummaReranker sets a new state of the art on CNN/DM with 47.16 R-1, 22.55 R-2 and 43.87 R-L, corresponding to gains of 2.60/1.65/2.29 R-1/2/L or +5.44% from our diverse beam search baseline. As expected, the relative gains in transfer setup are lower than in base setup. Next, we optimize model-based metrics, and note the difficulty in improving BERTScore, compared to BARTScore. Optimizing jointly ROUGE and model-based metrics improves all metrics, but does not match the results when training only ROUGE. Interestingly, performance gains saturate when adding two extra decoding methods (top-k and top-p sampling), despite gains in the oracle scores observed in Table 1.

To assert statistical significance of performance gains, we perform a t-test between SummaReranker scores and scores from the base model with *each* of the decoding methods being used, and mark with † results where the p-value is smaller than 0.05 for *all* these decoding methods.

We also show experts utilization (obtained with

		Decoding methods			XSum						Reddit TIFU					
Model	**Model stage**	$\mathbb{D}_{\textbf{train}}$	$\mathbb{D}_{\textbf{test}}$	m	**R-1**	**R-2**	**R-L**	**BS**	**BaS**	**Gain (%)**	**R-1**	**R-2**	**R-L**	**BS**	**BaS**	**Gain (%)**
PEGASUS (Zhang et al., 2020)	1	{1}	{1}	8	47.21	24.56	39.25	_	_	_	*26.63*	*9.01*	*21.60*	_	_	_
PEGASUS - *our setup*	1	{1}	{1}	15	**47.33**	**24.75**	**39.43**	**92.01**	**-1.92**	_	26.28	9.01	21.52	87.34	**-3.46**	_
PEGASUS - *our setup*	1	{2}	{2}	15	46.78	23.77	38.70	91.94	-2.00	_	25.67	8.07	20.97	87.47	-3.48	_
BART (Lewis et al., 2020)	1	{1}	{1}	5	45.14	22.27	37.25	_	_	_	_	_	_	_	_	_
BART - *our setup*	1	{1}	{1}	15	45.24	22.28	37.21	91.58	-1.97	_	**27.42**	**9.53**	**22.10**	87.43	-3.78	_
BART - *our setup*	1	{2}	{2}	15	44.15	20.84	35.88	91.51	-2.08	_	25.43	8.27	20.79	**87.48**	-4.19	_
BART + R3F (Aghajanyan et al., 2020)	1	{1}	{1}	5	_	_	_	_	_	_	*30.31*	*10.98*	*24.74*	_	_	_
GSum + RefSum (Liu et al., 2021)	2	{1}	{1}	4	47.45	24.55	39.41	_	_	_	_	_	_	_	_	_
PEGASUS + SimCLS (Liu and Liu, 2021)	2	{2}	{2}	16	47.61	24.57	39.44	69.81	_	_	_	_	_	_	_	_
PEGASUS **+ SR (new XSum SOTA)**	2	{1, 2}	{1}	15	**48.12**†	**24.95**	**40.00**†	**92.14**†	**-1.90**†	**1.31**	**29.57**†	9.70†	**23.29**†	87.63†	**-3.34**†	9.47
PEGASUS **+ SR**	2	{1, 2}	{2}	15	47.04	23.27	38.55	91.98	-2.01	-0.65	28.71†	8.73†	22.79†	**87.84**†	-3.42†	9.57
BART **+ SR**	2	{1, 2}	{1}	15	45.79†	22.17	37.31	91.69†	-1.97	0.33	28.99†	**9.82**	22.96†	87.53	-3.78	4.22
BART **+ SR**	2	{1, 2}	{2}	15	44.39	20.35	35.66	91.51	-2.16	-0.81	28.04†	8.66	22.41†	87.73†	-3.91†	7.59
PEGASUS **+ SR (best Reddit TIFU score)**	2	{1, 2}	{1, 2}	30	**47.72**	**24.16**	**39.42**	**92.10**†	**-1.94**	-0.53	**29.83**†	**9.50**†	**23.47**†	**87.81**†	**-3.33**†	**9.34**
BART **+ SR**	2	{1, 2}	{1, 2}	30	45.32	21.46	36.64	91.64	-2.04	-1.68	28.92†	9.16	22.87†	87.70†	-3.83†	1.69

Table 6: **Transfer setup results on XSum and Reddit TIFU**. **SR** refers to SummaReranker, m refers to the number of summary candidates, **BS** and **BaS** to BERTScore and BARTScore, respectively. Best scores for each type of model (single stage, second-stage) are in bold. † marks are results significantly better than the base model counterpart among metrics that SummaReranker was optimized for. Results for optimized metrics are shaded. **Gain** represents the mean relative gain over optimized metrics. Reddit TIFU results in italic are not directly comparable due to a different data split.

softmax weights from the gates) for the model optimized on all five metrics in Fig. 2. Notably, some experts specialize in certain metrics (for instance, expert 0 on R-2 and expert 4 on R-L).

Then, we apply SummaReranker on XSum and Reddit TIFU, as shown in Table 6. We train SummaReranker using the three ROUGE metrics $\mathbb{M} = \{\text{R-1, R-2, R-L}\}$ as objective, and $\mathbb{D} = \{\text{beam search, diverse beam search}\}$ to generate the candidates. On XSum, SummaReranker improves a base PEGASUS with beam search candidates by 1.31%, setting a new state-of-the-art of 48.12/24.95/40.00 R-1/2/L. On Reddit TIFU, we improve a base PEGASUS with beam search and diverse beam search (30 candidates) by 9.34%, reaching 29.83/9.50/23.47 R-1/2/L, and a base BART with beam search by 4.22%, reaching 28.99/9.82/22.96 R-1/2/L. Across datasets, training on a combination of beam search and diverse beam search candidates is consistently effective.

4.5 Ranking Evaluation

Beyond summary properties, we investigate the performance of re-ranking itself with rank-based evaluation measures. A perfect re-ranker should always single out the best summary from the rest, yielding oracle results. To evaluate how SummaReranker ranks the best summary, we compute the best summary candidate recall at different thresholds. Since several candidates might get the same metric scores (Appendix C), the best candidate recall at threshold k for the random uniform ranking baseline is not the standard $R@k = \frac{k}{m}$ anymore

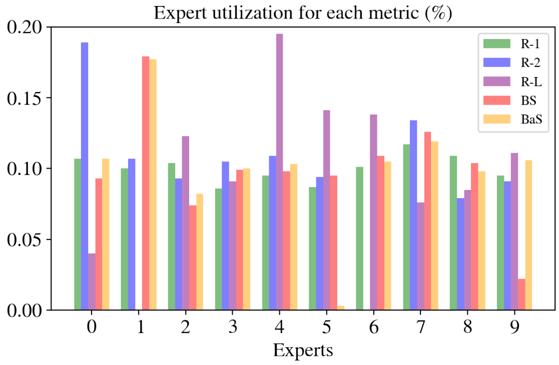


Figure 2: **Expert utilization** for a base PEGASUS with SummaReranker optimized with {R-1, R-2, R-L, BS, BaS} on **CNN/DM**, with 10 experts.

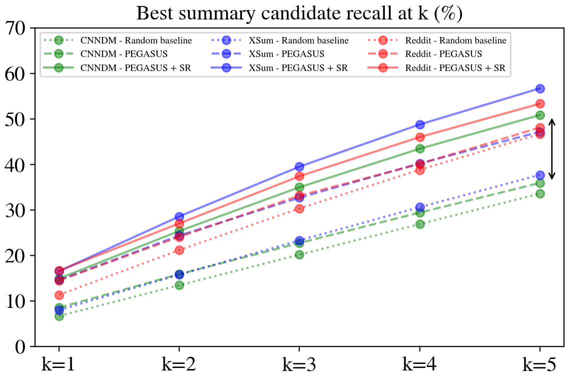
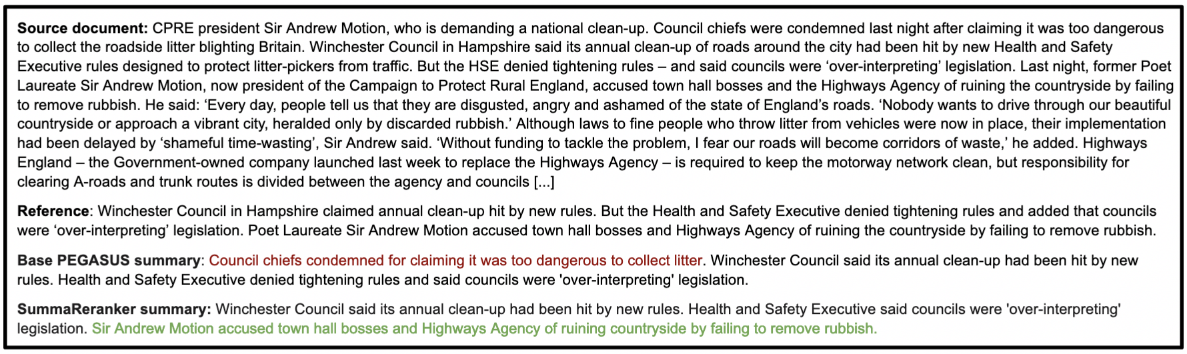


Figure 3: **Best summary candidate recall** with 15 diverse beam search candidates for PEGASUS on all three datasets. SR denotes SummaReranker. Dotted lines are random baselines, and dashed lines correspond to the base PEGASUS.



Source document: CPRE president Sir Andrew Motion, who is demanding a national clean-up. Council chiefs were condemned last night after claiming it was too dangerous to collect the roadside litter blighting Britain. Winchester Council in Hampshire said its annual clean-up of roads around the city had been hit by new Health and Safety Executive rules designed to protect litter-pickers from traffic. But the HSE denied tightening rules – and said councils were 'over-interpreting' legislation. Last night, former Poet Laureate Sir Andrew Motion, now president of the Campaign to Protect Rural England, accused town hall bosses and the Highways Agency of ruining the countryside by failing to remove rubbish. He said: 'Every day, people tell us that they are disgusted, angry and ashamed of the state of England's roads. 'Nobody wants to drive through our beautiful countryside or approach a vibrant city, heralded only by discarded rubbish.' Although laws to fine people who throw litter from vehicles were now in place, their implementation had been delayed by 'shameful time-wasting', Sir Andrew said. 'Without funding to tackle the problem, I fear our roads will become corridors of waste,' he added. Highways England – the Government-owned company launched last week to replace the Highways Agency – is required to keep the motorway network clean, but responsibility for clearing A-roads and trunk routes is divided between the agency and councils [...]

Reference: Winchester Council in Hampshire claimed annual clean-up hit by new rules. But the Health and Safety Executive denied tightening rules and added that councils were 'over-interpreting' legislation. Poet Laureate Sir Andrew Motion accused town hall bosses and Highways Agency of ruining the countryside by failing to remove rubbish.

Base PEGASUS summary: Council chiefs condemned for claiming it was too dangerous to collect litter. Winchester Council said its annual clean-up had been hit by new rules. Health and Safety Executive denied tightening rules and said councils were 'over-interpreting' legislation.

SummaReranker summary: Winchester Council said its annual clean-up had been hit by new rules. Health and Safety Executive said councils were 'over-interpreting' legislation. Sir Andrew Motion accused town hall bosses and Highways Agency of ruining countryside by failing to remove rubbish.

Figure 4: **Example of a summary** generated by SummaReranker trained for {R-1, R-2, R-L} on **CNN/DM**. The sentence in green is included in the SummaReranker summary, while the one in red is discarded.

but becomes instead:

$$R@k = \frac{\binom{m}{m_{\text{best}}} - \binom{m-k}{m_{\text{best}}}}{\binom{m}{m_{\text{best}}}} \tag{6}$$

where m_{best} is the number of best candidates.

Following Fig. 3, a PEGASUS with diverse beam search ranking of summary candidates (dashed lines) is not significantly better than the corresponding random baseline from eq. (6) (dotted lines) on CNN/DM and Reddit TIFU. However, it improves on it on XSum, confirming the observation made in Table 6 that it is harder to train a re-ranker on this dataset. On all three datasets, SummaReranker (solid lines) significantly pushes the recall at all thresholds. We note +14.90 absolute recall@5 improvement on CNN/DM (50.84 versus 35.94, indicated by the black arrow), +9.54 on XSum and +5.23 on Reddit TIFU.

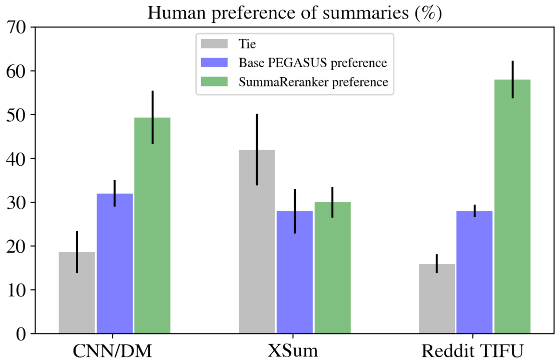


Figure 5: **Human evaluation** results on all three datasets. Black vertical bars are standard deviation across human raters.

4.6 Qualitative Evaluation

Lastly, we demonstrate that re-ranking improvements in quantitative metrics also translate to qualitatively better summaries. Fig. 4 shows an example of summary selected by SummaReranker, alongside its source document, ground-truth (reference) summary and output from the base model. SummaReranker is able to include a whole sentence which was missed by the base summary. We refer to Appendix K for full re-ranking demonstrations on each of the three datasets.

We also conduct a human evaluation. We asked three different humans to evaluate 50 randomly sampled test summaries for each dataset. Human raters were graduate students with professional English proficiency (TOEFL scores above 100 out of 120). Humans were shown the source document alongside the top beam search summary from PEGASUS, and the corresponding summary candidate selected by SummaReranker. They were asked to choose which one they believe is more faithful. They could choose a tie, because in some cases the base summary and the re-ranked one are very similar, or even identical (Appendix I). In Fig. 5, we see that on average, humans are more likely to pick the SummaReranker candidate.

5 Discussion

Abstractiveness Given that we are not modifying the base model nor its training procedure, we analyze whether our re-ranking system favors more abstractive candidates. In Fig. 6, we display the percentage of novel n-grams for n in {1,2,3,4}, for a base PEGASUS with beam search (blue) and diverse beam search (purple) decoding, and when adding SummaReranker in both cases (green and red, respectively). As first raised in (See et al., 2017), summary candidates are much less abstractive than ground truth summaries on CNN/DM. Yet, our re-ranker selects more abstractive candidates

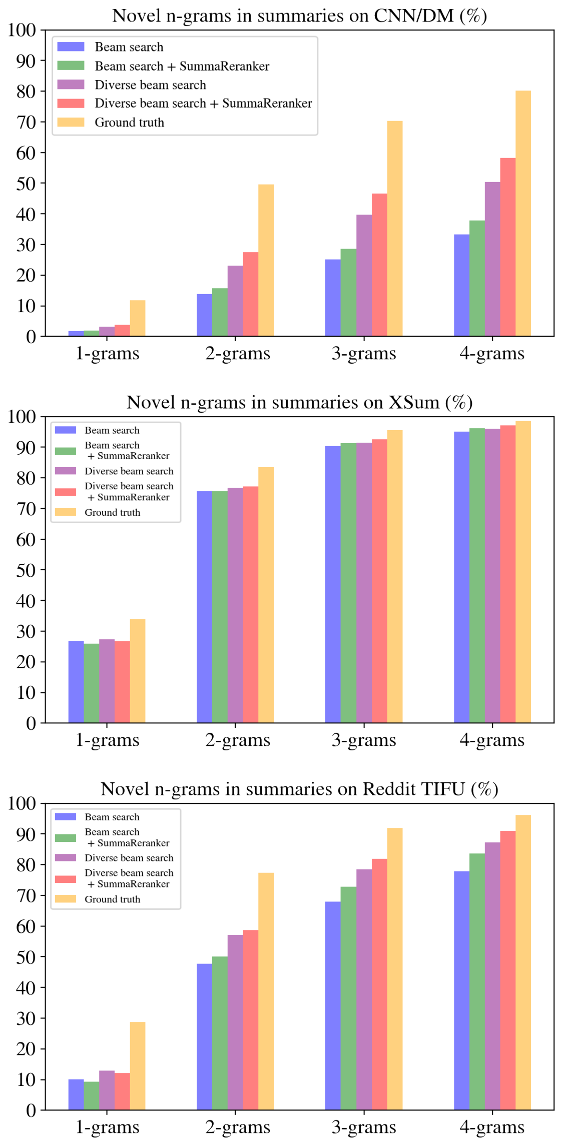


Figure 6: **Novel n-grams** with PEGASUS, across all datasets and with beam search and diverse beam search.

according to all n-grams metrics, even more so with diverse beam search, which is already more abstractive than beam search. This observation also holds on Reddit TIFU and XSum (other than 1-grams). XSum summary candidates are already almost as abstractive as the ground truth and it is harder to obtain significant abstractiveness gains through our re-ranking.

Speed/Performance trade-off On top of base model training and candidate generation, SummaReranker inference cost is linear in the number of candidates. A single candidate takes on average 38ms to be scored. As seen in Table 5 and Table 6, the performance gains from mixing several decoding methods to generate summary candidates are not scaling consistently (all four decoding methods are not better than just beam search and diverse beam search). To provide more insights on the speed/performance trade-off, we show in Appendix J SummaReranker performance when randomly sub-sampling $k \in \{1, \ldots, 15\}$ candidates. On CNN/DM, re-ranking as few as two candidates is sufficient to improve on the baseline PEGASUS. On XSum, it needs three to eight, and on Reddit TIFU three to four. As a rule of thumb, it is better to score all candidates when possible, but six to eight candidates provide a good trade-off between speed and performance across datasets.

Further Work To encode the source jointly with the summary candidate, we need to truncate the source to a fixed number of tokens. Thus, we are limited by the maximum context window of the language model encoder (512 in the case of RoBERTa-large). Applying SummaReranker to long-document summarization, such as scientific articles summarization (Cohan et al., 2018) would need better long-range modeling. In §3, we weighted metric-dependent losses uniformly. We leave to further work the exploration of more complex weight balancing or multi-task learning objectives (Lin et al., 2019).

6 Conclusion

We introduced SummaReranker, the first multi-task re-ranking framework for abstractive summarization. Encoding the source with the candidate, our model predicts whether the summary candidate maximizes each of the metrics optimized for. SummaReranker works well across diverse datasets, models, decoding methods and summarization evaluation metrics. Summaries selected by SummaReranker improve the ROUGE state-of-the-art on CNN/DM and XSum. In addition, we also show that they are more abstractive and more likely to be preferred by human evaluators over base model outputs.

Acknowledgements

This research was supported by the SINGA scholarship and partially supported by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme. We would like to thank anonymous reviewers for their insightful feedback on how to improve the paper.

References

Armen Aghajanyan, Akshat Shrivastava, Anchit Gupta, Naman Goyal, Luke Zettlemoyer, and Sonal Gupta. 2020. Better fine-tuning by reducing representational collapse. *arXiv preprint arXiv:2008.03156*.

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.

Sumanta Bhattacharyya, Amirmohammad Rooshenas, Subhajit Naskar, Simeng Sun, Mohit Iyyer, and Andrew McCallum. 2021. Energy-based reranking: Improving neural machine translation using energy-based models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4528–4537, Online. Association for Computational Linguistics.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 173–180.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Michael Collins and Terry Koo. 2005. Discriminative reranking for natural language parsing. *Computational Linguistics*, 31(1):25–70.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

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

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323. JMLR Workshop and Conference Proceedings.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28:1693–1701.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Srinivasan Iyer, Sewon Min, Yashar Mehdad, and Wen-tau Yih. 2021. RECONSIDER: Improved re-ranking using span-focused cross-attention for open domain question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1280–1287, Online. Association for Computational Linguistics.

Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. 2019. Abstractive summarization of Reddit posts with multi-level memory networks. 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 2519–2531, Minneapolis, Minnesota. Association for Computational Linguistics.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. Evaluating the efficacy of summarization evaluation across languages. *arXiv preprint arXiv:2106.01478*.

Bernhard Kratzwald, Anna Eigenmann, and Stefan Feuerriegel. 2019. RankQA: Neural question answering with answer re-ranking. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6076–6085, Florence, Italy. Association for Computational Linguistics.

Bernhard Kratzwald and Stefan Feuerriegel. 2018. Adaptive document retrieval for deep question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 576–581, Brussels, Belgium. Association for Computational Linguistics.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. 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 540–551, Hong Kong, China. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pretraining 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.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 150–157.

Xi Lin, Hui-Ling Zhen, Zhenhua Li, Qing-Fu Zhang, and Sam Kwong. 2019. Pareto multi-task learning. *Advances in neural information processing systems*, 32:12060–12070.

Xiang Lin, Simeng Han, and Shafiq Joty. 2021. Straight to the gradient: Learning to use novel tokens for neural text generation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6642–6653. PMLR.

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. *arXiv preprint arXiv: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.

Ramesh Nallapati. 2004. Discriminative models for information retrieval. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 64–71.

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

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*.

Vinay Pandramish and Dipti Misra Sharma. 2020. Checkpoint reranking: An approach to select better hypothesis for neural machine translation systems. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 286–291, Online. Association for Computational Linguistics.

Weizhen Qi, Yeyun Gong, Yu Yan, Can Xu, Bolun Yao, Bartuer Zhou, Biao Cheng, Daxin Jiang, Jiusheng Chen, Ruofei Zhang, Houqiang Li, and Nan Duan. 2021. ProphetNet-X: Large-scale pre-training models for English, Chinese, multi-lingual, dialog, and code generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 232–239, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

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, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR.

Shichao Sun and Wenjie Li. 2021. Alleviating exposure bias via contrastive learning for abstractive text summarization. *arXiv preprint arXiv:2108.11846*.

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

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

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.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *arXiv preprint arXiv:2106.11520*.

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. 2019a. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019b. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019a. 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.

Zhe Zhao, Lichan Hong, Li Wei, Jilin Chen, Aniruddh Nath, Shawn Andrews, Aditee Kumthekar, Maheswaran Sathiamoorthy, Xinyang Yi, and Ed Chi. 2019b. Recommending what video to watch next: a multitask ranking system. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pages 43–51.

A Hyper Parameters & Packages

For evaluation metrics, we used the following packages:

- For ROUGE metrics (Lin and Hovy, 2003), we used the public *rouge-score* package from Google Research: `https://github.com/google-research/google-research/tree/master/rouge`
- For BERTScore (Zhang et al., 2019a), we used the public *bert-score* package shared by the authors: `https://github.com/Tiiiger/bert_score`
- For BARTScore (Yuan et al., 2021), we used the public code shared by the authors: `https://github.com/neulab/BARTScore`

Dataset	Model	LR	Epochs	Opt.	BS	LS	MP	Source tokens	Summary tokens
CNN/DM	PEGASUS	5e-5	10	Adafactor	256	0.1	No	1024	128
	BART	3e-5	10	Adam	80	0.1	Yes	1024	128
XSum	PEGASUS	5e-5	10	Adafactor	256	0.1	No	512	64
	BART	3e-5	10	Adam	80	0.1	Yes	512	64
Reddit TIFU	PEGASUS	1e-4	15	Adafactor	256	0.1	No	512	128
	BART	3e-5	15	Adam	80	0.1	Yes	512	128

Table 7: **Hyper-parameters** for **fine-tuning** the base models. **LR** designates the *learning rate*, **Epochs** is the number of epochs, **Opt.** is the *optimizer*, **BS** is the *batch size*, **LS** means *label smoothing*, and **MP** means *mixed precision*. **Source tokens** is the maximum size of the input document, **Summary tokens** the maximum size of the output summary.

Dataset	Model	Source tokens	Summary tokens	Length penalty	Repetition penalty	Trigram blocking
CNN/DM	PEGASUS	1024	128	0.8	1.0	No
	BART	1024	128	0.8	1.0	No
XSum	PEGASUS	512	64	0.8	1.0	Yes
	BART	512	64	0.8	1.0	Yes
Reddit TIFU	PEGASUS	512	128	0.6	1.0	Yes
	BART	512	128	1.0	1.0	Yes

Table 8: **Hyper-parameters** for the summary candidates **generation** with the base models.

B Oracle Scores

Decoding methods	# Summary candidates	R-1	R-2	R-L	BS	BaS
Beam search (top beam)	1	47.33	24.75	39.43	92.01	-1.92
Beam search	15	56.07	33.80	48.33	93.19	-1.82
Diverse beam search	15	**57.82**	**35.28**	**50.95**	**93.65**	**-1.63**
Top-k sampling	15	55.57	32.54	48.35	93.18	-1.86
Top-p sampling	15	56.74	33.94	49.60	93.40	-1.77
All four above	60	**62.30**	**40.84**	**55.92**	**94.24**	**-1.48**

Table 9: **Oracle scores** for four popular decoding methods and five summarization evaluation measures for a base PEGASUS model on **XSum**.

Decoding methods	# Summary candidates	R-1	R-2	R-L	BS	BaS
Beam search (top beam)	1	26.28	9.01	21.52	87.34	-3.46
Beam search	15	36.08	14.93	29.70	88.64	-2.89
Diverse beam search	15	36.70	15.22	**30.88**	**89.08**	**-2.81**
Top-k sampling	15	36.76	14.37	29.49	88.53	-3.14
Top-p sampling	15	**37.54**	**15.24**	30.50	88.69	-3.03
All four above	60	**43.25**	**20.70**	**36.41**	**89.71**	**-2.58**

Table 10: **Oracle scores** for four popular decoding methods and five summarization evaluation measures for a base PEGASUS model on **Reddit TIFU**.

Observations from Table 9 and Table 10 are consistent with the ones made in Table 1: oracle scores are widely above the top beam baseline, and keep increasing when mixing several decoding methods.

C Unique Candidates Scores

Dataset	Model	Generation method	Scoring metric				
			R-1	R-2	R-L	BS	BaS
CNN/DM	PEGASUS	{1}	11.51	10.87	11.54	14.96	14.96
		{2}	14.34	14.09	14.34	14.99	14.99
		{3}	14.65	14.40	14.65	14.99	14.99
		{4}	14.68	14.41	14.69	15.00	15.00
	BART	{1}	11.51	10.90	11.54	14.93	14.95
		{2}	13.89	13.71	13.89	14.80	14.79
XSum	PEGASUS	{1}	8.90	7.91	8.56	14.99	14.99
		{2}	12.05	10.92	12.11	14.97	14.98
	BART	{1}	8.70	7.57	8.33	14.99	15.00
		{2}	7.37	6.63	7.37	14.59	14.99
Reddit TIFU	PEGASUS	{1}	9.19	6.31	8.85	14.99	14.99
		{2}	7.84	5.06	7.77	14.89	14.97
	BART	{1}	7.73	5.15	7.56	14.99	14.99
		{2}	7.42	3.92	7.38	14.89	14.97

Table 11: **Number of unique scores** among pools of 15 candidates generated on different datasets (CNN/DM, XSum, Reddit TIFU) with different base models (PEGASUS, BART) and different decoding methods ({1} stands for beam search, {2} is diverse beam search, {3} is top-p sampling and {4} top-k sampling). The lowest possible score of 1 indicates that all 15 candidates are assigned the same score under the metric being considered, while the highest of 15 means that all candidates are assigned a different score.

In Table 11, BERTScore (BS) and BARTScore (BaS) have results closer to 15, indicating that it is unlikely that two summary candidates share the exact metric score. This is understandable given that both these metrics are based on embeddings from pre-trained language models (BERT and BART, respectively), and embeddings values will vary whenever the input text is different, making it unlikely to have two candidates collude on the same score. In contrast, ROUGE measures n-gram overlaps, and two different summary candidates might get the same ROUGE score with the target summary (for instance if they only differ by n-grams not present in the target).

D Identical Candidates Scores

Dataset	Model	Generation method	Scoring metric				
			R-1	R-2	R-L	BS	BaS
CNN/DM	PEGASUS	{1}	0.00	0.50	0.00	0.00	0.00
		{2}	0.00	0.03	0.00	0.00	0.00
		{3}	0.03	0.06	0.03	0.03	0.00
		{4}	0.00	0.03	0.00	0.00	0.00
	BART	{1}	0.00	0.47	0.00	0.00	0.00
		{2}	0.00	0.05	0.00	0.00	0.00
XSum	PEGASUS	{1}	0.06	3.34	0.10	0.00	0.00
		{2}	0.04	1.11	0.04	0.00	0.00
	BART	{1}	0.17	4.31	0.19	0.00	0.00
		{2}	0.04	2.59	0.04	0.00	0.00
Reddit TIFU	PEGASUS	{1}	2.04	21.15	2.04	0.00	0.00
		{2}	1.52	17.03	1.52	0.00	0.00
	BART	{1}	2.32	24.14	2.32	0.00	0.00
		{2}	1.73	21.60	1.73	0.00	0.00

Table 12: **Fraction of sets of candidates with all identical scores (%)** for pools of 15 candidates generated on different datasets (CNN/DM, XSum, Reddit TIFU) with different base models (PEGASUS, BART) and different decoding methods ({1} stands for beam search, {2} is diverse beam search, {3} is top-p sampling and {4} top-k sampling.

We note that cases where all scores are identical are a small minority. ROUGE-2 is more likely than other metrics to lead to such a scenario of all identical scores.

E Metrics Correlation

	R-1	**R-2**	**R-L**	**BS**	**BaS**
R-1	1.000	0.888	0.905	0.850	0.657
R-2	0.888	1.000	0.911	0.790	0.628
R-L	0.905	0.911	1.000	0.847	0.620
BS	0.850	0.790	0.847	1.000	0.690
BaS	0.657	0.628	0.620	0.690	1.000

Table 13: **Pearson correlation coefficient** between the five evaluation metrics {R-1, R-2, R-L, BS, BaS} for a base PEGASUS decoded with beam search on **XSum**.

	R-1	**R-2**	**R-L**	**BS**	**BaS**
R-1	1.000	0.806	0.927	0.766	0.600
R-2	0.806	1.000	0.856	0.679	0.524
R-L	0.927	0.856	1.000	0.768	0.564
BS	0.766	0.679	0.768	1.000	0.646
BaS	0.600	0.524	0.564	0.656	1.000

Table 14: **Pearson correlation coefficient** between the five evaluation metrics {R-1, R-2, R-L, BS, BaS} for a base PEGASUS decoded with beam search on **Reddit TIFU**.

Metrics correlation from Table 13 and Table 14 follow the same pattern as in Table 2.

F Base Setup Results

Model	**Model stage**	**Decoding methods ($\mathbb{D}$)**	**R-1**	**R-2**	**R-L**	**Gain (%)**
PEGASUS - 1st half	1	{1}	46.02	23.38	38.10	_
PEGASUS - 1st half	1	{2}	45.41	22.37	37.22	_
PEGASUS - 2nd half	1	{1}	**46.26**	**23.45**	**38.22**	_
PEGASUS - 2nd half	1	{2}	45.53	22.42	37.31	_
BART - 1st half	1	{1}	42.76	20.22	35.00	_
BART - 1st half	1	{2}	40.93	18.75	33.44	_
BART - 2nd half	1	{1}	42.63	20.22	35.08	_
BART - 2nd half	1	{2}	40.65	18.65	33.38	_
PEGASUS - 1st half + SR	2	{1}	45.01	22.06	36.92	-3.63
PEGASUS - 1st half + SR	2	{2}	**46.35**	**22.64**	**38.05**	0.83
PEGASUS - 2nd half + SR	2	{1}	45.25	22.10	36.96	-3.77
PEGASUS - 2nd half + SR	2	{2}	46.25	22.50	37.93	1.17
BART - 1st half + SR	2	{1}	44.09	20.71	35.86	2.67
BART - 1st half + SR	2	{2}	43.70	19.90	35.21	6.07
BART - 2nd half + SR	2	{1}	44.30	20.88	36.23	3.50
BART - 2nd half + SR	2	{2}	43.96	20.03	35.49	7.27
PEGASUS - 1st half + SR	2	{1, 2}	46.74	23.10	38.35	0.37
PEGASUS - 2nd half + SR	2	{1, 2}	**47.00**	**23.30**	**38.54**	0.60
BART - 1st half + SR	2	{1, 2}	44.52	20.59	35.93	2.87
BART - 2nd half + SR	2	{1, 2}	44.68	20.76	36.20	3.57

Table 15: **Base setup results** for SummaReranker applied to PEGASUS and BART on the **XSum** dataset. **SR** refers to SummaReranker. **Decoding method {1} is beam search, {2} is diverse beam search**. Best scores for each type of model are in bold. **Gain** represents the mean relative gain over {R-1, R-2, R-L} compared to the best decoding method.

Model	**Model stage**	**Decoding methods ($\mathbb{D}$)**	**R-1**	**R-2**	**R-L**	**Gain (%)**
PEGASUS - 1st half	1	{1}	24.83	8.29	20.38	_
PEGASUS - 1st half	1	{2}	23.77	7.38	19.37	_
PEGASUS - 2nd half	1	{1}	25.16	8.42	20.53	_
PEGASUS - 2nd half	1	{2}	24.18	7.53	19.68	_
BART - 1st half	1	{1}	28.38	**9.60**	22.44	_
BART - 1st half	1	{2}	**28.60**	8.96	**22.49**	_
BART - 2nd half	1	{1}	26.94	9.13	21.65	_
BART - 2nd half	1	{2}	25.83	8.38	20.97	_
PEGASUS - 1st half + SR	2	{1}	28.78	9.20	22.74	12.83
PEGASUS - 1st half + SR	2	{2}	28.63	8.71	22.71	18.53
PEGASUS - 2nd half + SR	2	{1}	28.87	9.24	22.73	11.70
PEGASUS - 2nd half + SR	2	{2}	28.41	8.46	22.44	14.63
BART - 1st half + SR	2	{1}	28.98	**9.62**	**22.96**	1.53
BART - 1st half + SR	2	{2}	**28.89**	8.70	22.40	-0.77
BART - 2nd half + SR	2	{1}	27.93	9.48	22.45	3.73
BART - 2nd half + SR	2	{2}	28.24	8.77	22.43	6.98
PEGASUS - 1st half + SR	2	{1, 2}	**29.93**	**9.40**	**23.50**	16.37
PEGASUS - 2nd half + SR	2	{1, 2}	29.65	9.24	23.22	13.53
BART - 1st half + SR	2	{1, 2}	29.00	8.78	22.32	-2.63
BART - 2nd half + SR	2	{1, 2}	29.15	9.11	22.96	4.70

Table 16: **Base setup results** for SummaReranker applied to PEGASUS and BART on the **Reddit TIFU** dataset.

Tables Table 15 and Table 16 complement the base setup results exposed in Table 4.

G Recall Curves

Threshold k	**k=1**	**k=2**	**k=3**	**k=4**	**k=5**
CNN-DailyMail - Random baseline	6.75	13.49	20.20	26.91	33.60
CNN-DailyMail - PEGASUS	8.57	15.93	22.76	29.43	35.94
CNN-DailyMail - PEGASUS + SR	14.97	25.40	35.00	43.46	50.84
XSum - Random baseline	8.05	15.81	23.33	30.62	37.72
XSum - PEGASUS	14.60	24.40	32.70	40.23	47.17
XSum - PEGASUS + SR	16.57	28.60	39.53	48.78	56.71
Reddit TIFU - Random baseline	11.39	21.22	30.35	38.83	46.70
Reddit TIFU - PEGASUS	14.54	24.11	33.16	40.10	48.11
Reddit TIFU - PEGASUS + SR	16.70	27.07	37.42	46.02	53.34

Table 17: Values of **recall** curves plotted in Fig. 3.

H Human Evaluation

	Tie		**Base model**		**SummaReranker**	
	Mean	Std	Mean	Std	Mean	Std
CNN/DM	18.67	9.50	32.00	6.00	49.33	12.20
XSum	42.00	16.33	28.00	10.20	30.00	7.12
Reddit TIFU	16.00	4.32	28.00	2.82	58.00	4.32

Table 18: Numbers of the **human evaluation** in Fig. 5.

I Candidate Selection

Dataset	**Model**	**Generation method**	**SR pick** the base candidate (%)	**SR pick** the best candidate (%)
CNN/DM	PEGASUS	{1}	3.57	14.81
		{2}	11.11	15.00
	BART	{1}	**2.75**	**15.51**
		{2}	6.67	13.54
XSum	PEGASUS	{1}	**4.86**	9.97
		{2}	20.73	16.57
	BART	{1}	8.01	18.19
		{2}	22.23	**23.80**
Reddit TIFU	PEGASUS	{1}	6.16	18.21
		{2}	16.82	23.09
	BART	{1}	**3.22**	24.04
		{2}	3.32	**32.88**

Table 19: **Re-ranking overlap with base and best candidates.** Fraction of time that the re-ranked summary coincides with the base model one (left), and one of the best ones (oracle scores) among generated candidates (right). **SR** is SummaReranker.

In Table 19, we observe that SummaReranker is more likely to stick to the base model candidate with diverse beam search. Results in bold represent the most ideal scenario: SummaReranker differs the most from the base setup (lowest scores of the left column), and matches the most one of the best candidates (highest scores of the right column).

J Speed/Performance Trade-off

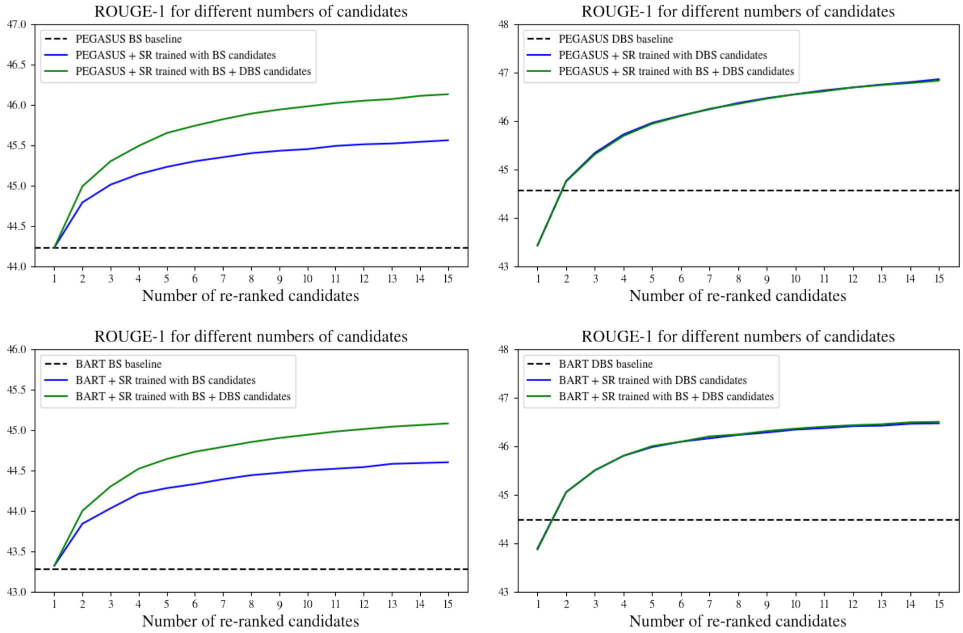


Figure 7: **ROUGE-1 on CNN/DM** for k sampled candidates at inference time, with $k \in \{1, \ldots, 15\}$. **SR** stands for SummaReranker, **BS** and **DBS** refer to beam search and diverse beam search, respectively.

In Fig. 8, we observe a failure mode of SummaReranker: on XSum and with PEGASUS when training the re-ranking with beam search candidates, performance decreases. However, the problem vanishes when SummaReranker is trained on a mixture of beam search and diverse beam search candidates.

Fig. 9 top left (PEGASUS with beam search) represents a curious case: re-ranking a *single* candidate is better than the top beam baseline. Since re-ranking a single candidate is equivalent to randomly sampling one candidate, this means that the top beam baseline is on average *lower* than sampling a random candidate. We observed that such cases are rare and usually the top beam baseline is better than the random baseline. When the top beam baseline is lower, it is of utmost importance to keep all candidate and use a second-stage method to identify a better one.

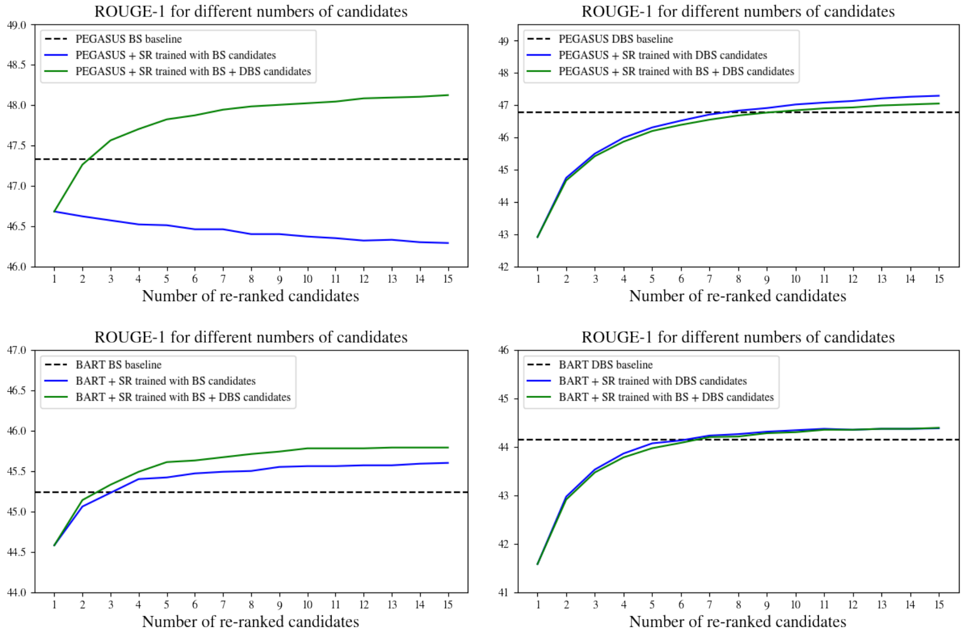


Figure 8: **ROUGE-1 on XSum** for k sampled candidates at inference time, with $k \in \{1, \ldots, 15\}$. **SR** stands for SummaReranker, **BS** and **DBS** refer to beam search and diverse beam search, respectively.

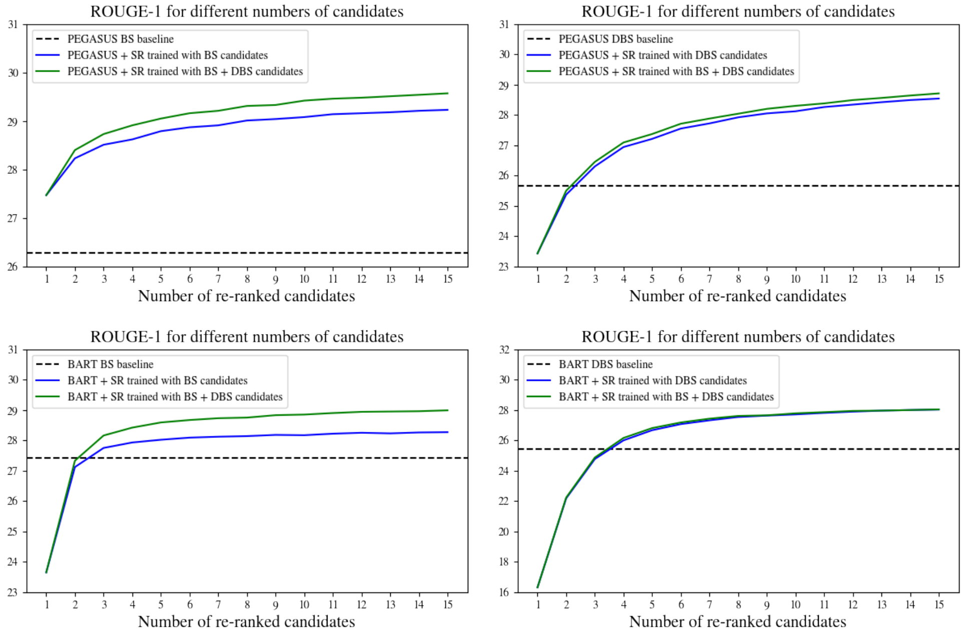


Figure 9: **ROUGE-1 on Reddit TIFU** for k sampled candidates at inference time, with $k \in \{1, \ldots, 15\}$. **SR** stands for SummaReranker, **BS** and **DBS** refer to beam search and diverse beam search, respectively.

K Re-ranking Examples

		CNN/DM
Source		Is this confirmation that Angel Di Maria is happy as a Manchester United player? The 27-year-old has endured a mixed start to his United career on-and-off the pitch since joining the club last summer - which has included an attempted burglary at his family home in Cheshire back in February. The midfielder has been linked with a move away from Old Trafford as a result, but speculation about his future could be squashed following his latest tattoo. Angel Di Maria (left) has a new No 7 tattoo which stands out among others on his left arm . Di Maria wears the No 7 shirt at Manchester United following his £60million from Real Madrid last summer. A new picture has been revealed on Twitter of Di Maria's latest piece of body art - the number seven which stands out strongly among others on his left arm. United's club record £60million signing of course adorns the No 7 shirt at the Red Devils - so could his latest tattoo suggest he's committed to Louis van Gaal's side for the long haul? However, before United fans get too carried away it must be noted that the former Real Madrid star does also wear the No 7 jersey for Argentina too. As well as adorning the No 7 shirt at United, 27-year-old (right) also wears that number for Argentina too.
Beam #1	Summary	Angel Di Maria has revealed his latest tattoo on Twitter. The 27-year-old has the No 7 shirt at Manchester United on his left arm. The Argentine has endured a mixed start to his United career. He has been linked with a move away from Old Trafford as a result.
	Reference scores	R-1: 38.6364, R-2: 18.6047, R-L: 34.0909 // Rank: 15
	Re-ranking	SummaReranker score: 0.1577 // SummaReranker score rank: 15
Beam #2	Summary	Angel Di Maria has a new tattoo of the No 7 shirt at Manchester United. The 27-year-old has endured a mixed start to his United career. The midfielder has been linked with a move away from Old Trafford. Di Maria also wears the No 7shirt for Argentina too.
	Reference scores	R-1: 59.0909, R-2: 34.8837, R-L: 56.8182 // Rank: 5
	Re-ranking	SummaReranker score: 0.3905 // SummaReranker rank: 6
Beam #3	Summary	Angel Di Maria has had a new No 7 tattoo on his left arm. The number stands out strongly among others on his arm. The 27-year-old joined Manchester United for a club record £60million last summer. Di Maria also wears the No 7 shirt for Argentina.
	Reference scores	R-1: 61.1765, R-2: 33.7349, R-L: 58.8235 // Rank: 4
	Re-ranking	SummaReranker score: 0.4447 // SummaReranker rank: 5
Beam #4	Summary	Manchester United's record signing has a new No 7 tattoo on his left arm. Angel Di Maria wears the number seven shirt at Old Trafford. The 27-year-old has endured a mixed start to his United career. He has been linked with a move away from Old Trafford as a result.
	Reference scores	R-1: 37.7778, R-2: 15.9091, R-L: 37.7778 // Rank: 14
	Re-ranking	SummaReranker score: 0.2528 // SummaReranker score: 10
Beam #5	Summary	Angel di Maria's latest tattoo shows him with the No 7 shirt at Manchester United. The 27-year-old has endured a mixed start to his United career. The midfielder has been linked with a move away from Old Trafford. Di Maria joined United for a club record £60million from Real Madrid.
	Reference scores	R-1: 53.3333, R-2: 27.2727, R-L: 48.8889 // Rank: 10
	Re-ranking	SummaReranker score: 0.2377 // SummaReranker rank: 12
Beam #6	Summary	Argentina star Angel Di Maria has a new No 7 tattoo on his left arm. The number stands out strongly among others on his arm. Di Maria joined Manchester United for a club record £60million last summe$. The 27-year-old does also wear the No 7 shirt for Argentina too.
	Reference scores	R-1: 61.3636, R-2: 37.2093, R-L: 59.0909 // Rank: 2
	Re-ranking	SummaReranker score: 0.3058 // SummaReranker rank: 8
Beam #7	Summary	Manchester United's Angel Di Maria has had a new No 7 tattoo on his left arm. The 27-year-old's latest body art was revealed on Twitter. Di Maria wears the No 7 shirt at Old Trafford following his £60million move from Real Madrid last summer.
	Reference scores	R-1: 56.4706, R-2: 31.3253, R-L: 47.0588 // Rank: 8
	Re-ranking	SummaReranker score: 0.8853 // SummaReranker rank: 2
Beam #8	Summary	The Manchester United star has revealed his latest tattoo on Twitter. Angel Di Maria has been linked with a move away from Old Trafford in recent weeks. Di Maria wears the No 7 shirt at United following his £60million move from Real Madrid last summer.
	Reference scores	R-1: 48.7805, R-2: 25.0000, R-L: 43.9024 // Rank: 12
	Re-ranking	SummaReranker score: 0.2473 // SummaReranker rank: 11
Beam #9	Summary	**Manchester United's Angel Di Maria has had a new No 7 tattoo on his left arm. The 27-year-old's latest body art was revealed on Twitter. Di Maria wears the No 7 shirt at Old Trafford following his £60million move from Real Madrid last summer. The Argentine also wears the number for Argentina too.**
	Reference scores	R-1: 61.7021, R-2: 34.7826, R-L: 53.1915 // Rank: 6
	Re-ranking	**SummaReranker score: 0.9135 // SummaReranker rank: 1**
Beam #10	Summary	The Manchester United star has revealed his latest tattoo on Twitter. Angel Di Maria has been linked with a move away from Old Trafford in recent weeks. Di Maria wears the No 7 shirt at United following his £60million move from Real Madrid last summer. The Argentine also wears the number for Argentina too.
	Reference scores	R-1: 54.9451, R-2: 29.2135, R-L: 50.5495 // Rank: 9
	Re-ranking	SummaReranker score: 0.1829 // SummaReranker rank: 14
Beam #11	Summary	Man Utd star Angel Di Maria has revealed his latest tattoo on Twitter. The 27-year-old has the No 7 shirt at Manchester United on his left arm. Di Maria joined United for a club record £60million from Real Madrid last summer. The Argentine also wears the No 7 shirt for Argentina too.
	Reference scores	R-1: 68.1319, R-2: 40.4494, R-L: 61.5385 // Rank: 1
	Re-ranking	SummaReranker score: 0.3383 // SummaReranker rank: 7
Beam #12	Summary	Manchester United star Angel Di Maria has had a new No 7 tattoo. The number stands out strongly among others on his left arm. Di Maria wears the No 7 shirt at Old Trafford following his £60million move. The Argentine also wears the number for his country too.
	Reference scores	R-1: 54.1176, R-2: 24.0964, R-L: 42.3529 // Rank: 11
	Re-ranking	SummaReranker score: 0.2172 // SummaReranker rank: 13
Beam #13	Summary	Manchester United midfielder Angel Di Maria has a new tattoo of the No 7 shirt at the club on his left arm. The 27-year-old has endured a mixed start to his United career on-and-off the pitch since joining the club last summer. Di Maria has been linked with a move away from Old Trafford as a result.
	Reference scores	R-1: 40.8163, R-2: 20.8333, R-L: 36.7347 // Rank: 13
	Re-ranking	SummaReranker score: 0.2782 // SummaReranker rank: 9
Beam #14	Summary	Angel Di Maria has revealed his latest tattoo on Twitter. The 27-year-old has the number seven inked on his left arm. Di Maria joined Manchester United for a club record £60million last summer. The Argentine also wears the No 7 shirt for Argentina.
	Reference scores	R-1: 58.5366, R-2: 35.0000, R-L: 56.0976 // Rank: 7
	Re-ranking	SummaReranker score: 0.7447 // SummaReranker rank: 3
Beam #15	Summary	Angel di Maria has a new No 7 tattoo on his left arm. The number seven is among others on his left arm. The 27-year-old wears the No 7 shirt at Manchester United. Di Maria joined United for a club record £60million from Real Madrid last summer.
	Reference scores	R-1: 62.7907, R-2: 33.3333, R-L: 58.1395 // Rank: 3
	Re-ranking	SummaReranker score: 0.4988 // SummaReranker rank: 4
Reference		Angel di Maria joined Manchester United from Real Madrid for £60million. Di Maria took the No 7 shirt upon his arrival at the English giants. 27-year-old also wears the No 7 jersey for Argentina too.

Table 20: **Diverse beam search summary candidates** of a base PEGASUS and their ground truth and SummaReranker re-ranking scores on **CNN/DM**.

XSum		
Source		Female officers will be able to wear a headscarf under their caps or berets, provided it is plain and is the same colour as the uniform. Headscarf bans on university campuses and state institutions - except for the judiciary, military and police - have also been lifted in recent years. The garment has been controversial in Turkey for years. Secularists regard it as a symbol of religious conservatism. Since the 1920s, Turkey has had a secular constitution with no state religion. The opposition have accused President Recep Tayyip Erdogan and his Islamist-rooted Justice and Development Party (AKP) of trying to reinterpret secularism. However, public debate has also evolved to accept the hijab as an expression of individual liberties, correspondents say. No strong opposition has been voiced against this latest move. President Erdogan has long embraced Turks' right to express their religious beliefs openly, but he says he is committed to secularism. In 2010, the country's universities abandoned an official ban on Muslim headscarves. Three years later, women were allowed to wear headscarves in state institutions - with the exception of the judiciary, military and police. That year, four MPs wore headscarves in parliament. Most people in Turkey are Sunni Muslims.
Beam #1	Summary	The Turkish authorities have lifted a ban on female police officers wearing headscarves.
	Reference scores	R-1: 50.0000, R-2: 27.2727, R-L: 41.6667 // Rank: 11
	Re-ranking	SummaReranker score: 0.6553 // SummaReranker rank: 12
Beam #2	Summary	Turkey has lifted a ban on female police officers wearing headscarves, the interior ministry says.
	Reference scores	R-1: 61.5385, R-2: 41.6667, R-L: 61.5385 // Rank: 2
	Re-ranking	SummaReranker score: 0.8562 // SummaReranker rank: 2
Beam #3	Summary	The Turkish authorities have lifted a ban on female police officers wearing headscarves, state media report.
	Reference scores	R-1: 53.8462, R-2: 25.0000, R-L: 53.8462 // Rank: 8
	Re-ranking	SummaReranker score: 0.5605 // SummaReranker rank: 1
Beam #4	Summary	Turkey has lifted its ban on female police officers wearing headscarves, the interior ministry says.
	Reference scores	R-1: 53.8462, R-2: 25.0000, R-L: 53.8462 // Rank: 8
	Re-ranking	SummaReranker score: 0.7049 // SummaReranker rank: 9
Beam #5	Summary	The Turkish government has lifted a ban on female police officers wearing headscarves.
	Reference scores	R-1: 58.3333, R-2: 36.3636, R-L: 50.0000 // Rank: 5
	Re-ranking	SummaReranker score: 0.7104 // SummaReranker rank: 8
Beam #6	Summary	The Turkish authorities have lifted a ban on police officers wearing headscarves.
	Reference scores	R-1: 52.1739, R-2: 28.5714, R-L: 43.4783 // Rank: 10
	Re-ranking	SummaReranker score: 0.7503 // SummaReranker rank: 7
Beam #7	Summary	**Turkey has lifted a ban on female police officers wearing headscarves.**
	Reference scores	R-1: 63.6364, R-2: 50.0000, R-L: 63.6364 // Rank: 1
	Re-ranking	**SummaReranker score: 0.9019 // SummaReranker rank: 1**
Beam #8	Summary	Turkey's police force has lifted its ban on female officers wearing headscarves.
	Reference scores	R-1: 50.0000, R-2: 18.1818, R-L: 50.0000 // Rank: 12
	Re-ranking	SummaReranker score: 0.6919 // SummaReranker rank: 10
Beam #9	Summary	Turkey's police force has lifted a ban on female officers wearing headscarves.
	Reference scores	R-1: 58.3333, R-2: 36.3636, R-L: 58.3333 // Rank: 4
	Re-ranking	SummaReranker score: 0.8103 // SummaReranker rank: 5
Beam #10	Summary	Turkey's police force has lifted its ban on female officers wearing headscarves, officials say.
	Reference scores	R-1: 46.1538, R-2: 16.6667, R-L: 46.1538 // Rank: 13
	Re-ranking	SummaReranker score: 0.5066 // SummaReranker rank: 15
Beam #11	Summary	The Turkish government has lifted a ban on female police officers wearing headscarves, state media report.
	Reference scores	R-1: 51.8519, R-2: 32.0000, R-L: 44.4444 // Rank: 9
	Re-ranking	SummaReranker score: 0.6522 // SummaReranker rank: 13
Beam #12	Summary	Turkey's police force has lifted a ban on female officers wearing headscarves, state media report.
	Reference scores	R-1: 51.8519, R-2: 32.0000, R-L: 51.8519 // Rank: 7
	Re-ranking	SummaReranker score: 0.7819 // SummaReranker rank: 6
Beam #13	Summary	Turkey has lifted its ban on female police officers wearing headscarves.
	Reference scores	R-1: 54.5455, R-2: 30.0000, R-L: 54.5455 // Rank: 6
	Re-ranking	SummaReranker score: 0.8140 // SummaReranker rank: 4
Beam #14	Summary	Turkey has lifted a ban on female police officers wearing headscarves, the interior ministry has said.
	Reference scores	R-1: 59.2593, R-2: 40.0000, R-L: 59.2593 // Rank: 3
	Re-ranking	SummaReranker score: 0.8298 // SummaReranker rank: 3
Beam #15	Summary	Turkey's police force has lifted its ban on female officers wearing headscarves, state media report.
	Reference scores	R-1: 44.4444, R-2: 16.0000, R-L: 44.4444 // Rank: 15
	Re-ranking	SummaReranker score: 0.6728 // SummaReranker rank: 11
Reference		Turkey has lifted a ban on police women wearing the Islamic headscarf.

Table 21: **Beam search summary candidates** of a base PEGASUS and their ground truth and SummaReranker re-ranking scores on **XSum**.

		Reddit TIFU
Source		here's my reconstruction of the fuck-up: during the visa application, i'm sifting through pages and pages of documentation with 15 tabs open on my browser and i arrive at a page with the title english requirement. it says something like "here's a list of approved test providers and you have to score a minimum cefr level of b1 to meet the english requirement." as someone who has taken many english exams such as toefl, ielts and pearson, i wonder what the hell a cefr level is, how come i've never heard of this and start popping new pages. turns out you have to score that much from ielts or this much from pearson or that much from other exams. cool. i'm thinking, currently i have 2 valid ielts exams that meet the criteria and a pearson's from which i've scored 90/90, sweet! i'll just submit pearson's and done. so i pay 2000aud and get an appointment, submit my documents and come back home. "hey wifey, it was really easy, let's do the same for you and get it done quickly." pay another 2000aud and my wife submits her application. 3 days after my submission, i get an e-mail saying a decision has been made, yay? more like nay, refused because we don't accept pearson's. 2 days later, wife gets refuses as well because we refused your husband. 2000 down the drain, luckily they are refunding the healthcare extras. turns out i failed to go back to that webpage and click on the link to get access to the most ridiculous list of approved english tests. there's not a mention of pearson, what's more, it doesn't have idp melbourne so my ielts exams are useless as well. on the plus side, i learn there are other ways to meet the requirement. if you have a diploma from an australian uni, just submit your diploma and you're good to go? why didn't you give me a call or send me an email and say "hey, you made a mistake silly" and i'd be like here's my diploma or something, wtf is refusing? this will stick to our international travel cv like a fucking bug. plus, i had to ask my future employer for another certificate, which takes another two weeks and makes me look like an idiot. then i'll have to submit another application, pay another 4000 and yada yada. A bit of background: we are turkish citizens living in australia as permanent residents obtained by using the same pearson's test. i'm working as a post-doctoral researcher and hopefully starting another post-doctoral position in scotland, not giving up!
Beam #1	Summary	i didn't go back to the page that says you have to score a cefr level of b1 to meet the english requirement.
	Reference scores	R-1: 4.4444, R-2: 0.0000, R-L: 4.4444 // Rank: 12
	Re-ranking	SummaReranker score: 0.8117 // SummaReranker rank: 12
Beam #2	Summary	i'm an idiot.
	Reference scores	R-1: 0.0000, R-2: 0.0000, R-L: 0.0000 // Rank: 15
	Re-ranking	SummaReranker score: 0.0612 // SummaReranker rank: 15
Beam #3	Summary	i didn't go back to the page that says you have to score a minimum cefr level of b1 to meet the english requirement.
	Reference scores	R-1: 4.3478, R-2: 0.0000, R-L: 4.3478 // Rank: 13
	Re-ranking	SummaReranker score: 0.7396 // SummaReranker rank: 14
Beam #4	Summary	i didn't go back to the page that says you have to score a cefr level of b1 to meet the english requirement
	Reference scores	R-1: 4.4444, R-2: 0.0000, R-L: 4.4444 // Rank: 11
	Re-ranking	SummaReranker score: 0.7940 // SummaReranker rank: 13
Beam #5	Summary	i didn't go back to the page where i found out i had to score a minimum cefr level of b1 to meet the english requirement.
	Reference scores	R-1: 4.1667, R-2: 0.0000, R-L: 4.1667 // Rank: 14
	Re-ranking	SummaReranker score: 0.8320 // SummaReranker rank: 11
Beam #6	Summary	i didn't go back to the page that says you have to score a cefr level of b1 to meet the english requirement and now i'm stuck in australia.
	Reference scores	R-1: 7.6923, R-2: 0.0000, R-L: 7.6923 // Rank: 7
	Re-ranking	SummaReranker score: 0.8954 // SummaReranker rank: 5
Beam #7	Summary	i didn't go back to the page that says you have to score a minimum cefr level of b1 to meet the english requirement and now i'm stuck in australia.
	Reference scores	R-1: 7.5472, R-2: 0.0000, R-L: 7.5472 // Rank: 8
	Re-ranking	SummaReranker score: 0.8890 // SummaReranker rank: 8
Beam #8	Summary	i didn't go back to the page that says you have to score a minimum cefr level of b1 to meet the english requirement and now i'm stuck in a foreign country.
	Reference scores	R-1: 7.2727, R-2: 0.0000, R-L: 7.2727 // Rank: 10
	Re-ranking	SummaReranker score: 0.8889 // SummaReranker rank: 9
Beam #9	Summary	i didn't go back to the page that says you have to score a cefr level of b1 to meet the english requirement on my visa application and now i'm stuck in australia.
	Reference scores	R-1: 7.2727, R-2: 0.0000, R-L: 7.2727 // Rank: 10
	Re-ranking	SummaReranker score: 0.8941 // SummaReranker rank: 7
Beam #10	Summary	**i didn't go back to the page that says you have to score a minimum cefr level of b1 to meet the english requirement on my visa application and now i'm stuck in australia.**
	Reference scores	R-1: 21.4286, R-2: 3.7037, R-L: 17.8571 // Rank: 1
	Re-ranking	**SummaReranker score: 0.9052 // SummaReranker rank: 1**
Beam #11	Summary	i didn't go back to the page that says you have to score a minimum cefr level of b1 to meet the english requirement on my visa application and now i'm stuck in australia.
	Reference scores	R-1: 21.0526, R-2: 3.6364, R-L: 17.5439 // Rank: 2
	Re-ranking	SummaReranker score: 0.9045 // SummaReranker rank: 2
Beam #12	Summary	i didn't go back to the page that says you have to score a minimum cefr level of b1 to meet the english requirement and now i'll have to submit another application, pay 4000 and look like an idiot.
	Reference scores	R-1: 12.9032, R-2: 0.0000, R-L: 12.9032 // Rank: 6
	Re-ranking	SummaReranker score: 0.8861 // SummaReranker ran: 10
Beam #13	Summary	i didn't go back to the page that says you have to score a cefr level of b1 to meet the english requirement and now i'm going to have to submit another application and pay 4000.
	Reference scores	R-1: 13.5593, R-2: 0.0000, R-L: 13.5593 // Rank: 3
	Re-ranking	SummaReranker score: 0.8994 // SummaReranker rank: 3
Beam #14	Summary	i didn't go back to the page that says you have to score a minimum cefr level of b1 to meet the english requirement and now i'm going to have to submit another application and pay 4000.
	Reference scores	R-1: 13.3333, R-2: 0.0000, R-L: 13.3333 // Rank: 4
	Re-ranking	SummaReranker score: 0.8947 // SummaReranker rank: 6
Beam #15	Summary	i didn't go back to the page that says you have to score a minimum cefr level of b1 to meet the english requirement and now i'm going to have to submit another application and pay 4000 dollars.
	Reference scores	R-1: 13.1148, R-2: 0.0000, R-L: 13.1148 // Rank: 5
	Re-ranking	SummaReranker score: 0.8964 // SummaReranker rank: 4
Reference		made a silly mistake and got refused on 2x tier 2 uk visa applications for me and my partner costing 2000aud.

Table 22: **Beam search summary candidates** of a base PEGASUS and their ground truth and SummaReranker re-ranking scores on **Reddit TIFU**.