Unsupervised Summarization Re-ranking

Mathieu Ravaut[1,2], **Shafiq Joty**[*1,3] **Nancy F. Chen**[2]
[1] Nanyang Technological University, Singapore
[2] Institute of Infocomm Research (I^{2}R), A*STAR, Singapore
[3] Salesforce AI
{mathieuj001@e.ntu, srjoty@ntu}.edu.sg
nfychen@i2r.a-star.edu.sg

Abstract

With the rise of task-specific pre-training objectives, abstractive summarization models like PEGASUS offer appealing zero-shot performance on downstream summarization tasks. However, the performance of such unsupervised models still lags significantly behind their supervised counterparts. Similarly to the supervised setup, we notice a very high variance in quality among summary candidates from these models while only one candidate is kept as the summary output. In this paper, we propose to re-rank summary candidates in an *unsupervised* manner, aiming to close the performance gap between unsupervised and supervised models. Our approach improves the unsupervised PEGASUS by up to 7.27% and ChatGPT by up to 6.86% relative mean ROUGE across four widely-adopted summarization benchmarks ; and achieves relative gains of 7.51% (up to 23.73% from XSum to WikiHow) averaged over 30 zero-shot transfer setups (finetuning on a dataset, evaluating on another).[1]

1 Introduction

Transformer-based encoder-decoder language models have achieved great success in abstractive summarization in the last few years, and produce fluent summaries which can be quite abstractive (Raffel et al., 2019; Lewis et al., 2020; Zhang et al., 2020). These models follow the *pre-train then fine-tune* paradigm: they are first pre-trained with a self-supervised objective on a large text corpus; then they are fine-tuned on the downstream dataset of interest, using the available supervision, which may be very scarce. Finding a better pre-training objective remains an active research area. Some models like T5 (Raffel et al., 2019) and BART (Lewis et al., 2020) adopt a more general language modeling objective (e.g., masked span generation), while others like PEGASUS (Zhang et al., 2020) or TED (Yang et al., 2020) are pre-trained specifically for the task of summarizing a document. PEGASUS uses salient sentences of the document as a proxy summary label, while TED leverages the lead bias to get the pseudo-summary target.

Generation method	**Summary candidate**	**R-1**	**R-2**	**R-L**
Beam search	First (top beam)	35.47	13.89	31.61
	Random	34.89	13.46	31.22
	Minimum	26.64	7.68	23.18
	Maximum (oracle)	**42.62**	**19.76**	**38.75**
Diverse beam search	First	34.35	13.02	30.65
	Random	31.73	11.22	28.4
	Minimum	21.25	4.45	18.61
	Maximum (oracle)	**41.87**	**19.29**	**38.22**
Nucleus sampling	First	32.14	11.29	28.66
	Random	32.12	11.29	28.64
	Minimum	24.09	6.49	21.19
	Maximum (oracle)	**40.19**	**17.47**	**36.43**

Table 1: ROUGE results with PEGASUS (unsupervised) on CNN/DM test set, for three generation methods to produce 20 summary candidates, and four candidate selection strategies. **R-1**, **R-2**, **R-L** stands for ROUGE-1/2/L.

Despite the impressive success on supervised abstractive summarization tasks, unsupervised summarization remains very challenging. The LEAD-3 (extractive) baseline which simply takes the first three sentences of a document as its summary, remains far ahead of unsupervised approaches on several news summarization datasets (See et al., 2017), especially the popular CNN/DM dataset (Hermann et al., 2015). In fact, it was only improved on by *supervised* abstractive models not more than five years ago (Narayan et al., 2018). It is expected that a model which has never seen *any* summarization example would struggle, as summarization is a task that is subjective and complex even for humans (Kryscinski et al., 2019). Since summarization labels are expensive to collect, it is essential to develop models with good zero-shot performance. Starting from instruction-tuned GPT-3, LLMs are offering promising performance in zero-shot summarization (Goyal et al., 2022), but remain an unscalable solution as these models are

[*]Work done when the author was on leave from NTU.
[1]Code for all experiments are available at https://github.com/ntunlp/SummScore.

Findings of the Association for Computational Linguistics: ACL 2023, pages 8341–8376

rarely open-source, and extremely computationally intensive.

Recently, in the supervised setup, second-stage approaches have gathered interest in abstractive summarization research. While the base encoder-decoder model is trained with maximum-likelihood estimation (MLE) to predict each token of the ground-truth summary in an autoregressive manner, second-stage methods work with a global view at the whole sequence level. SimCLS (Liu and Liu, 2021) and SummaReranker (Ravaut et al., 2022a) propose to train another neural model to rank summary candidates generated by decoding methods like beam search (Reddy, 1977) or diverse beam search (Vijayakumar et al., 2016). BRIO (Liu et al., 2022a) bypasses the need for another model, and re-uses the fine-tuned model for another fine-tuning stage in which the model also learns to rank candidates in the correct order. SummaFusion (Ravaut et al., 2022b) encodes each summary candidate separately and decodes into a new, abstractive second-stage summary. Such second-stage methods have improved ROUGE-1 state-of-the-art on CNN/DM by more than 3 points (Liu et al., 2022a).

In this paper, we propose to re-rank summary candidates in the *unsupervised* setup. Following observations made by second-stage summarization studies in the supervised setup (Liu et al., 2021; Ravaut et al., 2022a), we also observe large variance in performance among summary candidates in the unsupervised setup. In Table 1, the ***oracle*** for PEGASUS, which is the summary candidate maximizing the ROUGE score with the reference, reaches 42.62 when using beam search with 20 beams on CNN/DM (Hermann et al., 2015). This is in the same range (42-45 ROUGE-1) as the top beam of *supervised* leading models on this dataset (Lewis et al., 2020; Zhang et al., 2020). This observation implies strong potential motivating our work: ***with a perfect unsupervised summarization re-ranker, one could potentially by-pass supervised fine-tuning and just re-rank instead***.

The main challenge lies in the fact that the re-ranker must also not access any supervision. Our proposed model does not train any neural model, but simply computes features indicative of summary quality to score each summary candidate, some of them which also leverage the source document. A weighted average of these features is used for candidate re-ranking, and we explore several methods to estimate the feature weights. Our method, named SummScore, is lightweight, fast and easy to use as it does not rely on a neural network. Since it is purely unsupervised, the re-ranked results can provide more refined self-supervision to the pre-trained models, complementing the pre-training with rounds of self-training.

Our contributions in this paper are threefold:

- We propose SummScore, the first system to re-rank summarization candidates in an unsupervised setup and in an unsupervised manner.
- We demonstrate the strength of SummScore by consistent performance improvement: up to +7.27% with PEGASUS and +6.86% with ChatGPT[2] mean ROUGE gains over four unsupervised summarization datasets, +7.51% mean ROUGE gains averaged over 30 zero-shot transfer setups.
- Using the re-ranker, we derive an original and effective self-training method which continuously improves the base unsupervised summarization model, pushing PEGASUS from 35.47 to 39.76 ROUGE-1 (+12.09%).

2 Related Work

Unsupervised abstractive summarization In unsupervised abstractive summarization, SummAE (Liu et al., 2019a) proposes to auto-encode paragraphs with a sequence-to-sequence model and decode single-sentence summaries from the latent embeddings. SEQ3 (Baziotis et al., 2019) also uses an auto-encoder to compress the input then reconstruct it into a differentiable manner, the encoder output serving as a summary. However, both methods stick to unsupervised *sentence* summarization. More recent approaches typically rely on language models being pre-trained, then used in a zero-shot fashion. PEGASUS (Zhang et al., 2020) treats salient sentences as pseudo abstractive targets to build a pre-training objective. TED (Yang et al., 2020) exploits the lead bias in news articles and takes out the first sentences of the document as pseudo summary targets for pre-training. Due to their pre-training objective built for summary generation, these pre-trained models can be directly used for unsupervised summarization. The Summary Loop (Laban et al., 2020) uses reinforcement learning to train a model to fill-in deleted important words from the source document using the summary generated so far, then refines this summary.

[2] https://chat.openai.com/

Re-ranking in abstractive summarization Second-stage or sequence-level methods are gaining traction recently in *supervised* summarization. Among such methods, re-ranking consists in selecting a better summary candidate out of several of them produced by a base model (which has already been fine-tuned). RefSum (Liu et al., 2021) uses a meta-learning approach to learn how to rank summaries coming from multiple systems. SimCLS (Liu and Liu, 2021) trains a RoBERTa (Liu et al., 2019b) model with a ranking loss to learn how to rank summary candidates generated by a base BART or PEGASUS in their target metric order. SummaReranker (Ravaut et al., 2022a) also trains a RoBERTa re-ranker, but this time in a multi-label binary classification manner to predict whether each summary candidate maximizes each of the metrics of interest. To avoid using another neural network for re-ranking, BRIO (Liu et al., 2022b) performs a second fine-tuning stage with the re-ranking loss built in the base summarization system. Each of the four models above improves the SOTA on the CNN/DM benchmark, reaching 47.78 ROUGE-1 for BRIO.

To the best of our knowledge, there is no work on sequence-level unsupervised abstractive summarization. Concurrently to our work, MBRD (Suzgun et al., 2022) proposes to rank generated candidates in several generation tasks using majority voting based on BERTScore (Zhang et al., 2019).

3 Method

3.1 Unsupervised Summary Re-ranking

As an unsupervised summarization re-ranking approach, our method assumes access to a zero-shot self-supervised summarization model. We refer to it as the base model $\mathcal{M}_{\text{base}}$. Given a source document D, $\mathcal{M}_{\text{base}}$ will generate k *summary candidates* using a *generation method* to transform model predictions into a natural language summary. A widely used such generation approach is beam search, which maintains k top summary candidates throughout decoding, ranking them with decreasing mean log-probability of the sequence. In the end, practitioners keep the candidate maximizing the log-probability and discard the remaining, whereas we propose to keep *all* k candidates and re-rank them, following (Ravaut et al., 2022a).

Let $\mathbb{C} = \{C_1, \ldots, C_k\}$ be the pool of candidates. Our goal in (re-)ranking the candidates is to assign to each of them a score S, such that $S(C_i) > S(C_j)$ if C_i is a better candidate than C_j (for $1 \leq i, j \leq k$) according to some summary quality measures. We can then select the candidate maximizing the score as the best output:

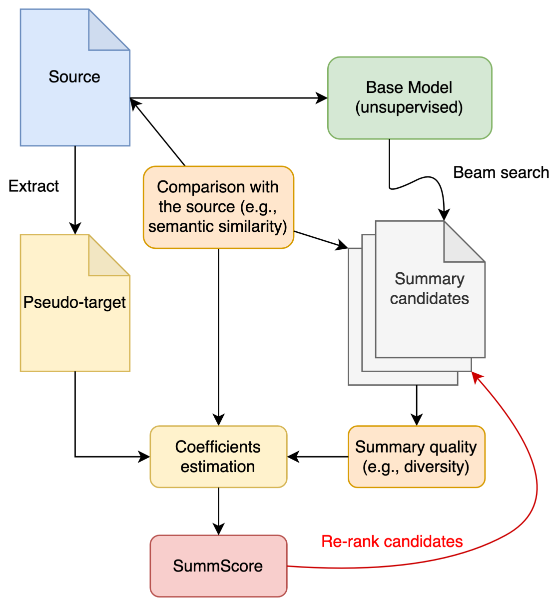

Figure 1: **SummScore (unsupervised) re-ranking** construction. SummScore leverages the source document for semantic similarity comparisons with summary candidates, as well as to extract a pseudo target.

$$C_S^* = \underset{C_i \in \mathbb{C}}{\arg\max} \; \{S(C_1), \ldots, S(C_k)\} \quad (1)$$

Unlike re-ranking in a supervised setup, where one can compute such scores by comparing with the ground truth summary or build models to optimize them (Liu and Liu, 2021; Ravaut et al., 2022a; Liu et al., 2022a), in our unspervised setup, we cannot assume access to the ground truth, which thus excludes scoring the candidate with regards to it (e.g., using ROUGE). In the following, we describe how we build our unsupervised scoring method (named *SummScore*) following principles assessing the quality of a summary.

3.2 Multi-Objective Re-ranking Score

We design our candidate-level SummScore as an aggregation of features, each representing desired properties for a summary. Features either come from the comparison between the summary candidate and the source, or from the candidate itself. Fig. 1 synthesizes the overall SummScore re-ranking process.

Comparison with the source One evident property of a summary is that it should stick to the source content, and contain as much of the important content as possible. The most straightforward way to measure this consists in using n-gram overlap metrics between the source document and each candidate. We use ROUGE-1 (noted *R-1*) (Lin, 2004), ROUGE-2 (*R-2*), and BLEU (Papineni et al., 2002), which form our first set of features:

$$S_{\text{overlap}} = \{\text{R-1}, \text{R-2}, \text{BLEU}\} \tag{2}$$

The above metrics only evaluate n-gram overlap, which can be helpful penalizing summary candidates departing too much from the source, potentially hallucinating. However, they have been shown to not be well suited at evaluating semantic similarity, and might encourage too much copying.

Thus, our next batch of SummScore features consists in model-based metrics designed to capture semantic similarity between two text items. We explore three such metrics: BERTScore (Zhang et al., 2019), BARTScore (Yuan et al., 2021) and BLEURT (Sellam et al., 2020). BERTScore (noted *BS*) computes token-level cosine similarity between the contextual embeddings of the pre-trained BERT (Devlin et al., 2019) of each text item to compare. BARTScore (noted *BaS*) uses BART (Lewis et al., 2020) token-level log-probabilities from the pre-trained BART to score the generated text. BLEURT (noted *BRT*) also leverages BERT but extends its pre-training with an additional multi-task pre-training on synthetic data. Our next features are:

$$S_{\text{semantic}} = \{\text{BS}, \text{BaS}, \text{BRT}\} \tag{3}$$

When each of these metrics is referred to, it is implicit that they are used to compare a summary candidate with the source document (in contrast to the supervised case, comparing with the target).

Summary quality A good summary should be *diverse*, meaning it should avoid repeated n-grams. We build a summary-level diversity score which measures the proportion of unique n-grams.

$$F_{\text{div}} = \frac{1}{N}\Sigma_{n=1}^{N}\frac{\text{unique } n\text{-grams}}{\text{total } n\text{-grams}} \tag{4}$$

We take $N = 3$ in practice. The summary should not be too short, nor too long. We penalize summaries which deviate a lot from the average summary length on the given dataset. To build a score with increasing values being desirable, we use a smooth inverse of the absolute length difference between the summary candidate and the mean length of summaries μ_{len}.

$$F_{\text{len}} = \frac{1}{\max(1, |\text{length} - \mu_{\text{len}}|)} \tag{5}$$

Final Score Our final set of summary features is:

$$\begin{aligned} S &= S_{\text{overlap}} \cup S_{\text{semantic}} \cup S_{\text{quality}} \\ &= \{F_1, \dots, F_{|S|}\} \end{aligned} \tag{6}$$

where $S_{\text{quality}} = \{F_{\text{div}}, F_{\text{len}}\}$. For data point x_i, SummScore simply outputs the summary candidate among the set $\mathbb{C}_i$ maximizing a weighted combination of all features above:

$$\text{SummScore}_{\theta}(\mathbb{C}_i) = \underset{C_i \in \mathbb{C}_i}{\arg\max} \sum_{j=1}^{|S|} \theta_j.F_j(C_i) \tag{7}$$

where we enforce coefficients to be $\sum_{j=1}^{|S|} \theta_j = 1.0$

3.3 Coefficients Estimation

SummScore is simply a linear combination of eight features in total. Yet a last crucial question remains: how to estimate the coefficients to assign to each feature? We propose to bootstrap a pseudo-summary using sentences from the source document. Coefficients are then tuned to maximize the mean of ROUGE-1/2/L between the summary candidate with the highest SummScore (e.g., SummScore output candidate), and the pseudo-target. We compare three approaches to extract pseudo-targets:

- **Random-3**: As a baseline, we randomly select three sentences from the source document to form a pseudo-target.
- **LEAD-3**: This consists in the first three sentences of the document. LEAD-3 is a strong baseline for lead-biased news summarization datasets (Hermann et al., 2015; See et al., 2017), and it has even been used as a pseudo-target for summarization pre-training in TED (Yang et al., 2020).
- **Salient Sentences**: We follow the *gap-sentences generation* idea introduced by PEGASUS pre-training objective (Zhang et al., 2020), and also used by SUPERT (Gao et al., 2020) for unsupervised summarization evaluation. A pseudo-target is constructed with salient sentences, which are defined as the source sentences maximizing the ROUGE with the rest of the document. The

Dataset	Domain	# Data points			# Words		# Tokens (PEGASUS)		New summary n-grams	
		Train	Val	Test	Doc.	Summ.	Doc.	Summ.	1-grams (%)	2-grams (%)
CNN/DM (Hermann et al., 2015)	News	287113	13334	11490	786.68	55.06	851.53	64.57	12.07	51.05
XSum (Narayan et al., 2018)	News	204045	11332	11334	430.18	23.19	456.96	26.01	33.98	83.33
WikiHow (Koupaee and Wang, 2018)	Wikipedia	157304	5600	5580	588.06	62.10	620.52	71.82	29.79	77.45
SAMSum (Gliwa et al., 2019)	Dialogue	14732	818	819	124.07	23.42	133.07	25.66	33.88	79.02

Table 2: Statistics on the datasets used for experiments. **Doc.** is the source document, **Summ.** the summary.

top 30% such sentences are extracted to form a pseudo-summary. We experiment with all three standard versions ROUGE-1, ROUGE-2 and ROUGE-L for salient sentences definition, referred to as **Salient-R1**, **Salient-R2** and **Salient-RL**, respectively.

We emphasize that none of these pseudo-targets definition makes any access to human supervision. Training SummScore amounts to estimating the coefficients θ in Eq. (7) using the pseudo-targets:

$$\hat{\theta} = \arg\max_{\theta} \sum_{i} \mathcal{R}(\tilde{y}_i, \text{SummScore}_{\theta}(\mathbb{C}_i)) \quad (8)$$

where $\mathcal{R}$ is the mean of ROUGE-1, ROUGE-2 and ROUGE-L, $\mathbb{C}_i$ is the set of candidates predicted by the base model $\mathcal{M}_{\text{base}}$ for data point x_i, and $\tilde{y}_i$ is the pseudo-target. To optimize coefficients, we hill climb with randomness to maximize $\mathcal{R}$ between the SummScore selected summary candidate, and the pseudo-target. Specifically, we estimate coefficients with stochastic local search on the validation set in a hierarchical manner: we first tune coefficients for $S_{overlap}$ and $S_{semantic}$ separately, then estimate coefficients for $S_{quality} \cup \{F_{overlap}, F_{semantic}\}$, where $F_{overlap}$ (resp. $F_{semantic}$) is the set $S_{overlap}$ (resp. $S_{semantic}$) after reduction to a single feature. Such hierarchical estimation is natural given that $S_{overlap}$ (resp. $S_{semantic}$) is made of features capturing similar properties, and dramatically reduces the search space.

4 Experiments

4.1 Setup

We experiment on four popular abstractive summarization datasets, from three different domains (see Table 2 for basic statistics on each dataset):

- **CNN-DailyMail** (Hermann et al., 2015; See et al., 2017) is made of 93k and 220k articles from the CNN and DailyMail newspapers, respectively. CNN/DM is the most extractive dataset among all the ones we consider and has the longest source documents.
- **XSum** (Narayan et al., 2018) has 227k articles from the BBC from 2010 to 2017. This is an extreme summarization task, compressing each article into a single, very abstractive sentence.
- **WikiHow** (Koupaee and Wang, 2018) contains 168k lists of short instructions from Wikipedia.
- **SAMSum** (Gliwa et al., 2019) is a dialogue summarization dataset containing 17k conversations. In this dataset, source length is significantly shorter than in the other datasets.

To estimate coefficients, we subsample randomly (on datasets other than SAMSum) 1,000 data points from the validation set. To avoid coefficients optimization to overfit, we cap each random search at 1,000 trials. Evaluation of summaries selected by SummScore is done with the standard ROUGE-1/2/L (Lin, 2004) (using summary-level ROUGE-LSUM variant for ROUGE-L) and BERTScore (Zhang et al., 2019). We use *transformers* (Wolf et al., 2020) and *datasets* (Lhoest et al., 2021) for pre-trained checkpoints and datasets, respectively.

4.2 Unsupervised Abstractive Summarization

We first apply SummScore to unsupervised abstractive summarization, using as base model ($\mathcal{M}_{\text{base}}$) two models of different capacity: the pre-traind PEGASUS (Zhang et al., 2020) (loading the *google/pegasus-large* checkpoint from *transformers*), and the recently introduced, highly-performing ChatGPT[3], accessed through OpenAI API (calling the *gpt-3.5-turbo* checkpoint). Due to its pre-training objective of generating gap-sentences, PEGASUS can directly be applied to the summarization task after pre-training. This is not the case of comparable sequence-to-sequence Transformer-based models T5 (Raffel et al., 2019) and BART (Lewis et al., 2020), which are pre-trained with token spans generation and sequence de-noising, respectively. For ChatGPT, to lower costs, we subsample randomly 1,000 data points from the test set on datasets other than SAMSum.

[3]https://chat.openai.com/. There is a chance that this checkpoint has been trained on the dataset above.

Backbone $\mathcal{M}_{\text{base}}$	Model Candidate Selection	CNN/DM R-1/R-2/R-L	CNN/DM BS	CNN/DM Gain (%)	XSum R-1/R-2/R-L	XSum BS	XSum Gain (%)	WikiHow R-1/R-2/R-L	WikiHow BS	WikiHow Gain (%)	SAMSum R-1/R-2/R-L	SAMSum BS	SAMSum Gain (%)
PEGASUS	Top beam (Zhang et al., 2020)	32.90/13.28/29.38	_	_	19.27/**3.00**/12.72	_	_	22.59/6.10/14.44	_	_	_	_	_
	Top beam	35.47/13.89/31.61	86.29	_	18.77/2.86/13.85	85.66	_	25.49/5.91/17.99	**84.98**	_	26.64/6.32/22.75	86.12	_
	Random beam	34.89/13.46/31.22	86.11	-1.67	18.58/2.81/13.90	85.29	-1.31	25.39/6.00/18.09	84.82	-0.38	25.27/5.80/21.78	85.31	-5.26
	SummScore - Random-3	35.92†/14.26†/32.34†	86.28	1.96	19.37†/**2.99**†/14.52†	85.78†	3.89	26.29†/**6.28**†/**18.78**†	**84.98**	3.89	28.09†/**7.26**†/**24.42**†	**86.39**†	**7.27**
	SummScore - LEAD-3	**36.92**†/**15.03**†/**33.19**†	**86.54**†	**5.19**	**19.62**†/**3.02**†/**14.71**†	**85.92**†	**5.24**	26.17†/6.19†/18.69†	**84.96**	3.16	**28.22**†/**7.16**/**24.39**†	**86.41**†	**7.27**
	SummScore - Salient-R1	35.54/14.05/32.04†	86.22	0.85	18.96/2.88/14.19†	85.65	1.52	**26.37**†/**6.32**†/**18.81**†	**84.92**	4.25	27.89†/7.08/24.08†	86.25	5.98
	SummScore - Salient-R2	35.65/14.12/32.14†	86.24	1.19	19.13†/**2.96**/14.34†	85.67	2.62	**26.40**†/**6.30**†/**18.83**†	**84.92**	**4.37**	27.93†/7.04/24.14†	86.24	6.09
	SummScore - Salient-RL	35.54/14.05/32.04†	86.22	0.85	19.29†/**2.99**†/14.48†	85.79†	3.63	**26.37**†/**6.32**†/**18.81**†	**84.92**	**4.31**	28.01†/7.08/24.21†	86.21	6.46
ChatGPT	First	40.79/16.61/36.92	**87.93**	_	**30.48**/10.00/**22.16**	88.78	_	29.61/7.28/22.14	**86.28**	_	40.82/15.57/35.15	90.67	_
	Random	40.79/16.61/36.92	**87.93**	0.00	**30.53**/**10.20**/**22.20**	88.77	0.48	29.99/7.57/**22.32**	**86.32**	_	40.60/15.28/34.78	90.63	-0.95
	SummScore - Random-3	41.82†/**18.11**†/37.88†	**87.91**	3.69	27.98/8.45/19.64	87.94	-10.49	30.09/7.85/22.16	86.15	1.78	**42.73**†/**17.45**†/**37.63**†	**90.93**†	**6.86**
	SummScore - LEAD-3	**42.05**†/**18.20**†/**38.06**†	**87.97**	**4.23**	27.97/8.42/19.76	88.05	-10.34	30.14/7.78/**22.22**	86.21	1.88	42.57†/17.29†/**37.54**†	**90.88**†	6.41
	SummScore - Salient-R1	40.30/17.10/36.37	87.67	-0.57	27.84/8.46/19.55	87.91	-10.87	**30.29**/**7.97**†/22.20	86.12	2.41	42.59†/17.26†/37.50†	**90.86**†	6.36
	SummScore - Salient-R2	40.20/17.06/36.23	87.65	-0.88	27.79/8.47/19.57	87.90	-10.87	**30.38**/**8.00**†/**22.27**	86.13	**2.74**	42.43†/17.00†/37.30†	**90.84**	5.67
	SummScore - Salient-RL	40.24/17.06/36.29	87.66	-0.76	27.82/8.51/19.58	87.90	-10.73	**30.29**/**7.97**†/22.20	86.12	2.39	42.59†/17.26†/37.50†	**90.86**†	6.36

Table 3: Unsupervised abstractive summarization results with SummScore re-ranking on the four datasets. Models are decoded to produce 20 summary candidates. **R-1/2/L** denotes ROUGE-1/2/L and **BS** denotes BERTScore. **Gain** represents the mean ROUGE relative gain compared to *our top beam or first candidate baseline*. † marks indicate significantly better results (p-value of paired t-test smaller than 0.05). Best results for each (backbone, dataset) pair within 0.1 are in bold.

We decode PEGASUS with beam search, and ChatGPT with top-p sampling with $p = 0.9$ and temperature 0.8 to enhance diversity, both models with 20 candidates. We report candidate selection baselines from Table 1: *top beam* or *first*, and *random* (a randomly sampled candidate).

We show unsupervised summarization results with PEGASUS and ChatGPT with 20 summary candidates in Table 3. SummScore improves the base PEGASUS by 4.37% to 7.27% across the four datasets. Notably, SummScore fails with ChatGPT on XSum, which we hypothesize is due to the nature of XSum and the fact that pseudo-labels from XSum source documents are too different from the ground truth labels, an issue not affecting PEGASUS because its performance range is far lower than ChatGPT. However, SummScore improves ChatGPT by 2.74% to 6.86% on the other datasets. We point out that SummScore gains are achieved *without using any human supervision*.

SummScore - LEAD-3 performs best for the news domain, which intuitively makes sense due to the lead bias and first sentences containing an overview of the article. On WikiHow, SummScore - Salient-R2 works the best, yet gains are more moderate and SummScore fails to improve the BERTScore on this dataset. SummScore - Random-3 is tied with SummScore - LEAD-3 on SAMSum: we attribute it to the fact that SAMSum source documents are very short (Table 2), and the LEAD-3, Random-3, and entire source document all overlap a lot. Appendix A confirms that SummScore re-ranking always finds a non-trivial (e.g., longest) candidate selection.

4.3 Zero-Shot Transfer

Next, we investigate SummScore performance in the transfer setup, with standard-size models (discarding ChatGPT or similar models). We perform zero-shot summarization inference followed by SummScore on a target dataset where the base model $\mathcal{M}_{\text{base}}$ was fine-tuned on *another* source dataset. As $\mathcal{M}_{\text{base}}$, we use three high-performing summarization models: PEGASUS (Zhang et al., 2020), BART (Lewis et al., 2020), and the recently introduced BRIO (Liu et al., 2022a), which achieves SOTA results on news summarization (CNN/DM & XSum). We use publicly available fine-tuned checkpoints on CNN/DM and XSum, and PEGASUS on WikiHow. We fine-tune ourselves PEGASUS on SAMSum, and BART on WikiHow and SAMSum. Generation and fine-tuning hyper-parameters and results are in Appendix B.

Given the findings from §4.2, we use SummScore - LEAD-3 on CNN/DM, XSum, and SAMSum, and SummScore - Salient-R2 on WikiHow. We tune coefficients in the same process described in §4.1. To stick to a **no supervision** scenario, we do not apply SummScore on a dataset on the which the base model was fine-tuned, which would fall into the supervised learning use case. We compare SummScore zero-shot transfer performance on CNN/DM with that of SOTA WikiTransfer (Fabbri et al., 2021), which fine-tunes BART on external data retrieved from Wikipedia before applying the model in zero-shot summarization.

Zero-shot transfer results are displayed in Table 4. SummScore consistently improves transfer performance, with ROUGE gains of 7.51% averaged over 30 setups: +9.43% on CNN/DM, +1.27% on XSum, +9.20% on WikiHow (up to +17.64% average when transferring from XSum) and +9.61% on SAMSum. Notably, on CNN/DM, BART transferred from SAMSum with SummScore improves on the ROUGE-1 and ROUGE-L of SOTA transfer model WikiTransfer (also using a BART backbone), despite WikiTransfer being fine-tuned on data specifically crafted to transfer better to the downstream task. We notice that SummScore helps

Fine-tuning dataset	Backbone $\mathcal{M}_{\text{base}}$	Candidate Selection	CNN/DM R-1/R-2/R-L	CNN/DM BS	CNN/DM Gain (%)	XSum R-1/R-2/R-L	XSum BS	XSum Gain (%)	WikiHow R-1/R-2/R-L	WikiHow BS	WikiHow Gain (%)	SAMSum R-1/R-2/R-L	SAMSum BS	SAMSum Gain (%)
CNN/DM	PEGASUS	Top beam				21.18/3.44/16.53	85.95	–	24.53/5.68/18.57	84.87	–	31.03/9.05/28.21	86.39	–
		SummScore				21.51†/3.49/16.69	86.05†	1.31	25.87†/6.04†/19.37†	84.94†	5.10	31.98/9.59/28.78	86.56	3.03
	BART	Top beam				20.32/3.10/15.95	86.03	–	26.13/6.03/19.69	85.18	–	30.78/9.60/28.28	86.81	–
		SummScore				20.61†/3.16/16.21†	86.27†	1.60	26.61†/6.24†/20.01†	85.24	1.97	30.77/9.56/28.20	86.87	-0.02
	BRIO	Top beam				23.91/5.41/19.51	87.07	–	29.67/8.01/22.73	**86.04**	–	35.04/13.04/32.42	**89.11**	–
		SummScore				23.72/5.33/19.38	87.06	-0.86	**30.08†/8.17/23.01**	**86.05**	1.39	**35.50/13.35/32.85**	**89.09**	1.50
XSum	PEGASUS	Top beam	23.10/8.03/20.18	85.88	–				15.32/3.54/11.98	85.38	–	23.05/4.75/19.89	87.03	–
		SummScore	26.60†/9.47†/23.13†	86.47†	**15.38**				19.36†/4.52†/14.27†	85.57†	**23.73**	26.82†/6.39†/22.91†	87.39	**17.61**
	BART	Top beam	25.60/8.10/22.16	86.37	–				18.31/4.30/13.71	85.63	–	26.92/5.98/22.20	88.03	–
		SummScore	27.80†/9.21†/23.97†	86.69†	9.18				20.52†/4.92†/14.94†	85.81†	11.24	30.03†/7.28†/24.71†	88.43†	12.52
	BRIO	Top beam	25.52/8.47/22.08	85.97	–				18.39/4.24/13.82	85.58	–	26.69/5.19/22.02	87.16	–
		SummScore	28.67†/9.82†/24.58†	86.42†	12.52				21.94†/5.31†/15.75†	85.66†	17.94	30.10†/7.13†/24.90†	87.62†	15.25
WikiHow	PEGASUS	Top beam	27.55/9.41/24.02	85.20	–	28.05/8.40/21.31	87.86	–				21.15/3.92/17.46	85.44	–
		SummScore	30.49†/10.97†/26.74†	85.95†	11.82	28.10/8.33/21.30	87.92	-0.05				23.62†/4.84†/19.26†	85.95	12.20
	BART	Top beam	29.39/10.52/25.26	85.87	–	23.79/7.19/19.05	87.99	–				19.51/4.52/17.29	87.07	–
		SummScore	31.30†/11.42†/26.72†	86.21†	6.54	25.57†/7.54†/20.11†	**88.18†**	**6.41**				22.48†/5.40/19.63†	87.15	14.80
SAMSum	PEGASUS	Top beam	36.40/15.48/32.52	87.16	–	24.30/6.31/18.75	87.41	–	22.17/5.10/16.29	85.08	–			
		SummScore	**39.15**†/16.89†/35.33†	**87.48**†	8.27	24.10/5.67/18.69	87.31	-1.52	24.44†/5.78†/18.03†	85.15	10.74			
	BART	Top beam	38.40/16.58/35.22	86.93	–	20.78/3.70/15.42	86.49	–	26.00/6.29/19.63	84.73	–			
		SummScore	**39.24**†/17.07†/**35.94**†	87.11†	2.26	21.22†/3.71/15.79†	86.59†	2.03	26.35†/6.43/19.91†	84.75	1.44			
WikiTransfer*		Top beam	39.11/**17.25**/35.73	–	–	**31.85/10.44/23.75**	–	–	–	–	–	–	–	–

Table 4: Zero-shot transfer results with SummScore re-ranking, across all twelve transfer directions over the four summarization datasets. Each model is decoded with beam search with 20 beams. **Top beam** refers to the base model performance, while **SummScore** is the candidate re-ranked by SummScore. **R-1/2/L** is ROUGE-1/2/L, **BS** denotes BERTScore, and **Gain (%)** is the relative mean ROUGE improvement compared to the base model performance. † marks indicate significantly better results (p-value of paired t-test smaller than 0.05). Best results within 0.1 are in bold. Greyed out cells correspond to the supervised setup, which is excluded. *WikiTransfer (Fabbri et al., 2021) is not directly comparable due to constructing the fine-tuning dataset specifically to optimize transfer to the downstream task.

more when the base model transfers less well, such as from single-sentence summaries XSum.

Appendix C evalutes re-ranking itself and shows that SummScore can also reach strong recall.

4.4 Self-Training with Unsupervised Paraphrasing

Using the selected summary candidate as a pseudo-target, one can naturally extend SummScore into a self-training summarization objective. Indeed, if γ parametrizes $\mathcal{M}_{\text{base}}$, we can further train $\mathcal{M}_{\text{base}}$ through the objective:

$$\tilde{\gamma} = \arg\max_{\gamma} \sum_{i} \log \left(p(\text{SummScore}(\mathbb{C}_i)|x_i;\gamma) \right) \quad (9)$$

This process can be repeated: if we denote new model weights by γ^k, we can re-apply SummScore and perform another round of self-training, yielding new model weights γ^{k+1}.

We notice that the unsupervised PEGASUS beam search summary candidates, including the one selected by SummScore, are quite extractive (see Appendix D). This could be because the self-supervised gap-sentences are extracts from the source document. To make the pseudo-summaries more abstractive and diverse enough to mitigate the confirmation bias in self-training (Tarvainen and Valpola, 2017), we use the paraphrasing approach proposed in FAR-RW (Zhang et al., 2022). On each dataset, we train a paraphrase model to generate the top n sentences maximizing the mean ROUGE with the top n most salient sentences, conditioning on these salient sentences. This yields an unsupervised, in-domain paraphrase model which we apply to the SummScore pseudo-labels on the training set to make them more abstractive and diverse. We refer to Appendix E for details on the paraphrasing model training, its performance and resulting abstractiveness and diversity levels on pseudo-labels. As the unsupervised process of paraphrasing may harm the pseudo-summary quality, in practice, we apply it to the x% most extractive training data points, where x is among {12.5%, 25%, 50%, 100%}. We use 25% for CNN/DM, 100% for XSum, 50% for WikiHow, and 12.5% on SAMSum, as these provide an ideal ROUGE/abstractiveness trade-off (see Appendix D).

For each dataset except SAMSum, we randomly subsample 50k data points from the training set and 1k from the validation set to self-train and validate the model, resulting in a self-training process much less computationally expensive than fine-tuning. We show self-training results on the test sets using PEGASUS as base model in Table 5. Self-training improves unsupervised summarization performance on all datasets, resulting in a self-trained model better than the base model although not as performing as SummScore. Notably, re-applying SummScore on the new model after self-training further improves performance drastically. Besides, paraphrasing self-training pseudo-labels helps maintain some degree of abstractiveness, as seen in Appendix D. On CNN/DM, one round of self-training followed by SummScore brings PE-

Dataset	Model	R-1	R-2	R-L	BS
CNN/DM	PEGASUS (Zhang et al., 2020)	32.90	13.28	29.38	_
	Summary Loop 45 (Laban et al., 2020)	37.70	14.80	34.70	_
	TED (Yang et al., 2020)	38.73	16.84	35.40	_
	FAR-RW* (Zhang et al., 2022) (SOTA)	**40.13**	**17.00**	**36.34**	_
	PEGASUS (ours)	35.47	13.89	31.61	86.29
	PEGASUS (ours) + SummScore	36.92	15.03	33.19	86.54
	Self-training (1st round)	36.68	14.52	32.72	86.49
	Self-training (1st round) + SummScore	38.75	16.11	34.78	86.88
	Self-training (2nd round)	38.17	15.77	34.25	86.87
	Self-training (2nd round) + SummScore	39.49	16.69	35.61	87.07
	Self-training (3rd round)	38.47	15.95	34.48	87.00
	Self-training (3rd round) + SummScore	39.76	16.79	35.85	**87.18**
XSum	PEGASUS (ours)	18.77	2.86	13.85	85.66
	PEGASUS (ours) + SummScore	19.62	**3.02**	14.71	85.92
	Self-training	19.33	2.76	14.18	86.03
	Self-training + SummScore	**20.02**	2.84	**14.93**	**86.23**
WikiHow	PEGASUS (ours)	25.49	5.91	17.99	**84.98**
	PEGASUS (ours) + SummScore	**26.40**	**6.30**	18.83	**84.92**
	Self-training	26.08	6.08	18.59	84.89
	Self-training + SummScore	**26.50**	**6.28**	**19.03**	**84.93**
SAMSum	PEGASUS (ours)	26.64	6.32	22.75	86.12
	PEGASUS (ours) + SummScore	28.22	7.16	24.39	86.41
	Self-training	26.96	6.41	23.40	86.25
	Self-training + SummScore	**28.91**	**7.55**	**25.54**	**86.58**

Table 5: Unsupervised abstractive summarization results with SummScore re-ranking and *self-training* for PEGASUS on the four datasets. We fine-tune the model with the unsupervised summary candidate which was selected by SummScore as pseudo-target, then apply again SummScore on the output. All models are decoded with beam search with 20 beams. **R-1/2/L** is ROUGE-1/2/L, and **BS** denotes BERTScore. Best results within 0.1 are in bold. *FAR-RW pipeline is not directly comparable due to relying on a SOTA unsupervised extractive summarization model first, then applying re-writing.

Use case	Attribute	PEGASUS	SummScore	Tie
Unsupervised abs. summ.	Informativeness	11.33 (1.15)	**20.67** (6.43)	18.00 (6.93)
	Factual consistency	14.67 (4.04)	**19.33** (5.03)	16.00 (9.00)
0-shot transfer from XSum	Informativeness	5.67 (2.89)	**24.00** (2.00)	20.33 (1.53)
	Factual consistency	4.67 (4.51)	18.67 (4.04)	**26.67** (3.51)

Table 6: Human evaluation on CNN/DM with PEGASUS. Mean number of times out of 50 that each model or a tie is selected, with standard deviation in parenthesis, across two use cases and two attributes.

GASUS performance above the Summary Loop, two rounds above TED, and three rounds to 39.76 ROUGE-1, within 1% of SOTA model FAR-RW.

4.5 Human Evaluation

We conduct a human evaluation on 50 data points randomly sampled from CNN/DM test set. We show human participants the source news article, alongside the summary candidate from the base PEGASUS model, and the one re-ranked by SummScore. Participants are asked to pick which summary is more informative, and which is more factually consisteny, with the option of choosing a tie. We cover two use cases: unsupervised abstractive summarization, and zero-shot transfer from a model fine-tuned on XSum. In the former use case, both summaries are identical in 7/50 data points, and 4/50 data points in the latter. Human raters are three volunteer graduate students, with full professional command of English. Results are displayed in Table 6. Although both summaries often overlap significantly (rightmost column), resulting in a high *Tie*, SummScore is strongly preferred over PEGASUS across both use cases and attributes.

Candidate selection	Dataset				Average
	CNN/DM	XSum	WikiHow	SAMSum	
PEGASUS	26.99	11.83	16.46	18.57	18.46
ROUGE-1 with source	26.90	12.03	**17.21**	**19.89**	19.01
ROUGE-2 with source	26.98	11.93	**17.16**	19.62	18.92
BLEU with source	26.90	11.99	**17.19**	**19.94**	19.01
BERTScore with source	28.19	**12.42**	**17.11**	19.43	19.29
BARTScore with source	28.11	12.23	16.60	19.70	19.16
BLEURT with source	27.45	12.12	16.79	19.69	19.01
Diversity score	25.33	11.36	14.52	15.67	16.72
Length score	27.07	11.67	16.66	18.60	18.50
Plain average	27.75	12.28	16.96	19.73	19.18
Random coefficients	27.75	12.25	16.84	19.72	19.14
SummScore	**28.38**	**12.45**	**17.18**	**19.92**	**19.48**

Table 7: Ablation study for unsupervised abstractive summarization with PEGASUS. We isolate each feature of SummScore and report its re-ranking performance (picking the candidate maximizing this feature), using the mean of ROUGE-1/2/L as reported metric. Best results within 0.1 are in bold.

5 Analysis

5.1 Ablation

To better understand SummScore performance gains, we perform an ablation study where re-ranking is done with each feature taken individually. Results for PEGASUS in unsupervised summarization are shown in Table 7. N-gram overlap features are very strong re-ranking baselines on WikiHow and SAMSum. In fact, ROUGE-1 with the source is even slightly better than SummScore on WikiHow. On news datasets, semantic similarity features such as BERTScore are strong baselines. Interestingly, our hand-crafted feature *diversity* has a *negative* contribution when used as standalone re-ranker ; however it can help a lot when combined with the other features, acting as a regularizer by encouraging some diversity. On average, SummScore performs the best. We also report trivial feature aggregation baselines *Plain average* and *Random coefficients*, which SummScore outpeforms, confirming the efficiency of estimating coefficients through pseudo-labels.

In Appendix F, we show that SummScore unsupervised re-ranking is also robust to other decoding methods diverse beam search (Vijayakumar et al., 2016) and nucleus sampling (Holtzman et al., 2019), and a different number of beams (5 to 20). We confirm that our default setup of beam search with 20 beams yields optimal ROUGE results. Echoing SummaReranker findings (Ravaut et al., 2022a), gains further increase when mixing in several decoding methods.

Source document: Reports speak of at least four people injured. The city is at the heart of the conflict between the Turkish government and Kurdish separatists. Interior Minister Suleyman Soylu said the blast happened at a vehicle repair unit, and appeared to be an accident. He said "it seems there is no outside interference, and the explosion came from the vehicle under repair". Mr Soylu said one person was trapped under rubble, another was seriously injured, and others had minor injuries. The blast brought a roof down, left a huge crater and a pall of smoke drifted over part of the city. The cause remains unclear. The banned Kurdistan Workers' Party (PKK) is active in the area. Turkey is five days away from a key referendum on granting President Recep Tayyip Erdogan sweeping new powers [...]
PEGASUS summary (ROUGE-1**: 10.53**): Interior Minister Suleyman Soylu said the blast happened at a vehicle repair unit, and appeared to be an accident.
Self-training summary (ROUGE-1**: 32.43**): The blast happened at a vehicle repair unit in the city of Diyarbakir, near the border with Syria.
Ground truth summary: A large explosion has struck a police headquarters in the mainly Kurdish city of Diyarbakir in south-eastern Turkey.

Table 8: Qualitative sample with self-training PEGASUS from the XSum dataset, after a single round of self-training.

5.2 Qualitative Samples

We refer to Appendix H for full qualitative unsupervised re-ranking examples on all datasets, and to Table 8 for an example of summary generated by the self-trained PEGASUS model on XSum. As seen, both re-ranking and self-training can improve dramatically from the unsupervised PEGASUS baseline, capturing entirely new phrases.

5.3 Factual Consistency

As noted in Table 6, SummScore summaries tend to be more factually consistent than the baseline. There is strong intuition to this result: since SummScore is built to maximize features of n-gram overlap and semantic similarity with the source, it should yield summaries closer to the source, and more factually consistent as a result. We investigate this further, and use two popular models to evaluate summarization factuality: the established *factCC* (Kryscinski et al., 2020) and the recently introduced state-of-the-art *QAFactEval* (Fabbri et al., 2022). factCC uses a BERT model to classify each summary sentence as consistent or inconsistent with regards to the source, and reports the average accuracy over 100%. QAFactEval improves each step of the QA evaluation pipeline (answer selection, question generation, etc) and combines entailment with QA-based metrics into a learned metric. In Table 9, we observe that SummScore QAFactEval is consistently above PEGASUS, and SummScore factCC is better on news datasets too.

5.4 Learned Coefficients

We analyze coefficients learned by SummScore from a high level perspective in Table 10, gathering features from a same group together. Semantic similarity features are dominating (except for WikiHow), encouraging further research using newer semantic similarity metrics for re-ranking.

A finer-grain analysis, covering all SummScore pseudo-labeling techniques, can be viewed in Tables 19 and 20 of Appendix G. SummScore - Salient-R1 and SummScore - Salient-RL place much more emphasis on n-gram overlap with the source. In contrast, SummScore - LEAD-3 (which we use for self-training on CNN/DM, XSum, SAMSum) uses relatively more semantic similarity features like BERTScore, suggesting that it is able to exploit key semantic content contained in initial sentences.

Dataset	Factual consistency model	PEGASUS	SummScore
CNN/DM	factCC	92.45	**93.66**
	QAFactEval	4.53	**4.55**
XSum	factCC	96.78	**97.53**
	QAFactEval	4.54	**4.64**
WikiHow	factCC	**96.48**	95.85
	QAFactEval	4.33	**4.36**
SAMSum	factCC	**98.35**	96.28
	QAFactEval	3.26	**3.50**

Table 9: Factual consistency evaluation of SummScore with PEGASUS in unsupervised abstractive summarization. We use the entire test set for *factCC*, and a random sample of 500 test data points for *QAFactEval*.

Dataset	PEGASUS			ChatGPT		
	N-gram	Semantic	Quality	N-gram	Semantic	Quality
CNN/DM	0.025	**0.900**	0.075	0.100	**0.775**	0.125
XSum	0.050	**0.950**	0.000	0.250	**0.725**	0.025
WikiHow	**0.875**	0.100	0.025	**0.900**	0.100	0.000
SAMSum	0.000	**1.000**	0.000	0.000	**1.000**	0.000

Table 10: Coefficients learned by SummScore in unsupervised abstractive summarization. We sum weights assigned to all features of each category defined in §3.2.

6 Conclusion

We introduced SummScore, the first unsupervised abstractive summarization re-ranking system. SummScore does not rely on a neural network: instead, it builds features for each summary candidate, some of them using the source as well, and aggregates them into a final re-ranking score. Feature coefficients are estimated through tuning against a pseudo-label derived from the source document. It is a simple framework which easily supports the addition of new features.

SummScore significantly improves the performance of the base summarization model, in terms of ROUGE, BERTScore, factual consistency, and human preference ; in both unsupervised and zero-shot transfer scenarios. Moreover, SummScore selected summary candidate naturally extends into a self-training objective for abstractive summarization, which improves unsupervised summarization.

Limitations

As a second-stage method, SummScore requires access to a base abstractive summarization model generating summary candidates. Generating up to 20 summary candidates per data point can take a long time, especially on training sets, which is needed for the self-training use case. Besides, even though SummScore does not need to train a new neural network, we also need to generate all eight features for each summary candidate once all candidates are generated. N-gram overlap features are very fast, but model-based semantic similarity features (e.g, BERTScore) can be time-consuming to extract, once again, especially on entire training sets.

While SummScore will significantly improve the quality of the base model across base models and datasets, ultimately, the performance of the final selected summary is bounded by the capacity of this base model: SummScore improves more PEGASUS than it does on ChatGPT ; but PEGASUS performance drags ChatGPT.

Another limitation lays in the metric used to compare summary candidates with the pseudo-target. We used mean ROUGE, although a model-based semantic similarity metric would make sense too, but at a much greater computational cost.

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 thank anonymous reviewers for a fruitful discussion, especially with regards to evaluation of the factual consistency. We also thank Florian Le Bronnec and Jiajing Zhang for their proofreading.

References

Christos Baziotis, Ion Androutsopoulos, Ioannis Konstas, and Alexandros Potamianos. 2019. SEQ^3: Differentiable sequence-to-sequence-to-sequence autoencoder for unsupervised abstractive sentence compression. 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 673–681, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Alexander Fabbri, Simeng Han, Haoyuan Li, Haoran Li, Marjan Ghazvininejad, Shafiq Joty, Dragomir Radev, and Yashar Mehdad. 2021. Improving zero and few-shot abstractive summarization with intermediate fine-tuning and data augmentation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 704–717, Online. Association for Computational Linguistics.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

Yang Gao, Wei Zhao, and Steffen Eger. 2020. SUPERT: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1347–1354, Online. Association for Computational Linguistics.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

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.

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

Mahnaz Koupaee and William Yang Wang. 2018. Wikihow: A large scale text summarization dataset. *arXiv preprint arXiv:1810.09305*.

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.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Wojciech Kryściński, Romain Paulus, Caiming Xiong, and Richard Socher. 2018. Improving abstraction in text summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1808–1817, Brussels, Belgium. Association for Computational Linguistics.

Philippe Laban, Andrew Hsi, John Canny, and Marti A. Hearst. 2020. The summary loop: Learning to write abstractive summaries without examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5135–5150, Online. 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 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.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. 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.

Peter J Liu, Yu-An Chung, and Jie Ren. 2019a. Summae: Zero-shot abstractive text summarization using length-agnostic auto-encoders. *arXiv preprint arXiv:1910.00998*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. 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.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022a. 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.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022b. Brio: Bringing order to abstractive summarization. *arXiv preprint arXiv:2203.16804*.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. 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*.

Mathieu Ravaut, Shafiq Joty, and Nancy Chen. 2022a. 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.

Mathieu Ravaut, Shafiq Joty, and Nancy Chen. 2022b. Towards summary candidates fusion. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8488–8504, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Raj Reddy. 1977. Speech understanding systems: A summary of results of the five-year research effort at carnegie mellon university. *Pittsburgh, Pa*.

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.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

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

Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. 2022. Follow the wisdom of the crowd: Effective text generation via minimum bayes risk decoding. *arXiv preprint arXiv:2211.07634*.

Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 1195–1204, Red Hook, NY, USA. Curran Associates Inc.

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*.

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.

Ziyi Yang, Chenguang Zhu, Robert Gmyr, Michael Zeng, Xuedong Huang, and Eric Darve. 2020. TED: A pretrained unsupervised summarization model with theme modeling and denoising. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1865–1874, 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. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhihao Zhang, Xinnian Liang, Yuan Zuo, and Zhoujun Li. 2022. Unsupervised abstractive summarization via sentence rewriting. *Computer Speech & Language*, page 101467.

A Overlap with Simple Baselines

Simple Candidate Selection	CNN/DM	XSum	WikiHow	SAMSum
Max R-1 w. source	16.33	21.38	60.36	38.71
Max R-2 w. source	21.45	24.32	**66.94**	44.93
Max BLEU w. source	16.68	18.92	58.44	38.95
Max BS w. source	43.46	**69.98**	35.50	**58.61**
Max BaS w. source	**47.15**	46.06	13.85	52.50
Max BRT w. source	14.74	13.43	15.75	23.32
Max diversity feature	5.40	5.95	1.45	4.40
Max length feature	11.80	7.46	14.19	13.68
Top beam	15.05	12.85	9.18	30.28
Oracle candidate	15.24	12.27	10.73	18.44
Worst candidate	5.33	7.48	7.65	7.20
Longest candidate	20.58	22.74	64.43	51.28

Table 11: Overlap with simple re-reranking methods (%) in unsupervised abstractive summarization with PEGASUS. We report the fraction (in percentage) of test set data points on the which SummScore falls back to a trivial summary candidate selection: maximizing one of the input features, picking the top beam, one oracle or worst candidate, or the longest one. All setups are with beam search with 20 candidates, thus a random baseline corresponds to 5% overlap.

We perform a sanity check counting the percentage of time that SummScore falls back to a *trivial* method of re-ranking summary candidates. For each feature described in §3.2, we report the overlap between SummScore and a re-ranking approach consisting in picking the summary candidate maximing this feature. We also report baselines consisting in picking the top beam, an oracle or a *worst* candidate, and the longest candidate. As seen in Tables 11 and 12, across both backbones PEGASUS and ChatGPT, SummScore never collapses

Simple Candidate Selection	CNN/DM	XSum	WikiHow	SAMSum
Max R-1 w. source	16.00	32.10	58.70	14.53
Max R-2 w. source	33.50	50.30	**79.80**	17.34
Max BLEU w. source	17.80	31.20	57.10	12.58
Max BS w. source	**54.50**	**75.50**	44.80	24.05
Max BaS w. source	52.20	26.50	24.60	**54.09**
Max BRT w. source	10.20	14.60	14.20	29.79
Max diversity feature	9.60	1.90	1.00	3.30
Max length feature	3.50	0.80	2.10	11.48
Oracle candidate	9.00	1.80	9.00	12.21
Worst candidate	4.80	12.50	6.10	3.17
Longest candidate	10.90	22.70	39.60	6.47

Table 12: Overlap with simple re-reranking methods (%) in unsupervised abstractive summarization with ChatGPT. We report the fraction (in percentage) of test set data points on the which SummScore falls back to a trivial summary candidate selection: maximizing one of the input features, picking one oracle or worst candidate, or the longest one. All setups are with beam search with 20 candidates, thus a random baseline corresponds to 5% overlap.

to a trivial candidate selection, and we see similar patterns on the same dataset (e.g., highest overlap with a single feature selection is with BERTScore with source feature on CNN/DM).

B Generation & Fine-Tuning Details

In Table 13, we show generation hyper-parameters used for each dataset to generate beam search summary candidates used in Table 3. For the transfer setup shown in Table 4, we use as generation hyper-parameters on each target dataset the parameters used on that dataset for Table 3. For instance, PEGASUS-XSum, PEGASUS-WikiHow and PEGASUS-SAMSum, when transferred to CNN/DM, are decoded with hyper-parameters of PEGASUS-CNN/DM shown in Table 13.

Dataset	Model	Max source length	Max target length	Length penalty	Trigram blocking
CNN/DM	PEGASUS	1024	128	0.8	Yes
	BART			1.0	Yes
	BRIO			1.0	Yes
XSum	PEGASUS	512	64	0.8	Yes
	BART			1.0	Yes
	BRIO			0.8	Yes
WikiHow	PEGASUS	512	128	0.6	No
	BART			1.0	Yes
SAMSum	PEGASUS	512	64	0.8	No
	BART			1.0	Yes

Table 13: Generation hyper-parameters for each dataset and model used to produce summary candidates.

For experiments shown in Table 4, we fine-tune ourselves BART on WikiHow dataset, and PEGASUS and BART on SAMSum dataset. Fine-tuning hyper-parameters are shown in Table 14. We perform early stopping with regards to the mean ROUGE on the validation set. Our BART reaches 44.21/19.31/34.67 ROUGE-1/2/L on WikiHow test set, our PEGASUS 52.33/27.97/44.02 ROUGE-1/2/L on SAMSum test set, and our BART 52.78/28.28/44.08 ROUGE-1/2/L.

Dataset	Model	Epochs	Optimizer	Scheduler	LR	BS	LS	Eval steps
WikiHow	BART	15	Adam	none	1e-5	80	0.1	250
SAMSum	PEGASUS	30	Adam	none	1e-4	256	0.1	50
	BART	30	Adam	linear	1e-5	80	0.1	50

Table 14: Fine-tuning hyper-parameters used to fine-tune BART on WikiHow and PEGASUS and BART on SAMSum.

C Recall Analysis

Besides the quality of the selected summary, we also analyze re-ranking performance itself. In Fig. 2, Fig. 3, Fig. 4 and Fig. 5, we show recall curves on each dataset and for all unsupervised and zero-shot summarization setups. Recall@k is defined as the probability of outputting *one* of the oracle summary candidates (candidates maximizing the mean ROUGE with the target) among the first k candidates. We compare SummScore with the baseline beam search output, and a random candidate selection baseline.

In most cases, SummScore (green curves) provides higher recall, with the notable exception of XSum, where both beam search and SummScore and XSum can fail to improve the random baseline.

D Abstractiveness Analysis

In Table 15, we show ROUGE results from Table 5 alongside abstractiveness results, as measured per the fraction of novel n-grams in output summaries, for re-ranking and self-training experiments. Maximizing both ROUGE and abstractiveness is notoriously difficult, as easy solutions for abstractiveness optimization can deviate a lot from the source, resulting in a harmed ROUGE score.

The unsupervised PEGASUS (first row of each block) is very extractive and only produces a small fraction of novel n-grams. SummScore selected summaries, despite maximizing a score which maximizes the mean ROUGE with pseudo-labels extracted from the source document, both improve the ROUGE and the abstractiveness level. However, SummScore re-ranking applied to self-trained models tends to reduce their abstractiveness level, although it stays above the level of the baseline PEGASUS. Paraphrased summaries drastically increase abstractiveness, at the expense of ROUGE -

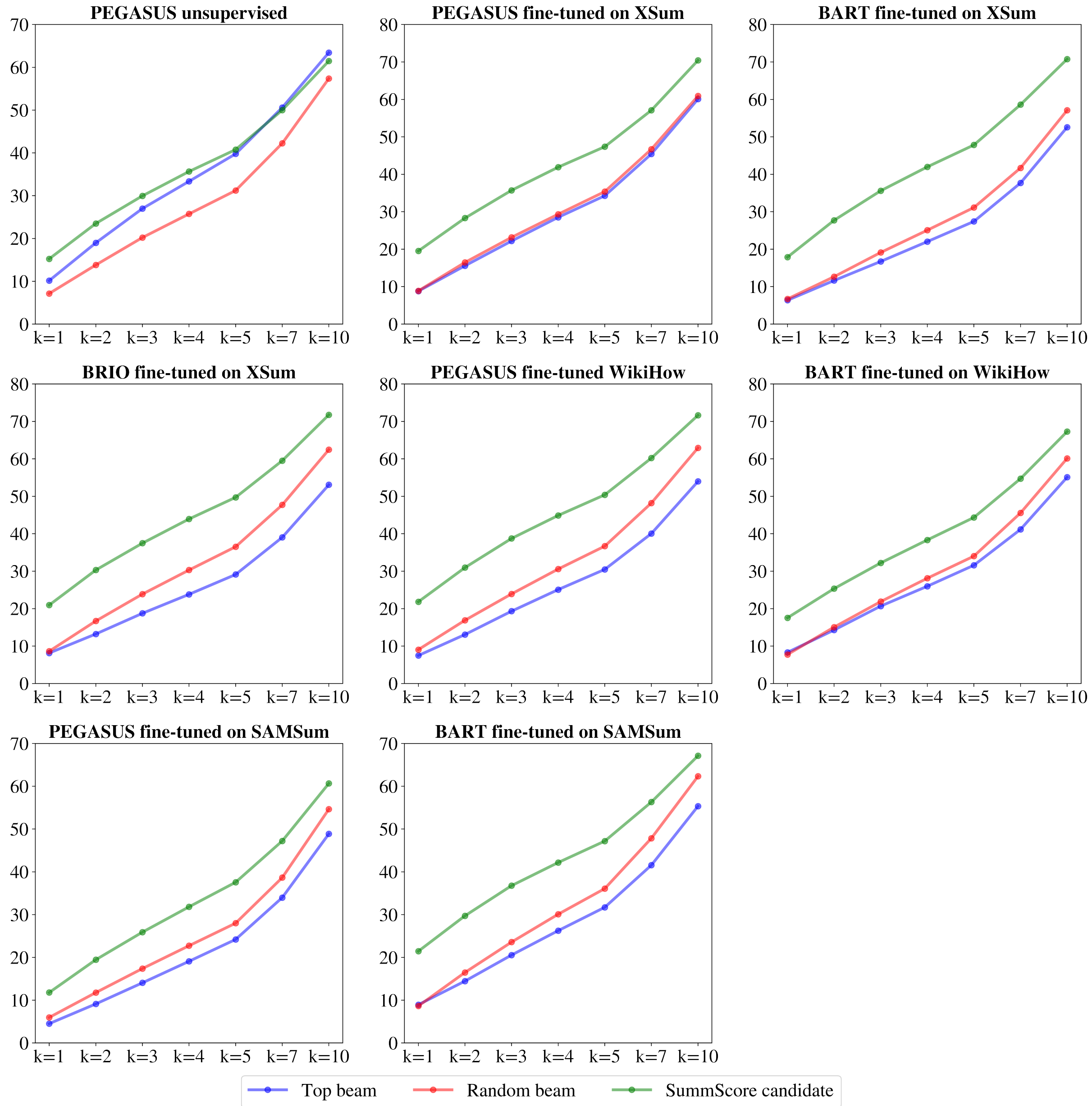

Figure 2: Recall curves on CNN/DM with PEGASUS backbone. The top left plot corresponds to unsupervised summarization re-ranking from Table 3, and the next seven plots to all zero-shot transfer summarization setups from Table 4. Each re-ranking setup has 20 summary candidates, and we show recall over *any oracle candidate* for several thresholds $k \in \{1, 2, 3, 4, 5, 7, 10\}$.

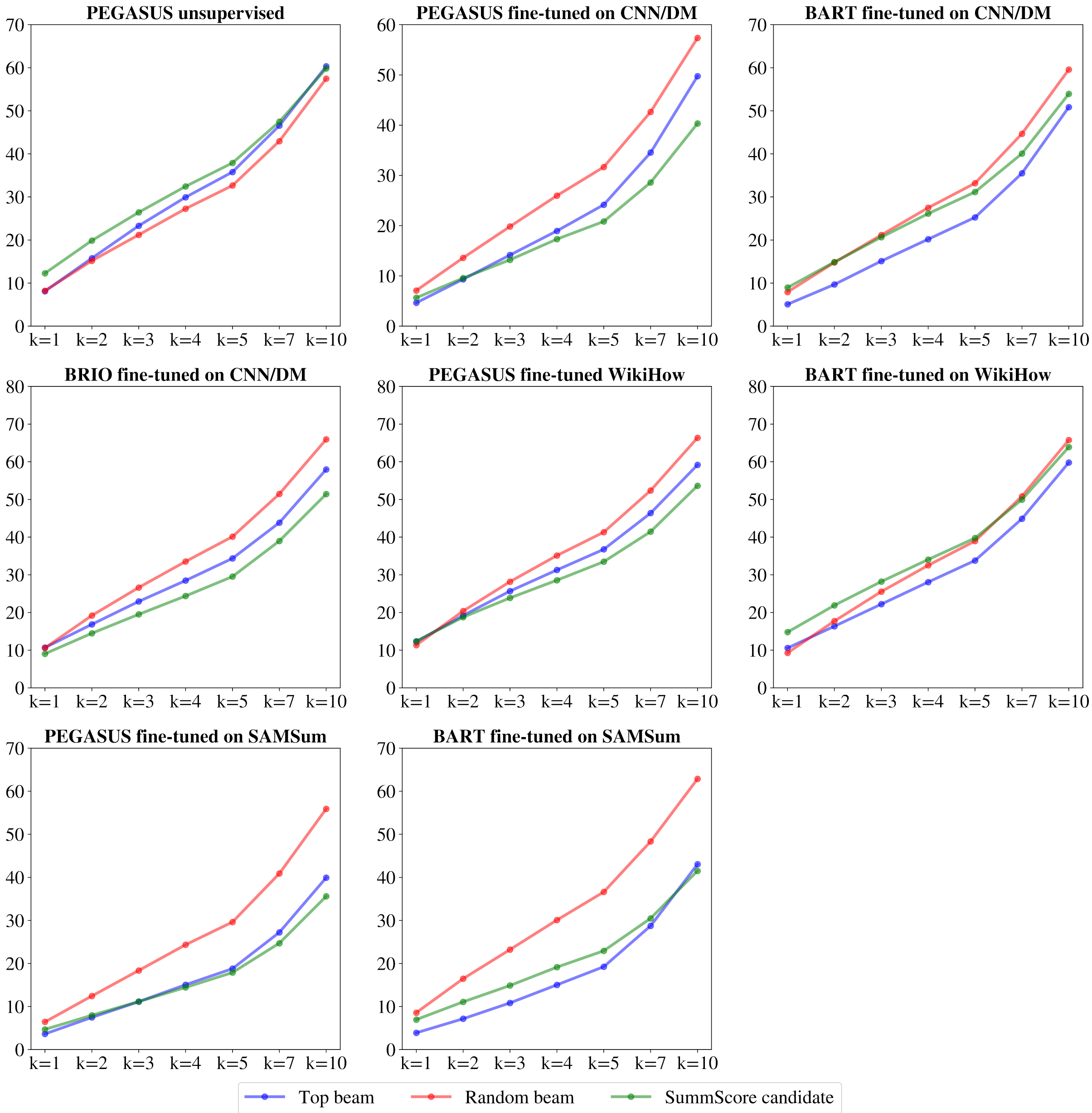

Figure 3: Recall curves on XSum with PEGASUS backbone. The top left plot corresponds to unsupervised summarization re-ranking from Table 3, and the next seven plots to all zero-shot transfer summarization setups from Table 4. Each re-ranking setup has 20 summary candidates, and we show recall over *any oracle candidate* for several thresholds $k \in \{1, 2, 3, 4, 5, 7, 10\}$.

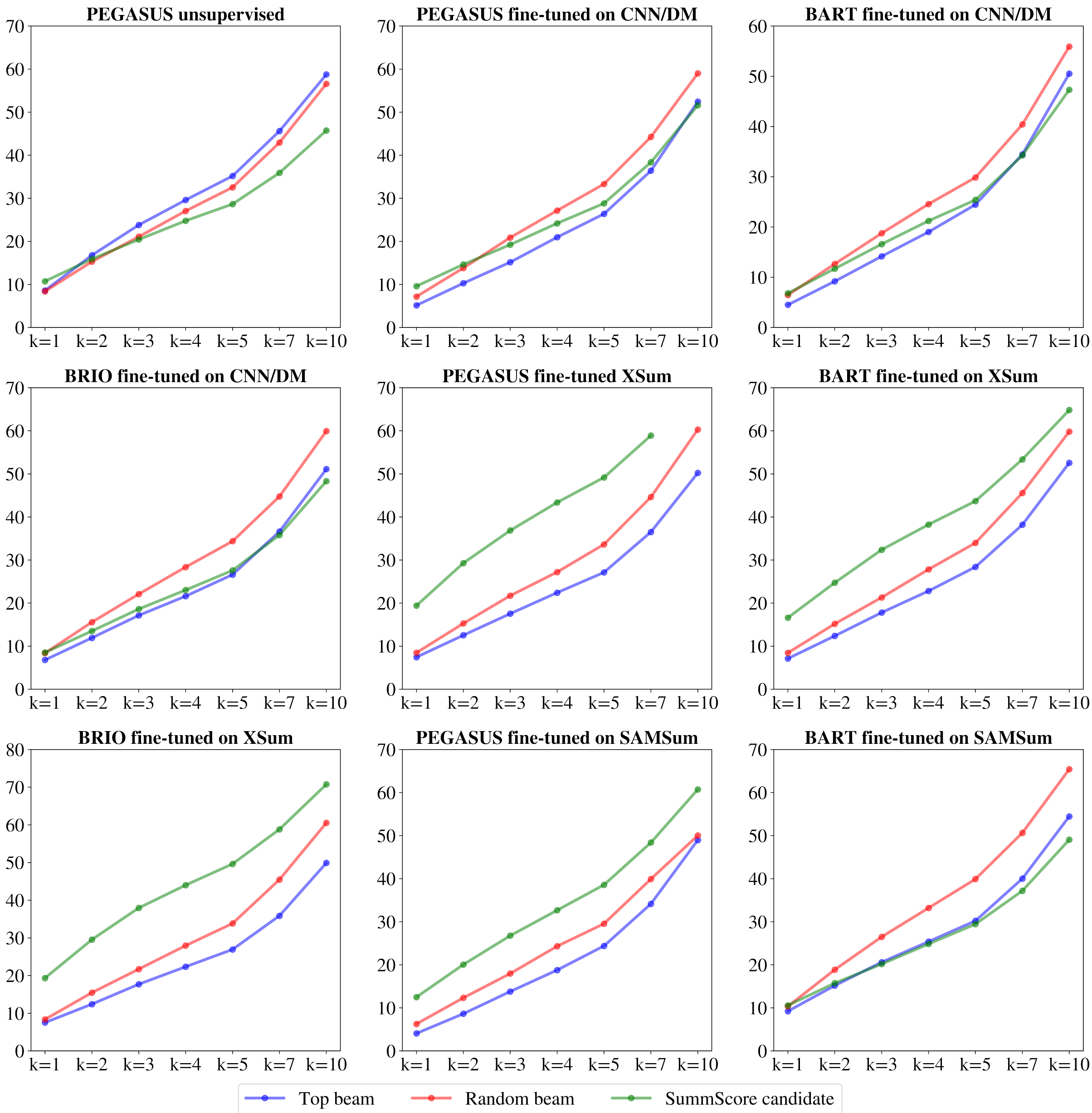

Figure 4: Recall curves on WikiHow with PEGASUS backbone. The top left plot corresponds to unsupervised summarization re-ranking from Table 3, and the next eight plots to all zero-shot transfer summarization setups from Table 4. Each re-ranking setup has 20 summary candidates, and we show recall over *any oracle candidate* for several thresholds $k \in \{1, 2, 3, 4, 5, 7, 10\}$.

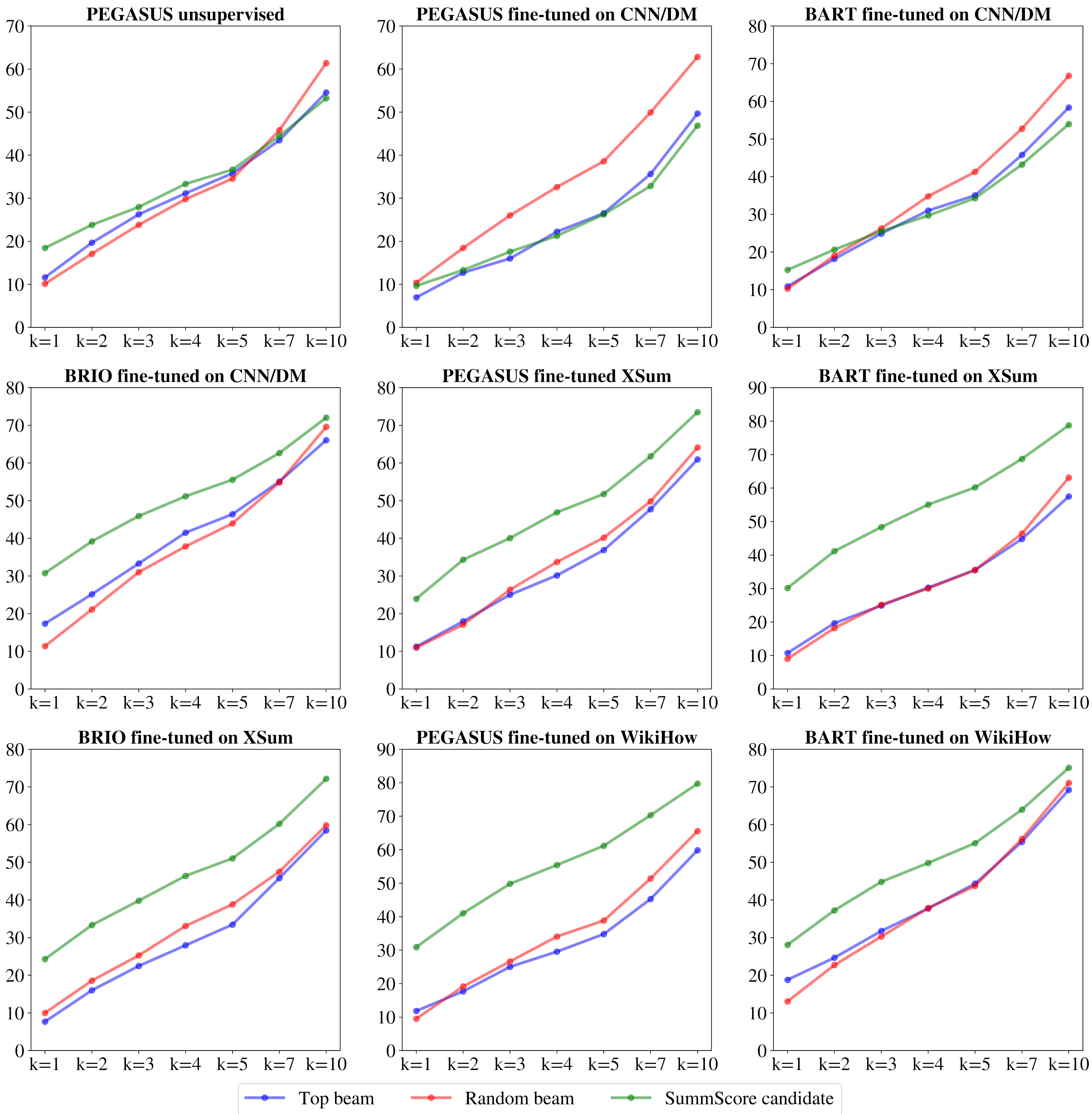

Figure 5: Recall curves on SAMSum with PEGASUS backbone. The top left plot corresponds to unsupervised summarization re-ranking from Table 3, and the next eight plots to all zero-shot transfer summarization setups from Table 4. Each re-ranking setup has 20 summary candidates, and we show recall over *any oracle candidate* for several thresholds $k \in \{1, 2, 3, 4, 5, 7, 10\}$.

Dataset	Model	ROUGE				Abstractiveness (new n-grams)		
		Mean R	R-1	R-2	R-L	New 1-grams	New 2-grams	New 3-grams
CNN/DM	PEGASUS	26.99	35.47	13.89	31.61	0.19	0.89	2.44
	PEGASUS + SummScore LEAD-3	28.38	36.92	15.03	33.19	0.19	0.94	2.73
	PEGASUS + SummScore LEAD-3 - paraphrasing 100%	22.46	29.72	11.07	26.58	**14.01**	**35.18**	**44.23**
	PEGASUS + SummScore LEAD-3 - paraphrasing 50%	25.37	33.24	13.02	29.83	7.29	18.34	23.77
	PEGASUS + SummScore LEAD-3 - paraphrasing 25% (**pseudo-labels**)	<u>26.85</u>	<u>35.06</u>	<u>13.99</u>	<u>31.49</u>	<u>3.73</u>	<u>9.71</u>	<u>13.36</u>
	PEGASUS + SummScore LEAD-3 - paraphrasing 12.5%	27.61	35.99	14.50	32.35	1.95	5.29	7.98
	PEGASUS self-trained (1st round)	27.98	36.68	14.52	32.72	0.25	0.66	1.84
	PEGASUS self-trained (1st round) + SummScore LEAD-3	29.88	38.75	16.11	34.78	0.10	0.43	1.60
	PEGASUS self-trained (2nd round)	29.40	38.17	15.77	34.25	0.66	1.49	2.61
	PEGASUS self-trained (2nd round) + SummScore LEAD-3	30.59	39.49	**16.69**	35.61	0.21	0.93	2.15
	PEGASUS self-trained (3rd round)	29.63	38.47	15.95	34.48	0.68	1.72	2.74
	PEGASUS self-trained (3rd round) + SummScore LEAD-3	**30.80**	**39.76**	**16.79**	**35.85**	0.11	0.99	2.25
XSum	PEGASUS	11.83	18.77	2.86	13.85	0.20	0.44	1.16
	PEGASUS + SummScore LEAD-3	12.45	19.62	3.02	14.71	0.19	0.60	2.04
	PEGASUS + SummScore LEAD-3 - paraphrasing 100% (**pseudo-labels**)	<u>**12.98**</u>	<u>**20.19**</u>	<u>**3.60**</u>	<u>**15.16**</u>	<u>**12.94**</u>	<u>**30.30**</u>	<u>**37.63**</u>
	PEGASUS + SummScore LEAD-3 - paraphrasing 50%	12.75	19.94	3.32	14.97	6.55	15.46	19.87
	PEGASUS + SummScore LEAD-3 - paraphrasing 25%	12.61	19.79	3.18	14.86	3.41	8.06	10.96
	PEGASUS + SummScore LEAD-3 - paraphrasing 12.5%	12.52	19.71	3.10	14.77	1.83	4.36	6.53
	PEGASUS self-trained	12.09	19.33	2.76	14.18	1.49	3.20	4.43
	PEGASUS self-trained + SummScore LEAD-3	12.60	20.02	2.84	14.93	0.66	1.99	3.55
WikiHow	PEGASUS	16.46	25.49	5.91	17.99	0.48	1.12	2.36
	PEGASUS + SummScore R-2	**17.17**	26.40	**6.30**	18.83	0.80	2.47	5.05
	PEGASUS + SummScore R-2 - paraphrasing 100%	16.75	25.59	6.19	18.47	**4.65**	**17.13**	**26.14**
	PEGASUS + SummScore R-2 - paraphrasing 50% (**pseudo-labels**)	<u>16.97</u>	<u>26.01</u>	<u>**6.26**</u>	<u>18.62</u>	<u>2.79</u>	<u>9.82</u>	<u>15.55</u>
	PEGASUS + SummScore R-2 - paraphrasing 25%	17.08	26.24	**6.27**	18.73	1.81	6.14	10.28
	PEGASUS + SummScore R-2 - paraphrasing 12.5%	17.13	26.32	**6.28**	18.79	1.31	4.34	7.71
	PEGASUS self-trained	16.92	26.08	6.08	18.59	0.84	1.80	3.56
	PEGASUS self-trained + SummScore R-2	**17.27**	**26.50**	**6.28**	**19.03**	0.61	1.71	4.02
SAMSum	PEGASUS	18.57	26.64	6.32	22.75	0.30	1.35	2.81
	PEGASUS + SummScore LEAD-3	19.92	28.22	7.16	24.39	0.54	1.73	3.85
	PEGASUS + SummScore LEAD-3 - paraphrasing 100%	15.95	22.84	4.14	20.88	**15.08**	**37.45**	**50.66**
	PEGASUS + SummScore LEAD-3 - paraphrasing 50%	17.77	25.34	5.55	22.43	7.45	18.83	26.22
	PEGASUS + SummScore LEAD-3 - paraphrasing 25%	18.88	26.83	6.40	23.41	3.93	9.75	14.23
	PEGASUS + SummScore LEAD-3 - paraphrasing 12.5% (**pseudo-labels**)	<u>19.33</u>	<u>27.41</u>	<u>6.73</u>	<u>23.84</u>	<u>2.28</u>	<u>5.85</u>	<u>9.29</u>
	PEGASUS self-trained	18.92	26.96	6.41	23.40	0.36	1.51	3.35
	PEGASUS self-trained + SummScore LEAD-3	**20.67**	**28.91**	**7.55**	**25.54**	0.60	2.18	4.93

Table 15: ROUGE and abstractiveness for several models: the unsupervised PEGASUS (first sub-block), re-ranking with SummScore (first sub-block), paraphrasing the resulting pseudo-labels (second sub-block), self-training with the pseudo-labels (third sub-block), then re-ranking self-training outputs with SummScore again (third sub-block). All results are on the test set, results of self-training pseudo-labels are underlined, and highest numbers within 0.1 are in bold.

except on XSum where paraphrasing also improves ROUGE, motivating our choice to use 100% paraphrased summaries as pseudo-labels. We confirm that our pseudo-labels for self-training, made of a blend of SummScore selected summaries and selected summaries being paraphrased, maintains high ROUGE while being much more abstractive than the baseline PEGASUS.

E Paraphrasing Model

For each dataset, we fine-tune BART-large (Lewis et al., 2020) (from the pre-training checkpoint *facebook/bart-large* in HuggingFace transformers (Wolf et al., 2020)) for paraphrasing. The model is trained to paraphrase blocks of $n = 3$ sentences on CNN/DM, $n = 1$ sentence on XSum, and $n = 2$ sentences on WikiHow and SAMSum, in line with average summary lengths on these datasets. We train the model with Adafactor (Shazeer and Stern, 2018) for 5 epochs, with effective batch size 32, learning rate 2e-5, and no weight decay nor label smoothing. We evaluate every 500 optimization steps on CNN/DM, XSum, and WikiHow, and every 100 steps on SAMSum. At inference, we use beam search with beam width 5 and length penalty of 1.0, and block repeated trigrams like in (Kryściński et al., 2018).

Dataset	CNN/DM	XSum	WikiHow	SAMSum
Paraphrasing model	32.88	15.58	20.34	17.44

Table 16: ROUGE results of the paraphrasing model, on the validation set of each dataset. We report the mean of ROUGE-1/2/L.

We track the mean of ROUGE-1, ROUGE-2 and ROUGE-L between the generated paraphrase and target paraphrase on the validation set during training, and perform early stopping. Best mean ROUGE results are shown in Table 16.

Next, we study the impact of the paraphrasing model on the SummScore pseudo-targets. In Ta-

Dataset	Mean R	New 1-grams	New 2-grams	New 3-grams
CNN/DM	55.80	17.28	34.58	39.61
XSum	62.13	20.93	34.60	38.59
WikiHow	81.26	7.96	20.14	25.60
SAMSum	50.64	22.52	41.29	52.02

Table 17: Impact of paraphrasing on the pseudo-targets. We report mean ROUGE and percentage of novel n-grams between the paraphrased pseudo-targets and the original pseudo-targets, on the *training* set of each dataset since this is the subset that paraphrasing is applied to.

ble 17, we compute the mean ROUGE between pseudo-targets and their paraphrase, and analyze the novel n-grams. We point out that the paraphrasing is only applied to the *training* pseudo-labels as the goal of paraphrasing is to encourage the model to learn diversity during self-training, hence Table 17 reporting results on *training sets*. On each dataset, the mean ROUGE is in the 50-80 range, indicating that the paraphrased pseudo-labels do not deviate too much from the original pseudo-labels and yet is able to re-write some content. Besides, there is a high proportion of new n-grams: more than 10% new 1-grams (with the exception of WikiHow on the which the paraphrasing model seems to struggle more to rephrase the input), and more than 20% 2-grams.

F Other Summary Candidates Setups

Decoding method	Candidate Selection	# Candidates 5	10	15	20
Beam search	PEGASUS	26.74	27.00	27.00	26.99
	SummScore	27.46	28.01	28.33	28.38
Diverse beam search	PEGASUS	26.08	26.08	26.07	26.01
	SummScore	26.98	27.48	27.76	27.87
Nucleus sampling	PEGASUS	23.92	23.95	24.04	24.03
	SummScore	26.13	26.57	26.85	27.11
		15	**30**	**45**	**60**
All three methods	SummScore	27.87	28.35	28.34	**28.59**

Table 18: Candidate generation setups. We compare several summary candidates generation setups with PEGASUS on CNN/DM, varying both the decoding method and the number of candidates. We report the mean of ROUGE-1/2/L. Best results within 0.1 are in bold.

In Table 18, we apply SummScore outside of the standard beam search with 20 beams setup. Results show that SummScore performance continuously improves with more summary candidates, whereas the top beam stays around the same level. Besides, SummScore relative gains are stronger with lower quality decoding methods diverse beam search and nucleus sampling. Lastly, combining 20 summary candidates from each of the three decoding methods yields a pool of 60 summary candidates, out of the which SummScore re-ranking can improve by an extra +0.21 mean ROUGE the performance compared to re-ranking 20 beam search candidates (28.59 mean ROUGE vs 28.38). Overall, we recommend our default setup of beam search with 20 beams to apply SummScore re-ranking. A greater number of beams becomes difficult to fit into a standard GPU with 16 GB memory.

G Learned Coefficients

In Table 19 (PEGASUS backbone) and Table 20 (ChatGPT backbone), we show coefficients found by SummScore (for each of the five methods to select pseudo-labels which we studided), and on each dataset, including when applying SummScore again on top the self-trained models. For the sake of conciseness, we do not include SummScore coefficients obtained in zero-shot setups. *BERTScore with source* appears as the feature which consistently receives the highest weight for SummScore - Random-3 and SummScore - LEAD-3 ; while *ROUGE-2 with source* dominates for SummScore - Salient-R1/R2/RL. *Diversity* and *Length* features are significantly less used.

H Re-ranking Examples

In the following, beam search output (for PEGASUS) or the first candidate from top-p sampling (for ChatGPT) is in orange, SummScore selected summary candidate in blue, and oracle candidate(s) in teal. On each dataset, we show one re-ranking example on the unsupervised PEGASUS and/or ChatGPT (Table 3), one zero-shot re-ranking example selected from Table 4, and one re-ranking example applied on top of the self-trained PEGASUS (Table 5).

Dataset	Model	ROUGE-1	ROUGE-2	BLEU	BERTScore	BARTScore	BleuRT	Diversity	Length
CNN/DM	SummScore - Random-3	0.0000	**0.5700**	0.0300	0.2681	0.0000	0.0069	0.1250	0.0000
	SummScore - LEAD-3 (**selected SummScore version**)	0.0000	0.0250	0.0000	**0.4275**	0.3375	0.1350	0.0500	0.0250
	SummScore - Salient-R1	0.0850	**0.7650**	0.0000	0.1000	0.0031	0.0219	0.0000	0.0250
	SummScore - Salient-R2	0.1444	0.1856	**0.4950**	0.1050	0.0000	0.0450	0.0000	0.0250
	SummScore - Salient-RL	0.1062	**0.7438**	0.0000	0.1000	0.0031	0.0219	0.0000	0.0250
	Self-training (1st round) + SummScore - LEAD-3	0.0000	0.0000	0.0000	**0.4500**	0.4275	0.0225	0.1000	0.0000
	Self-training (2nd round) + SummScore - LEAD-3	0.0000	0.0000	0.0000	**0.6338**	0.2925	0.0488	0.0250	0.0000
	Self-training (3rd round) + SummScore - LEAD-3	0.0000	0.0500	0.0000	**0.8075**	0.1425	0.0000	0.0000	0.0000
XSum	SummScore - Random-3	0.0287	**0.5462**	0.0000	0.1200	0.0900	0.1900	0.0250	0.0000
	SummScore - LEAD-3 (**selected SummScore version**)	0.0500	0.0000	0.0000	**0.7837**	0.1425	0.0238	0.0000	0.0000
	SummScore - Salient-R1	0.1275	**0.7225**	0.0000	0.0338	0.0000	0.0413	0.0000	0.0750
	SummScore - Salient-R2	**0.8000**	0.0000	0.0000	0.0000	0.0000	0.2000	0.0000	0.0000
	SummScore - Salient-RL	0.1200	0.1600	**0.5200**	0.1550	0.0000	0.0450	0.0000	0.0000
	Self-training (1st round) + SummScore - LEAD-3	0.0000	0.0000	0.0000	**0.5550**	0.3700	0.0000	0.0250	0.0500
WikiHow	SummScore - Random-3	0.0100	0.0400	0.0000	**0.9025**	0.0238	0.0238	0.0000	0.0000
	SummScore - LEAD-3	0.0000	0.0000	0.0000	**0.7312**	0.2437	0.0000	0.0250	0.0000
	SummScore - Salient-R1	0.1094	**0.7656**	0.0000	0.0825	0.0000	0.0175	0.0250	0.0000
	SummScore - Salient-R2 (**selected SummScore version**)	**0.8750**	0.0000	0.0000	0.0825	0.0000	0.0175	0.0250	0.0000
	SummScore - Salient-RL	0.2625	**0.6125**	0.0000	0.0825	0.0000	0.0175	0.0250	0.0000
	Self-training (1st round) + SummScore - Salient-R2	**0.5031**	0.1750	0.1969	0.0625	0.0050	0.0325	0.0250	0.0000
SAMSum	SummScore - Random-3	0.0300	0.2625	0.0075	**0.4900**	0.2100	0.0000	0.0000	0.0000
	SummScore - LEAD-3 (**selected SummScore version**)	0.0000	0.0000	0.0000	**0.7750**	0.2250	0.0000	0.0000	0.0000
	SummScore - Salient-R1	0.1650	**0.6600**	0.0000	0.0000	0.0000	0.0000	0.1250	0.0500
	SummScore - Salient-R2	0.0731	**0.8044**	0.0975	0.0000	0.0000	0.0000	0.0000	0.0250
	SummScore - Salient-RL	0.1950	**0.7800**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0250
	Self-training (1st round) + SummScore - LEAD-3	0.0000	0.0000	0.0000	**0.8500**	0.1500	0.0000	0.0000	0.0000

Table 19: Coefficients learned by SummScore with PEGASUS for each feature in each dataset and with each pseudo-labels construction technique. Highest feature values for each model are in bold.

Dataset	Model	ROUGE-1	ROUGE-2	BLEU	BERTScore	BARTScore	BleuRT	Diversity	Length
CNN/DM	SummScore - Random-3	0.0600	0.2400	0.0000	**0.3881**	0.1437	0.0431	0.1250	0.0000
	SummScore - LEAD-3 (**selected SummScore version**)	0.0000	0.0975	0.0025	**0.5038**	0.2712	0.0000	0.1250	0.0000
	SummScore - Salient-R1	0.2925	**0.6075**	0.0000	0.0925	0.0025	0.0050	0.0000	0.0000
	SummScore - Salient-R2	**0.3825**	0.3375	0.1800	0.0850	0.0075	0.0075	0.0000	0.0000
	SummScore - Salient-RL	0.2925	**0.6075**	0.0000	0.0825	0.0000	0.0175	0.0000	0.0000
XSum	SummScore - Random-3	0.0581	**0.4844**	0.2325	0.1350	0.0150	0.0500	0.0250	0.0000
	SummScore - LEAD-3 (**selected SummScore version**)	0.0250	0.2250	0.0000	**0.6525**	0.0544	0.0181	0.0250	0.0000
	SummScore - Salient-R1	0.1575	**0.6525**	0.0900	0.0775	0.0025	0.0200	0.0000	0.0000
	SummScore - Salient-R2	0.2700	**0.4950**	0.1350	0.0800	0.0050	0.0150	0.0000	0.0000
	SummScore - Salient-RL	0.3600	**0.5400**	0.0000	0.0750	0.0050	0.0200	0.0000	0.0000
WikiHow	SummScore - Random-3	0.0600	**0.4800**	0.0600	0.3000	0.0281	0.0469	0.0250	0.0000
	SummScore - LEAD-3	0.0000	0.1187	0.0063	**0.7200**	0.0800	0.0000	0.0750	0.0000
	SummScore - Salient-R1	**0.4950**	0.3150	0.0900	0.0850	0.0050	0.0100	0.0000	0.0000
	SummScore - Salient-R2 (**selected SummScore version**)	0.3825	**0.4950**	0.0225	0.0875	0.0050	0.0075	0.0000	0.0000
	SummScore - Salient-RL	**0.4950**	0.3150	0.0900	0.0850	0.0050	0.0100	0.0000	0.0000
SAMSum	SummScore - Random-3 (**selected SummScore version**)	0.0000	0.0000	0.0000	0.0925	0.3006	**0.5319**	0.0000	0.0750
	SummScore - LEAD-3	0.0000	0.0000	0.0000	0.0750	0.3250	**0.6000**	0.0000	0.0000
	SummScore - Salient-R1	0.0000	0.0000	0.0000	0.1500	0.2250	**0.6250**	0.0000	0.0000
	SummScore - Salient-R2	0.0000	0.0000	0.0000	0.0250	0.2250	**0.7500**	0.0000	0.0000
	SummScore - Salient-RL	0.0000	0.0000	0.0000	0.1500	0.2250	**0.6250**	0.0000	0.0000

Table 20: Coefficients learned by SummScore with ChatGPT for each feature in each dataset and with each pseudo-labels construction technique. Highest feature values for each model are in bold.

CNN/DM: re-ranking from the unsupervised PEGASUS		
Source		Royal Dutch Shell Plc said it . has filed a complaint in federal court in Alaska seeking an . order to remove Greenpeace activists who climbed aboard an oil . rig in the Pacific Ocean bound for the Arctic on Monday in a . protest against Arctic drilling. The environmental group said in a statement its team would . occupy the underside of the main deck of the Polar Pioneer, which is under contract to Shell, and plans to unfurl a banner . with the names of millions of people opposed to Arctic drilling. The group said the activists would not interfere with the . vessel's navigation. Scroll down for video . On the rig: Greenpeace activists scale the Polar Pioneer drill rig in the Pacific Ocean . Map: The activists boarded the rig just 750 miles northwest of Hawaii as it makes its journey to the Arctic . At dawn on Monday, the six, from the USA, Germany, New Zealand, Australia, Sweden and Austria, sped towards the Polar Pioneer in inflatable boats launched from the Greenpeace ship Esperanza . Climbers: All Greenpeace activists aboard the rig are experienced climbers and say they don't plan to interfere with the ship's course . 'We're here to highlight that in less than 100 days Shell is . going to the Arctic to drill for oil,' 32-year-old Johno Smith, . one of the six to board the Blue Marlin, the ship carrying the . rig, said in the statement. 'Shell's actions are exploiting the melting ice to increase . a man-made disaster. Climate change is real,' he added. Shell said in an emailed statement that it has met with .groups against oil drilling off Alaska's shores and 'respect . their views' but condemned the boarding. 'We can confirm that protesters from Greenpeace have . illegally boarded the Polar Pioneer, under contract to Shell, jeopardizing not only the safety of the crew on board, but the . protesters themselves,' Shell said. The move comes just days after the U.S. Interior Department . upheld a 2008 lease sale in the Chukchi Sea off Alaska, moving. Shell a step closer to returning to oil and gas exploration in . the Arctic since it suffered mishaps in the region in 2012. The people vs shell: The activists hope they will draw media attention to oil drilling in the Arctic . Reveal a list: Greenpeace activists scale the Polar Pioneer drill rig in the Pacific Ocean to unfurl a banner with the names of millions of people opposed to Arctic drilling . Long haul: The activists used ropes and climbing equipment to scale the 38,000-tonne platform . Many environmentalists oppose offshore energy exploration in . the Arctic, saying that once production begins any oil spill . would be extremely difficult to clean up. Oil industry interests say the Arctic will be important to . the United States' energy security in coming decades when output . from shale formations is expected to wane. Images published by Greenpeace showed the activists using . climbing gear to move from an inflatable boat onto the Blue . Marlin heavy-lift vessel towing the Pioneer, one of two drill . rigs heading to the region, as it cruised some 750 miles (1,207 . km) northwest of Hawaii. The six activists planned to camp on the 38,000-tonne Polar Pioneer platform, which they boarded using inflatable boats from he Greenpeace vessel 'Esperanza.' Tweeting from the rig: Aliyah Field tweeted she'd love some coffee but that the sunrise over the Pacific is gorgeous even from the side of the oil rig . Many names: Aliyah maybe referring to the list of names the activists will hang showing all the people who are opposed to oil drilling in the arctic . The six – from the United States, Germany, New Zealand, Australia, Sweden and Austria – have supplies for several days and can communicate with the outside world, Greenpeace said. 'We made it! We're on Shell's platform. And we're not alone. Everyone can help turn this into a platform for people power!' tweeted one of the six, Aliyah Field. Johno Smith from New Zealand added: 'We're here to highlight that in less than 100 days Shell is going to the Arctic to drill for oil. 'This pristine environment needs protecting for future generations and all life that will call it home. But instead Shell's actions are exploiting the melting ice to increase a man-made disaster.' A Shell spokeswoman, Kelly op de Weegh, blasted the action. 'We can confirm that protestors from Greenpeace have illegally boarded the 'Polar Pioneer,'under contract to Shell, jeopardizing not only the safety of the crew on board, but the protestors themselves,' she said on Monday. She added: 'Shell has met with organizations and individuals who oppose energy exploration offshore Alaska. We respect their views and value the dialogue. 'We will not, however, condone the illegal tactics employed by Greenpeace. Nor will we allow these stunts to distract from preparations underway to execute a safe and responsible exploration program,' she said in a statement.
Beam #1	Summary	'We're here to highlight that in less than 100 days Shell is.
	Scores	Mean ROUGE: **6.55** (rank 11) || SummScore rank: **20**
Beam #2	Summary	'We're here to highlight that in less than 100 days Shell is.' Greenpeace activists boarded the Polar Pioneer drill rig in the Pacific Ocean to unfurl a banner with the names of millions of people opposed to oil drilling in the Arctic. Greenpeace activists climbed aboard an oil drilling rig off the coast of Alaska on Monday in a protest against Shell's plans to drill for oil in Arctic waters, the environmental group said in a statement. The group said the activists would not interfere with the vessel's navigation, but the company said it had filed a complaint in federal court in Alaska seeking an order to remove the activists.
	Scores	Mean ROUGE: 43.17 (rank 3) || SummScore rank: 2
Beam #3	Summary	'We're here to highlight that in less than 100 days Shell is.' Greenpeace activists boarded the Polar Pioneer drill rig in the Pacific Ocean to unfurl a banner with the names of millions of people opposed to oil drilling in the Arctic. Greenpeace activists climbed aboard an oil drilling rig off the coast of Alaska on Monday in a protest against Shell's plans to drill for oil in Arctic waters, the environmental group said in a statement. The group said the activists would not interfere with the ship's navigation, but the company said it had filed a complaint in federal court in Alaska seeking an order to remove the activists.
	Scores	Mean ROUGE: **42.85** (rank 4) || SummScore rank: **1** (**SummScore output**)
Beam #4	Summary	'We're here to highlight that in less than 100 days Shell is.' Greenpeace activists boarded the Polar Pioneer drill rig in the Pacific Ocean to unfurl a banner with the names of millions of people opposed to oil drilling in the Arctic. Greenpeace activists climbed aboard an oil drilling rig off the coast of Alaska on Monday in a protest against Shell's plans to drill for oil in Arctic waters, the environmental group said in a statement. The group said the activists would not interfere with the vessel's navigation, but the company said it had filed a complaint in federal court in Alaska seeking an order to remove the activists from
	Scores	Mean ROUGE: 43.59 (rank 2) || SummScore rank: 12
...		
Beam #10	Summary	'We're here to highlight that in less than 100 days Shell is.' Greenpeace activists boarded the Polar Pioneer drill rig in the Pacific Ocean to unfurl a banner with the names of millions of people opposed to oil drilling in the Arctic. Greenpeace activists climbed aboard an oil drilling rig off the coast of Alaska on Monday in a protest against Shell's plans to drill for oil in Arctic waters, the environmental group said in a statement. The group said the activists would not interfere with the vessel's navigation, but the company said it had filed a complaint in federal court in Alaska seeking an order to remove them from the
	Scores	Mean ROUGE: **43.91** (rank 1) || SummScore rank: **6**
...		
Reference		Shell has filed a complaint in federal court in Alaska seeking an order to remove Greenpeace activists who climbed aboard an oil rig in the Pacific . The environmental group said in a statement its team would occupy the underside of the main deck of the Polar Pioneer . The six activists are camping on the 38,000-tonne Polar Pioneer platform, which they boarded using inflatable boats from the Greenpeace vessel 'Esperanza' 'We made it! We're on Shell's platform. And we're not alone. Everyone can help turn this into a platform for people power!' tweeted Aliyah Field .

Table 21: SummScore re-ranking applied to the unsupervised PEGASUS with beam search on CNN/DM.

CNN/DM: re-ranking from ChatGPT		
Source		Although Hillary Clinton boasts a robust 3.6 million Twitter followers, not even a vast right-wing conspiracy would be able to interact with 2 million of them. According to two popular online measuring tools, no more than 44 per cent of her Twitter fan base consists of real people who are active in using the social media platform. And at least 15 per cent – more than 544,000 – are completely fake. StatusPeople.com, the oldest publicly available Twitter-auditing tool, reports that 44 per cent of the former secretary of state's followers are 'good'; 15 per cent are 'fake'; and 41 per cent are 'inactive,' meaning that they never tweet or reply to any tweets. SCROLL DOWN FOR VIDEO . FAKERS: According to one popular online audit tool, only 44 per cent of Hillary Clinton's Twitter followers are real people who participate on the social media platform . 'I'M RUNNING FOR PRESIDENT': Clinton has cast herself as a champion of 'everyday Americans' Another Twitter sleuthing website sampled more than 320,000 of Clinton's followers and found that a much larger number of them were 'fake' Just 4 per cent of President Barack Obama's Twitter followers, by comparison, are considered fake. The White House worked overtime to purge most of them after a September 2013 report found that more than half of his followers didn't really exist. Michelle Obama's Twitter audience is 25 per cent fake, according to StatusPeople, along with 21 per cent of Vice President Joe Biden's. Another tool, TwitterAudit.com, sampled 320,000 of Mrs. Clinton's followers and found that 18 per cent were fake. The new measurements will add to the Clinton presidential campaign's embarrassment following news on Tuesday that a large number of her Facebook fans may represent 'likes' that were purchased rather than earned. REALLY? Hillary Clinton's Twitter follower-count appears to be significantly inflated . FACEBOOK FAKERY: Clinton boasts two-thirds of a million Facebook 'likes,' but more than 46,000 of them list 'Baghdad' as their hometown . Vocativ reported that at least 7 per cent of them listed Baghdad, Iraq as their hometown, a larger number than any U.S. city. That would represent more than 46,000 people. Additional evidence of digital astroturfing is that while most of her U.S. Facebook fans are older than 55, most of the Baghdad contingent is in the 18-34 age range. While Clinton was America's top diplomat, her State Department was buying Facebook 'likes,'according to an Associated Press report from last year. 'In 2013, the State Department, which has more than 400,000 likes and was recently most popular in Cairo, said it would stop buying Facebook fans after its inspector general criticized the agency for spending $630,000 to boost the numbers,' the wire service reported. Clinton left the agency while the inspector general was investigating. ASTROTURF? A woman cast as a kindly, gardening grandma in Clinton's campaign launch video is actually a former abortion lobbyist who once helmed a campaign for Texas state Senator Wendy Davis . WHERE'S SCOOBY? Clinton's caravan was spotted in Maumee, Ohio on Monday when she lunched at Chipotle, but no one recognized her . Facebook says it has entire teams of employees tasked with rooting out schemes to artificially boost statistics. 'When we catch fraudulent activity, we work to counter and prevent it – including blocking accounts and removing fake likes,' the company reports on its website. Clinton is in Monticello, Iowa on Tuesday for the first stop on her nascent presidential 'listening' tour. She was driven from New York in a van that her campaign aides have dubbed 'Scooby,' after the 'Mystery Machine' van in the Scooby-Doo cartoon series. Daily Mail Online visited Kirkwood Community College on Monday, the site of the event, and heard from students that the former first lady is a 'control freak' who may be pursuing legal status for illegal immigrants because the Democratic Party needs a new pool of loyal voters.
Candidate #1	Summary	Online measuring tools show that only 44% of Hillary Clinton's Twitter followers are real people who actively use the platform, with 15% completely fake. Comparatively, just 4% of President Obama's Twitter followers are fake. Clinton's Facebook audience has also come under scrutiny, with reports of purchased likes and a large number of followers from Baghdad, Iraq.
	Scores	Mean ROUGE: **23.88** (rank 9) ‖ SummScore rank: **18**
Candidate #2	Summary	At least 15% of Hillary Clinton's Twitter followers are fake, according to two online measuring tools, with more than 544,000 completely fake. In comparison, just 4% of Barack Obama's Twitter followers are fake. Clinton's recent embarrassment of having a large number of Facebook likes purchased rather than earned further adds to the digital astroturfing speculation.
	Scores	Mean ROUGE: 19.60 (rank 26) ‖ SummScore rank: 15
Candidate #3	Summary	An investigation has revealed that at least 15% of Hillary Clinton's Twitter followers are completely fake. The same investigation also found that only 44% of her Twitter followers are real people who actively use the platform. This adds to the Clinton presidential campaign's embarrassment following news on Tuesday that a large number of her Facebook fans may represent 'likes' that were purchased rather than earned.
	Scores	Mean ROUGE: 23.88 (rank 8) ‖ SummScore rank: 5
Candidate #4	Summary	According to two online measuring tools, only 44% of Hillary Clinton's 3.6 million Twitter followers are real people who actively use the platform, with at least 15% being fake. In comparison, only 4% of President Barack Obama's Twitter followers are considered fake. Clinton's Facebook page has also been called into question, with evidence of 'likes' being purchased rather than earned.
	Scores	Mean ROUGE: 27.09 (rank 3) ‖ SummScore rank: 10
Candidate #5	Summary	According to online measuring tools, over half of Hillary Clinton's Twitter followers are either fake or inactive. One tool reported that only 44% of her followers are real people who are active on the platform. Additionally, a large number of Clinton's Facebook fans may have been purchased rather than earned, with over 46,000 of them listing Baghdad, Iraq as their hometown.
	Scores	Mean ROUGE: 24.82 (rank 7) ‖ SummScore rank: 7
...		
Candidate #17	Summary	According to online measuring tools, only 44% of Hillary Clinton's 3.6 million Twitter followers consist of real people who are active on the social media platform. At least 15% of her followers, or more than 544,000, are completely fake. This comes after news that a large number of her Facebook fans may represent 'likes' that were purchased rather than earned.
	Scores	Mean ROUGE: **31.34** (rank 1) ‖ SummScore rank: **1** (**SummScore output**)
...		
Reference		Two different online audit tools say no more than 44 per cent of Hillary's 3.6 million Twitter fans are real people who participate in the platform . The newly minted presidential candidate is fending off accusations that her Facebook page is full of fake 'likes' Her Facebook fan base includes more people from Baghdad, Iraq than any US city . When she was secretary of state, her agency paid $630,000 to bulk up its Facebook likes, but pledged to stop after she left .

Table 22: SummScore re-ranking applied to ChatGPT with top-p sampling on CNN/DM.

CNN/DM: re-ranking from the PEGASUS trained on WikiHow		
Source		Assault: Dr Sahar Hussain attacked two Tube workers because she didn't want to miss the last train home . A GP attacked two Tube workers while screaming 'I'm a doctor' because she did not want to miss the last train home on a Friday night. Dr Sahar Hussain, 53, panicked when she was unable to get through the gates at Leicester Square station, and started ranting at staff. She denied assaulting the two workers, saying she was worried about being stranded on her own in central London because she is a Muslim woman. But Hussain has now been found guilty and ordered to pay a total of £2,250 in fines, compensation and court costs - and she could face disciplinary action from the General Medical Council. In video footage captured on her own mobile phone, Hussain could be heard to shout: 'I'm a doctor actually, I work for the NHS. I'm a doctor. Get me through the gate, I'm going to miss my train.' City of London Magistrates' Court heard Hussain arrived at the station around 11.30pm on June 20 last year, trying to get home to Woodford Green after socialising with friends in the West End. When she was refused entry by the automatic gates, she demanded that ticket seller Malcolm Shaw let her through before lashing out at his colleague Indira Ramsaroop, who was trying to help. Hussain, originally from Iraq, screamed and shouted at Mrs Ramsaroop as she thrust a camera phone into her face before grabbing her by the arm. The 24-year-old Transport for London worker was then chased by the doctor as she tried to flee to the control room, bumping her head on the way. In the video on Hussain's phone she was heard shouting: 'This woman is on something, she's not sober is she? You're in work and you're not sober. Get me through the gate.' During the scuffle Hussain, a mother of one who helps train GPs at two universities, also grabbed Mr Shaw by the arms, leaving him with scratches. Mrs Ramsaroop was close to tears in court as she told how she had to take almost two weeks off work following the incident, adding: 'I had a lot of sleepless nights. It had an impact on myself with customers when I came back to work. 'I have felt very let down to have been threatened and been running away in my place of work. It actually affected me for a very long time and I got quite ill just at the worrying and fear.' Row: The assault took place on a Friday night at Leicester Square station in central London . Hussain admitted losing her temper, telling the court: 'I'm very sorry about the way I expressed myself with my agitation and frustration.' District Judge Quentin Purdy found her guilty of two counts of assault by beating, saying: 'The evidence is overwhelming. You completely lost your self-control. 'Unusually for the sort of incident this court regularly deals with, there is no hint of intoxication in any way. 'But the explanation is there. You were anxious about missing the last train, concerned about being late for work or unable to work the next day, cultural embarrassment by being left in London as a married Muslim woman. 'It was completely out of character, but as we have all seen, you acted in an unpleasant and hysterical fashion towards numerous members of staff.' He said she had subjected Mrs Ramsaroop to 'completely unjustified violence', as well as 'humiliation' for falsely suggesting she was intoxicated at work. The judge added: 'You have lost a great deal by your lack of self-control. Your good name is now lost with these two convictions for assault.' Hussain was fined £500 for each assault and ordered her to pay £500 compensation to Mrs Ramsaroop, £200 to Mr Shaw, £500 prosecution costs and a £50 victim surcharge.
Beam #1	Summary	Try to get through the gate because you didn't want to miss the last train home. Attack the workers because you didn't want to miss the last train home.
	Scores	Mean ROUGE: **13.80** (rank 6) ‖ SummScore rank: **17**
Beam #2	Summary	Tell the court she was worried about being stranded on her own in central London because she is a Muslim woman. Describe to the court how she screamed: 'I'm a doctor, get me through the gate.'
	Scores	Mean ROUGE: 40.97 (rank 2) ‖ SummScore rank: 2
Beam #3	Summary	Tell the court she was worried about being stranded on her own in central London because she is a Muslim woman. Describe to the court how she screamed: 'Get me through the gate, get me through the gate.'
	Scores	Mean ROUGE: 39.29 (rank 4) ‖ SummScore rank: 10
Beam #4	Summary	Try to get through the gate because you don't want to miss the last train home. Attack the workers because you didn't want to miss the last train home.
	Scores	Mean ROUGE: 13.80 (rank 6) ‖ SummScore rank: 18
Beam #5	Summary	Try to get through a gate because you don't want to miss the last train home. Attack the workers because you don't want to miss the last train home.
	Scores	Mean ROUGE: 13.09 (rank 7) ‖ SummScore rank: 20
Beam #6	Summary	Try to get through the gate because you don't want to miss the last train home. Attack the workers because you don't want to miss the last train home.
	Scores	Mean ROUGE: 13.80 (rank 6) ‖ SummScore rank: 19
Beam #7	Summary	Tell the court she was worried about being stranded on her own in central London because she is a Muslim woman. Describe to the court how she yelled: 'Get me through the gate, get me through the gate.'
	Scores	Mean ROUGE: 39.29 (rank 3) ‖ SummScore rank: 8
...		
Beam #10	Summary	Tell the court she was worried about being stranded on her own in central London because she is a Muslim woman. Describe to the court how she screamed: 'I'm a doctor, get me through the gate!'
	Scores	Mean ROUGE: **40.97** (rank 2) ‖ SummScore rank: **1** (**SummScore output**)
...		
Beam #14	Summary	Tell the court she was worried about being stranded on her own in central London because she is a Muslim woman. Describe to the court how she screamed: 'I'm a doctor, get through the gate.'
	Scores	Mean ROUGE: **42.04** (rank 1) ‖ SummScore rank: **9**
...		
Reference		Shell has filed a complaint in federal court in Alaska seeking an order to remove Greenpeace activists who climbed aboard an oil rig in the Pacific . The environmental group said in a statement its team would occupy the underside of the main deck of the Polar Pioneer . The six activists are camping on the 38,000-tonne Polar Pioneer platform, which they boarded using inflatable boats from the Greenpeace vessel 'Esperanza' 'We made it! We're on Shell's platform. And we're not alone. Everyone can help turn this into a platform for people power!' tweeted Aliyah Field .

Table 23: SummScore re-ranking applied to the PEGASUS fine-tuned on WikiHow with beam search on CNN/DM.

CNN/DM: re-ranking from the self-trained PEGASUS		
Source		Grandparents have pleaded for the safe return to Australia of two young children whose mother took them from Melbourne to the Islamic State capital in Syria. Former Melbourne woman Dullel Kassab fled to Raqqa in Syria with her children last year, and she regularly boasts on Twitter that her four-year-old daughter and two-year-old son sleep with toy guns next to their beds and her daughter likes watching IS videos of 'Muslims killing bad ppl.' The children's paternal grandparents say they are worried Kassab, 28, is 'brainwashing' the children, after their father was killed near the Syria-Turkey border last year, The Herald Sun reported. Former Melbourne woman Dullel Kassab fled to Raqqa in Syria from Melbourne with her children last year . Kassab posts pictures to Twitter of airstrikes hitting blocks away from their Raqqa apartment . 'We miss the children a lot. Their safety and religion has been compromised and we are deeply worried but unable to do anything about it,' a family spokesman told the Herald Sun. 'We pray they come back but it does not look good.' Kassab's Twitter paints a picture of their life in the city the terrorist group IS have made their headquarters, where the children cannot go to school and airstrikes hit blocks away from their apartment. The 28-year-old has a new husband, as the Islamic State does not permit unmarried foreign women to stay in Raqqa. In social media posts she boasts about her children's distaste for Kuffar (non-believers). A photo of another airstrike a day later. The children's paternal grandparents say they are worried Kassab, 28, is 'brainwashing' the children, after their father was killed near the Syria-Turkey border last year . On her Twitter account she boasts about her children's distaste for Kuffar (non-believers) 'My 4y/o encouraging her little bro to eat his eggs – "C'mon eat ur eggs so u can be big & strong & fight the Kuffar!" Allah yehmikum! [sic]' she wrote in December. '#Awkward Just asked my 4yo wat she wants 2 watch.. "Muslims killing bad ppl" (i.e. #IS vids obv not beheading ones) LOL [sic],' she wrote in October. Kassab has also complained the 12 to 17-year-olds are now regarded as children when 'in the past they were warriors'. And during the Sydney Lindt café siege in December last year she sent a series of tweets joking that it was exciting. 'This is the most excitement Sydney has seen since the 2000 Olympics!' she posted. Kassab also posts pictures of the Islamic State capital - including this of a 'double rainbow' And during the Sydney Lindt café siege last year Kassab sent a series of tweets joking that it was 'exciting' 'I guess attack the coffee shop wasn't a bad idea, It's a long night… One needs caffeine and chocolate!! [sic]' Kassab also posts pictures of the Islamic State capital, and of Nutella and Twix and Snickers chocolate bars with the caption: 'Im really appreciating #globalization right about now! #SimplePleasures Another reason to love #IS [sic].' The 28-year-old's father Jalal Kassab said he was worried about his grandchildren living in a war zone, but said the threat of imprisonment made it difficult for his daughter to return to Australia. 'I know she wants to come back and we are trying everything we can to bring her back,' Mr Kassab told the Herald Sun. Another former Melbourne woman Zehra Duman last month shared a series of propaganda pictures she says shows her 'five star jihad' lifestyle . In photographs posted to a Twitter several women are pictured standing under an Islamic State flag, reclining against a clean white BMW M5 and wielding machine guns . In one tweet, Duman said: 'US + Australia, how does it feel that all 5 of us were born n raised in your lands, & now here thirsty for ur blood?'
Beam #1	Summary	The children's paternal grandparents say they are worried Kassab, 28, is 'brainwashing' the children, after their father was killed near the Syria-Turkey border last year. 'We pray they come back but it does not look good.' Kassab's Twitter paints a picture of their life in the city the terrorist group IS have made their headquarters, where the children cannot go to school and airstrikes hit blocks away from their apartment.
	Scores	Mean ROUGE: **14.89** (rank 4) ‖ SummScore rank: **6**
Beam #2	Summary	'We pray they come back but it does not look good.' Kassab's Twitter paints a picture of their life in the city the terrorist group IS have made their headquarters, where the children cannot go to school and airstrikes hit blocks away from their apartment. The children's paternal grandparents say they are worried Kassab, 28, is 'brainwashing' the children, after their father was killed near the Syria-Turkey border last year.
	Scores	Mean ROUGE: 14.89 (rank 4) ‖ SummScore rank: 11
Beam #3	Summary	The children's paternal grandparents say they are worried Kassab, 28, is 'brainwashing' the children, after their father was killed near the Syria-Turkey border last year, The Herald Sun reported. 'We pray they come back but it does not look good.' Kassab's Twitter paints a picture of their life in the city the terrorist group IS have made their headquarters, where the children cannot go to school and airstrikes hit blocks away from their apartment.
	Scores	Mean ROUGE: 14.41 (rank 6) ‖ SummScore rank: 5
Beam #4	Summary	'We pray they come back but it does not look good.' Kassab's Twitter paints a picture of their life in the city the terrorist group IS have made their headquarters, where the children cannot go to school and airstrikes hit blocks away from their apartment. 'My 4y/o encouraging her little bro to eat his eggs ˘ 'C'mon eat ur eggs so u can be big & strong & fight the Kuffar!' Allah yehmikum!
	Scores	Mean ROUGE: 9.92 (rank 10) ‖ SummScore rank: 13
...		
Beam #9	Summary	Former Melbourne woman Dullel Kassab fled to Raqqa in Syria with her children last year, and she regularly boasts on Twitter that her four-year-old daughter and two-year-old son sleep with toy guns next to their beds and her daughter likes watching IS videos of 'Muslims killing bad ppl.' The children's paternal grandparents say they are worried Kassab, 28, is 'brainwashing' the children, after their father was killed near the Syria-Turkey border last year.
	Scores	Mean ROUGE: **57.48** (rank 1) ‖ SummScore rank: **1** (**SummScore output**)
...		
Reference		Grandparents have pleaded for the safe return of two children in Syria . Former Melbourne woman Dullel Kassab fled to Raqqa in Syria with her four-year-old daughter and two-year-old son last year . She said her daughter likes watching IS videos of 'Muslims killing bad ppl'

Table 24: Self-trained PEGASUS with beam search on CNN/DM.

XSum: re-ranking from the unsupervised PEGASUS		
Source		Acting Taoiseach Enda Kenny of Fine Gael and Micheál Martin of Fianna Fáil hope to avoid a second election. Mr Martin has said his party will facilitate a minority government, but will not support a programme for government. Water charges are one of the main sticking points to reaching agreement. A commission to consider the future of national water utility Irish Water is one of the proposals being considered. Fianna Fáil want to see the immediate removal of water charges, but Fine Gael see a role for them. Following the election, almost two months ago, Fine Gael had 50 seats, Fianna Fáil 44, Sinn Féin 23 and the Labour Party got seven. But no party was able to form a majority government and TDs have so far failed to elect a taoiseach.
Beam #1	Summary	Fianna Fil want to see the immediate removal of water charges, but Fine Gael see a role for them.
	Scores	Mean ROUGE: **8.77** (rank 5) ‖ SummScore rank: **14**
Beam #2	Summary	Following the election, almost two months ago, Fine Gael had 50 seats, Fianna Fil 44, Sinn Féin 23 and the Labour Party got seven.
	Scores	Mean ROUGE: 6.06 (rank 9) ‖ SummScore rank: 6
Beam #3	Summary	Acting Taoiseach Enda Kenny of Fine Gael and Michel Martin of Fianna Fil hope to avoid a second election.
	Scores	Mean ROUGE: 7.02 (rank 7) ‖ SummScore rank: 15
Beam #4	Summary	After the election, almost two months ago, Fine Gael had 50 seats, Fianna Fil 44, Sinn Féin 23 and the Labour Party got seven.
	Scores	Mean ROUGE: 6.06 (rank 9) ‖ SummScore rank: 7
Beam #5	Summary	The election, almost two months ago, Fine Gael had 50 seats, Fianna Fil 44, Sinn Féin 23 and the Labour Party got seven.
	Scores	Mean ROUGE: 6.20 (rank 8) ‖ SummScore rank: 12
Beam #6	Summary	A commission to consider the future of national water utility Irish Water is one of the proposals being considered. Fianna Fil want to see the immediate removal of water charges, but Fine Gael see a role for them.
	Scores	Mean ROUGE: 10.53 (rank 4) ‖ SummScore rank: 4
Beam #7	Summary	Mr Martin has said his party will facilitate a minority government, but will not support a programme for government. Fianna Fil want to see the immediate removal of water charges, but Fine Gael see a role for them.
	Scores	Mean ROUGE: 17.63 (rank 3) ‖ SummScore rank: 2
Beam #8	Summary	following the election, almost two months ago, Fine Gael had 50 seats, Fianna Fil 44, Sinn Féin 23 and the Labour Party got seven.
	Scores	Mean ROUGE: 6.06 (rank 9) ‖ SummScore rank: 8
Beam #9	Summary	Follow the election, almost two months ago, Fine Gael had 50 seats, Fianna Fil 44, Sinn Féin 23 and the Labour Party got seven.
	Scores	Mean ROUGE: 6.06 (rank 9) ‖ SummScore rank: 13
Beam #10	Summary	During the election, almost two months ago, Fine Gael had 50 seats, Fianna Fil 44, Sinn Féin 23 and the Labour Party got seven.
	Scores	Mean ROUGE: 6.06 (rank 9) ‖ SummScore rank: 9
Beam #11	Summary	acting Taoiseach Enda Kenny of Fine Gael and Michel Martin of Fianna Fil hope to avoid a second election.
	Scores	Mean ROUGE: 7.02 (rank 7) ‖ SummScore rank: 20
Beam #12	Summary	Fianna Fil wants to see the immediate removal of water charges, but Fine Gael see a role for them.
	Scores	Mean ROUGE: 8.77 (rank 5) ‖ SummScore rank: 16
Beam #13	Summary	Mr Martin has said his party will facilitate a minority government, but will not support a programme for government. Fianna Fil want to see the immediate removal of water charges, but Fine Gael see a role for them. However, no party was able to form a majority government and TDs have so far failed
	Scores	Mean ROUGE: **19.28** (rank 2) ‖ SummScore rank: **1** (**SummScore output**)
Beam #14	Summary	While Fianna Fil want to see the immediate removal of water charges, but Fine Gael see a role for them.
	Scores	Mean ROUGE: 8.55 (rank 6) ‖ SummScore rank: 19
Beam #15	Summary	Fianna Fil wanted to see the immediate removal of water charges, but Fine Gael see a role for them.
	Scores	Mean ROUGE: 8.77 (rank 5) ‖ SummScore rank: 17
Beam #16	Summary	Mr Martin has said his party will facilitate a minority government, but will not support a programme for government.
	Scores	Mean ROUGE: **21.25** (rank 1) ‖ SummScore rank: **10**
Beam #17	Summary	Fianna Fil hope to see the immediate removal of water charges, but Fine Gael see a role for them.
	Scores	Mean ROUGE: 8.77 (rank 5) ‖ SummScore rank: 18
Beam #18	Summary	Mr Martin has said his party will facilitate a minority government, but will not support a programme for government. Fianna Fil want to see the immediate removal of water charges, but Fine Gael see a role for them. However, no party was able to form a majority government and TDs so far failed to
	Scores	Mean ROUGE: 19.28 (rank 2) ‖ SummScore rank: 3
Beam #19	Summary	Following the election, almost two months ago, Fine Gael had 50 seats, Fianna Fil 44, Sinn Féin 23 and the Labour Party got seven..
	Scores	Mean ROUGE: 6.06 (rank xx) ‖ SummScore rank: 11
Beam #20	Summary	Mr Martin has said his party will facilitate a minority government, but will not support a programme for government. Fianna Fil want to see the immediate removal of water charges, but Fine Gael see a role for them. However, no party was able to form a majority government and TDs will so far failed
	Scores	Mean ROUGE: 19.28 (rank 2) ‖ SummScore rank: 5
Reference		Negotiators from Ireland's two biggest parties are to meet for a final effort to form a minority government.

Table 25: SummScore re-ranking applied to the unsupervised PEGASUS with beam search on XSum.

XSum: re-ranking from the BART trained on WikiHow		
Source		Cardiff narrowly missed out on the Championship play-offs last term, finishing eighth in the table. But under new head coach Paul Trollope, Pilkington thinks the Welsh side are capable of better. "We fell just short [last season] and we were all really disappointed with that. We felt we had the squad to get in the play-offs definitely," he said. "It's going to be a tough season again, just like it was last year, but hopefully we can be right up there come the end of the season." Pilkington, 28, scored Cardiff's first goal in their 4-0 friendly win at Shrewsbury on Saturday. The Republic of Ireland international played in attack, having moved from the wing to a central forward role last season. Pilkington made the transition under former Cardiff boss Russell Slade, who was replaced by Trollope in May. Trollope had been working under Slade before then, and Pilkington believes that smooth transition has helped the Bluebirds in their preparations for the new campaign. "The gaffer was working with us last season so we know how he works," he added. "We know what he's about and we know we have to be on our toes all the time. "He's come in and changed a few things. It's been good up to now so hopefully we can carry on working really well."
Beam #1	Summary	Aim to improve on last season's finish., Expect a tough season.
	Scores	Mean ROUGE: **5.13** (rank 11) || SummScore rank: **12**
Beam #2	Summary	Aim to improve on last season's finish.,Pilkington believes the Bluebirds are capable of reaching the play-offs.
	Scores	Mean ROUGE: 24.63 (rank 2) || SummScore rank: 5
Beam #3	Summary	Aim to improve on last season's finish., Believe the Bluebirds have the squad to get back up the table.
	Scores	Mean ROUGE: 17.89 (rank 7) || SummScore rank: 6
Beam #4	Summary	Reach the play-offs again.
	Scores	Mean ROUGE: 7.02 (rank 10) || SummScore rank: 18
Beam #5	Summary	Aim to improve on last season's finish.,Pilkington believes the Bluebirds have the squad to get back up the table.
	Scores	Mean ROUGE: **23.20** (rank 4) || SummScore rank: **1** (**SummScore output**)
Beam #6	Summary	Aim to improve on last season's finish.,Pilkington believes the Bluebirds have the squad to reach the play-offs.
	Scores	Mean ROUGE: 23.90 (rank 3) || SummScore rank: 2
Beam #7	Summary	Expect to improve on last season's finish.,Pilkington believes the Bluebirds have the squad to get back up the table.
	Scores	Mean ROUGE: 23.20 (rank 4) || SummScore rank: 3
Beam #8	Summary	Aim to improve on last season's finish.,Pilkington believes the Bluebirds have the squad to challenge for promotion.
	Scores	Mean ROUGE: **41.06** (rank 1) || SummScore rank: **7**
Beam #9	Summary	Aim to improve on last season's finish.,Pilkington believes the Bluebirds are capable of reaching the play-offs again.
	Scores	Mean ROUGE: 23.90 (rank 3) || SummScore rank: 4
Beam #10	Summary	Aim to improve on last season's finish., Believe in the squad.
	Scores	Mean ROUGE: 12.82 (rank 9) || SummScore rank: 9
Beam #11	Summary	Aim to improve on last season's finish., Expect a tough season again.
	Scores	Mean ROUGE: 4.94 (rank 12) || SummScore rank: 8
Beam #12	Summary	Aim to improve on last season's finish., Believe in the squad.
	Scores	Mean ROUGE: 12.82 (rank 9) || SummScore rank: 11
Beam #13	Summary	Aim to improve on last season's finish., Expect to challenge for promotion again.
	Scores	Mean ROUGE: 21.79 (rank 6) || SummScore rank: 16
Beam #14	Summary	Prepare for a tough season.
	Scores	Mean ROUGE: 14.04 (rank 8) || SummScore rank: 13
Beam #15	Summary	Aim to improve on last season's finish., Believe in the squad.
	Scores	Mean ROUGE: 12.82 (rank 9) || SummScore rank: 10
Beam #16	Summary	Prepare for a tough season.
	Scores	Mean ROUGE: 14.04 (rank 8) || SummScore rank: 14
Beam #17	Summary	Aim to improve on last season's finish., Expect to challenge for promotion.
	Scores	Mean ROUGE: 22.62 (rank 5) || SummScore rank: 17
Beam #18	Summary	Prepare for a tough season.
	Scores	Mean ROUGE: 14.04 (rank 8) || SummScore rank: 15
Beam #19	Summary	Prepare for another tough season.
	Scores	Mean ROUGE: 14.04 (rank 8) || SummScore rank: 20
Beam #20	Summary	Prepare for another tough season.
	Scores	Mean ROUGE: 14.04 (rank 8) || SummScore rank: 19
Reference		Cardiff City forward Anthony Pilkington believes the Bluebirds can challenge for promotion this season.

Table 26: SummScore re-ranking applied to the BART fine-tuned on WikiHow with beam search on XSum.

XSum: re-ranking from the self-trained PEGASUS		
Source		In a statement seen by the BBC, his wife said he had been out drinking with another man in the city of Yangon on Friday night before his death. The Foreign Office said it was investigating the death of a British national and was providing support to the victim's family. News agency AFP quoted police as saying the dead man was a 47-year-old who had suffered wounds to his head and chest. According to reports, the man - who had been working at an international school - was found dead at an apartment in Yangon, previously known as Rangoon. The Foreign Office says it is working with local authorities in Myanmar, formerly Burma, to determine what happened.
Beam #1	Summary	The Foreign Office said it was investigating the death of a British national and was providing support to the victim's family.
	Scores	Mean ROUGE: **13.07** (rank 10) ‖ SummScore rank: **3**
Beam #2	Summary	In a statement seen by the BBC, his wife said he had been out drinking with another man in the city of Yangon on Friday night before his death.
	Scores	Mean ROUGE: 10.85 (rank 12) ‖ SummScore rank: 4
Beam #3	Summary	According to reports, the man - who had been working at an international school - was found dead at an apartment in Yangon, previously known as Rangoon.
	Scores	Mean ROUGE: 20.61 (rank 8) ‖ SummScore rank: 7
Beam #4	Summary	The man, who has not been named, was found dead at an apartment in Yangon, Myanmar, on Saturday.
	Scores	Mean ROUGE: **31.39** (rank 1) ‖ SummScore rank: **14**
Beam #5	Summary	The man, who has not been named, was found dead at an apartment in Yangon, formerly known as Rangoon, on Saturday.
	Scores	Mean ROUGE: 24.88 (rank 6) ‖ SummScore rank: 12
Beam #6	Summary	According to reports, the man - who had been working at an international school - was found dead at an apartment in Yangon, formerly known as Rangoon.
	Scores	Mean ROUGE: 20.61 (rank 8) ‖ SummScore rank: 5
Beam #7	Summary	The man, who has not been named, was found dead at an apartment in Yangon, previously known as Rangoon.
	Scores	Mean ROUGE: **26.39** (rank 4) ‖ SummScore rank: **1** (**SummScore output**)
Beam #8	Summary	The man, who has not been named, was found dead at an apartment in Yangon, formerly known as Rangoon.
	Scores	Mean ROUGE: 26.39 (rank 4) ‖ SummScore rank: 2
Beam #9	Summary	The man, who has not been named, was found dead at an apartment in Yangon, previously known as Rangoon, on Saturday.
	Scores	Mean ROUGE: 24.88 (rank 6) ‖ SummScore rank: 11
Beam #10	Summary	The Foreign Office said it was working with local authorities in Myanmar, formerly Burma, to determine what happened.
	Scores	Mean ROUGE: 12.64 (rank 11) ‖ SummScore rank: 10
Beam #11	Summary	The Foreign Office says it is working with local authorities in Myanmar, formerly Burma, to determine what happened.
	Scores	Mean ROUGE: 12.64 (rank 11) ‖ SummScore rank: 11
Beam #12	Summary	The man, who has not been named, was found dead at an apartment in Yangon, formerly Burma, on Saturday.
	Scores	Mean ROUGE: 26.39 (rank 4) ‖ SummScore rank: 18
Beam #13	Summary	Media playback is unsupported on your device 11 August 2015 Last updated at 08:00 BST The Foreign Office said it was investigating the death of a British national in the city of Yangon.
	Scores	Mean ROUGE: 9.78 (rank 13) ‖ SummScore rank: 19
Beam #14	Summary	Media playback is unsupported on your device 11 August 2015 Last updated at 08:00 BST The man, who has not been named, was found dead at an apartment in Yangon.
	Scores	Mean ROUGE: 19.33 (rank 9) ‖ SummScore rank: 20
Beam #15	Summary	The man, who has not been named, was found dead at an apartment in Yangon, the capital of Myanmar, on Saturday.
	Scores	Mean ROUGE: 28.69 (rank 2) ‖ SummScore rank: 16
Beam #16	Summary	According to reports, the man - who had been working at an international school - was found dead at an apartment in Yangon, previously known as Burma.
	Scores	Mean ROUGE: 20.61 (rank 8) ‖ SummScore rank: 15
Beam #17	Summary	The man, who has not been named, was found dead at an apartment in the city of Yangon on Saturday.
	Scores	Mean ROUGE: 25.61 (rank 5) ‖ SummScore rank: 13
Beam #18	Summary	The Foreign Office said the man, who has not been named, was found dead at an apartment in Yangon, previously known as Rangoon.
	Scores	Mean ROUGE: 23.53 (rank 7) ‖ SummScore rank: 9
Beam #19	Summary	The Foreign Office said the man, who has not been named, was found dead at an apartment in Yangon, formerly known as Rangoon.
	Scores	Mean ROUGE: 23.53 (rank 7) ‖ SummScore rank: 8
Beam #20	Summary	The man, who has not been named, was found dead at an apartment in Yangon on Saturday.
	Scores	Mean ROUGE: 28.11 (rank 3) ‖ SummScore rank: 17
Reference		A British man believed to be a teacher has been found dead in Myanmar.

Table 27: Self-trained PEGASUS with beam search on XSum.

WikiHow: re-ranking from the unsupervised PEGASUS		
Source		On your scanner tool, set the PPI (Pixel per inch) to 350 so that it will create a high-quality image.; , Click on the program's shortcut icon on your desktop or programs list to launch it. Wait until the program starts. Any version will do, but the latest one is better because they have additional useful functions. , On the Menu bar located at the top of the screen, go to File and click "Open." Locate the image, select it, then click "Open." , To change it, go to "Image" on the Menu bar then click "Mode" and choose "RGB." To turn your scanned image fully grayscale, just in case you didn't change it on your scanner setting and the image has traces of colors, change the Adjustment, under "Image" on the Menu bar, into "Black and White." , On the Layer panel on the palette dock located at the bottom left of the screen, create a new layer by clicking a small paper-like icon beside the folder icon. , Change the layer name to "Skin" as this will be the first part of the image that you will add a color to. , Press "B" on your keyboard or get the Brush tool on the toolbar located at the left side of the screen. The icon is a brush. , Pick a nice skin tone color on the Swatches, which are also on the Palette dock. You may also set the foreground color to get more colors on the Toolbar, below the Zoom Tool, which is the two square that indicates as foreground and background color. A skin tone color is the lightest Orange or Yellow Orange. , Drag and move your mouse to the skin tone of your image. To avoid misplacing the colors, use the Zoom tool or simply press CTRL + Z. Use the Eraser tool, which is also located on the toolbar, to delete some misplaced part; but change the Brush Preset, which is on the Palette Dock, and choose "Soft Round" so it will create pixelated color when you erase them. , Once you're done with the skin, repeat Steps 1 to 6 with other parts of the body. For example, if you want to color the clothes, add a layer and rename the layer "Clothes." After renaming, proceed with selecting the right tool and color, and coloring the clothes. Continue with other parts, like lips, hair, background, etc. , To merge all layers, go to Layer dock, right-click, and choose "Flatten Image" or "Merge Visible" on the options. , Adjust the image tone, level, or brightness on the Image Menu to enhance the photo. , Press CTRL + S, choose JPEG as file type, add a file name, and click "Save" when you're done!
Beam #1	Summary	You may also set the foreground color to get more colors on the Toolbar, below the Zoom Tool, which is the two square that indicates as foreground and background color.
	Scores	Mean ROUGE: **13.01** (rank 7) ‖ SummScore rank: **20**
Beam #2	Summary	, On your scanner tool, set the PPI (Pixel per inch) to 350 so that it will create a high-quality image.:, Click on the program's shortcut icon on your desktop or programs list to launch it., To change it, go to "Image" on the Menu bar then click "Mode" and choose "RGB." To turn your scanned image fully grayscale, just in case you didn't change it on your scanner setting and the image has traces of colors, change the Adjustment, under "Image"—on the Palette dock located at the bottom left of the screen—into "
	Scores	Mean ROUGE: **23.18** (rank 1) ‖ SummScore rank: **2**
Beam #3	Summary	, On your scanner tool, set the PPI (Pixel per inch) to 350 so that it will create a high-quality image.;, Click on the program's shortcut icon on your desktop or programs list to launch it., To change it, go to "Image" on the Menu bar then click "Mode" and choose "RGB." To turn your scanned image fully grayscale, just in case you didn't change it on your scanner setting and the image has traces of colors, change the Adjustment, under "Image"—on the Palette dock located at the bottom left of the screen, into "
	Scores	Mean ROUGE: **23.18** (rank 1) ‖ SummScore rank: **5**
Beam #4	Summary	, On your scanner tool, set the PPI (Pixel per inch) to 350 so that it will create a high-quality image.;, Click on the program's shortcut icon on your desktop or programs list to launch it., To change it, go to "Image" on the Menu bar then click "Mode" and choose "RGB." To turn your scanned image fully grayscale, just in case you didn't change it on your scanner setting and the image has traces of colors, change the Adjustment, under "Image"—on the menu bar, into "Black and White.",On the
	Scores	Mean ROUGE: 21.05 (rank 5) ‖ SummScore rank: 11
Beam #5	Summary	, On your scanner tool, set the PPI (Pixel per inch) to 350 so that it will create a high-quality image.;, Click on the program's shortcut icon on your desktop or programs list to launch it., To change it, go to "Image" on the Menu bar then click "Mode" and choose "RGB." To turn your scanned image fully grayscale, just in case you didn't change it on your scanner setting and the image has traces of colors, change the Adjustment, under "Image"— on the menu bar, into "Black and White.",On the
	Scores	Mean ROUGE: 21.05 (rank 5) ‖ SummScore rank: 13
Beam #6	Summary	, On your scanner tool, set the PPI (Pixel per inch) to 350 so that it will create a high-quality image.;, Click on the program's shortcut icon on your desktop or programs list to launch it., To change it, go to "Image" on the Menu bar then click "Mode" and choose "RGB." To turn your scanned image fully grayscale, just in case you didn't change it on your scanner setting and the image has traces of colors, change the Adjustment, under "Image"—on the Palette dock located at the bottom left of the screen—to "
	Scores	Mean ROUGE: 22.83 (rank 2) ‖ SummScore rank: 4
Beam #7	Summary	, On your scanner tool, set the PPI (Pixel per inch) to 350 so that it will create a high-quality image.;, Click on the program's shortcut icon on your desktop or programs list to launch it., To change it, go to "Image" on the Menu bar then click "Mode" and choose "RGB." To turn your scanned image fully grayscale, just in case you didn't change it on your scanner setting and the image has traces of colors, change the Adjustment, under "Image"—on the menu bar—into "Black and White.",On the
	Scores	Mean ROUGE: 21.05 (rank 5) ‖ SummScore rank: 7
Beam #8	Summary	On your scanner tool, set the PPI (Pixel per inch) to 350 so that it will create a high-quality image.;, Click on the program's shortcut icon on your desktop or programs list to launch it., To change it, go to "Image" on the Menu bar then click "Mode" and choose "RGB." To turn your scanned image fully grayscale, just in case you didn't change it on your scanner setting and the image has traces of colors, change the Adjustment, under "Image"—on the Palette dock located at the bottom left of the screen— into "
	Scores	Mean ROUGE: **23.18** (rank 1) ‖ SummScore rank: **3**
...		
Beam #11	Summary	, On your scanner tool, set the PPI (Pixel per inch) to 350 so that it will create a high-quality image.;, Click on the program's shortcut icon on your desktop or programs list to launch it., To change it, go to "Image" on the Menu bar then click "Mode" and choose "RGB." To turn your scanned image fully grayscale, just in case you didn't change it on your scanner setting and the image has traces of colors, change the Adjustment, under "Image"—on the Palette dock located at the bottom left of the screen—and choose
	Scores	Mean ROUGE: **22.71** (rank 3) ‖ SummScore rank: **1** (**SummScore output**)
...		
Reference		Negotiators from Ireland's two biggest parties are to meet for a final effort to form a minority government.

Table 28: SummScore re-ranking applied to the unsupervised PEGASUS with beam search on WikiHow.

WikiHow: re-ranking from the PEGASUS trained on CNN/DM		
Source		Gently stabilize it by holding it steady with one or both hands. Pet your cat and talk to it in a soothing voice to calm and reassure it.If your cat resists you or is prone to scratching, then wrap your cat in the towel so that just its head is sticking out. , Once your cat is calm, place your non-dominant hand on top of your cat's head in front of its ears. Wrap your fingers around the bottom of its cheekbones for leverage., This should cause your cat's mouth to open involuntarily. Pick up the pill with your other hand. Hold the pill between your index finger and thumb. Then place your middle or ring finger on the lower molars to keep its jaw open. Do not place your finger on top of the canine tooth, i.e., the sharp fang, to keep its jaw open.If your cat will not open its mouth, then you will need to pry it open. Do this using the middle finger of the same hand holding the pill. Place your middle finger over the small incisor teeth in the front of your cat's mouth at the bottom. Then apply gentle pressure to push your cat's jaw open. , Drop the pill as far back as you can on your cat's tongue, i.e., the base of the tongue. Or, use your index finger and thumb to push the pill over the back of your cat's tongue. If you do it this way, you must do it quickly to prevent your fingers from getting bitten.Alternatively, you can use a pill syringe to place the pill at the base of your cat's tongue. This way you can avoid having to put your fingers in your cat's mouth. Coat the pill with butter to make swallowing it easier. , Once you have placed the pill at the base of its tongue, use your hands to keep its mouth closed by applying gentle pressure. Also re-position your cat's head so that it is level instead of tilted back; this will make it easier for your cat to swallow the pill. Gently rub your cat's throat or blow in its nose to encourage it to swallow the pill.Try not to let your cat go until the pil lis swallowed. Place a drop of water on its nose. If your cat licks the water off, then this means that it has swallowed the pill. , Tricyclic antidepressants should be given with food to decrease side effects such as vomiting. Therefore, feed your cat directly after giving it its medication. If your cat will not eat, then entice it to eat by feeding it its favorite food like cheese, or by giving it a treat.Make sure to provide your cat with water as well. You can also give your cat flavored water like juice from a tuna can to entice it to drink.
Beam #1	Summary	Pet your cat and talk to it in a soothing voice to calm and reassure it. If your cat resists you or is prone to scratching, then wrap your cat in a towel so that just its head is sticking out.
	Scores	Mean ROUGE: **17.17** (rank 8) ‖ SummScore rank: **12**
Beam #2	Summary	Pet your cat and talk to it in a soothing voice to calm and reassure it. If your cat resists you or is prone to scratching, then wrap your cat in the towel so that just its head is sticking out.
	Scores	Mean ROUGE: 18.78 (rank 6) ‖ SummScore rank: 8
Beam #3	Summary	Pet your cat and talk to it in a soothing voice to calm and reassure it. If your cat resists you or is prone to scratching, then wrap your cat in a towel so that just its head is sticking out.
	Scores	Mean ROUGE: 17.17 (rank 8) ‖ SummScore rank: 11
Beam #4	Summary	Pet your cat and talk to it in a soothing voice to calm and reassure it. gently stabilize it by holding it steady with one or both hands. Rub your cat's throat or blow in its nose to encourage it to swallow the pill.
	Scores	Mean ROUGE: 18.68 (rank 7) ‖ SummScore rank: 7
Beam #5	Summary	Pet your cat and talk to it in a soothing voice to calm and reassure it. Wrap your fingers around the bottom of its cheekbones for leverage. Alternatively, you can use a pill syringe to place the pill at the base of your cat's tongue.
	Scores	Mean ROUGE: **26.93** (rank 3) ‖ SummScore rank: **1** (**SummScore output**)
Beam #6	Summary	Tricyclic antidepressants should be given with food to decrease side effects such as vomiting. If your cat resists you or is prone to scratching, then wrap your cat in a towel so that just its head is sticking out.
	Scores	Mean ROUGE: 15.09 (rank 10) ‖ SummScore rank: 17
Beam #7	Summary	Pet your cat and talk to it in a soothing voice to calm and reassure it. gently stabilize it by holding it steady with one or both hands.
	Scores	Mean ROUGE: 8.72 (rank 12) ‖ SummScore rank: 19
Beam #8	Summary	Gently stabilize it by holding it steady with one or both hands. If your cat resists you or is prone to scratching, then wrap your cat in a towel so that just its head is sticking out.
	Scores	Mean ROUGE: 17.18 (rank 7) ‖ SummScore rank: 14
...		
Beam #14	Summary	If your cat resists you or is prone to scratching, then wrap your cat in a towel so that just its head is sticking out. Alternatively, you can use a pill syringe to place the pill at the base of your cat's tongue.
	Scores	Mean ROUGE: **30.74** (rank 1) ‖ SummScore rank: **2**
...		
Reference		Negotiators from Ireland's two biggest parties are to meet for a final effort to form a minority government.

Table 29: SummScore re-ranking applied to the PEGASUS fine-tuned on CNN/DM with beam search on WikiHow.

WikiHow: re-ranking from the self-trained PEGASUS		
Source		Talking in private will help you feel less nervous and more comfortable about asking for help.It will also give your teacher more time to help you because they aren't trying to teach at the same time. For example, you could say, "When you have time later today, could we talk about a problem I'm having?" If you're afraid to approach your teacher, put a note on their chair. The note could say, "Can we talk later about something personal? Thanks, Mark." You could also send your teacher an email or message letting them know you would like to ask their help with a personal issue. , Sometimes teachers will try to offer advice on what you should do, when all you wanted was for them to listen. If you let your teacher know how they can help, it will make it easier for them to give you the help you want and need.Think about what kind of help you want. Ask yourself, "Do I want her to listen, to give me advice, or to do something about the problem?" Tell your teacher how they can help. For example, "Can you help me come up with ways to make more friends?" If you don't know how you want your teacher to help, you it's okay for you to say that, too. Try saying, "I need your help with a problem, but I don't know what kind of help I need." , Also, being honest will help your teacher figure out the right way to help you. The more truthful information your teacher has about what is going on, the more they will be able to help. If you are afraid you will get in trouble, then say that. You could say, "I need your help with something but I'm scared I'll get in trouble." If you are asking for help, but also trying not to get someone else in trouble, you could leave out names, but still be honest about what is going on. For example, "My friend is thinking about cheating on a test and I need advice on how to stop them from doing something so stupid."
Beam #1	Summary	You could say, "I need your help with something but I'm scared I'll get in trouble." If you are asking for help, but also trying not to get someone else in trouble, you could leave out names, but still be honest about what is going on. For example, "My friend is thinking about cheating on a test and I need advice on how to stop them from doing something so stupid."
	Scores	Mean ROUGE: **22.54** (rank 14) || SummScore rank: **14**
Beam #2	Summary	Think about what kind of help you want. You could say, "I need your help with something but I'm scared I'll get in trouble." If you are asking for help, but also trying not to get someone else in trouble, you could leave out names, but still be honest about what is going on.If you're afraid to approach your teacher, put a note on their chair.
	Scores	Mean ROUGE: 30.92 (rank 7) || SummScore rank: 18
Beam #3	Summary	For example, you could say, "When you have time later today, could we talk about a problem I'm having?" If you're afraid to approach your teacher, put a note on their chair. Try saying, "I need your help with a problem, but I don't know what kind of help I need.", Also, being honest will help your teacher figure out the right way to help you.
	Scores	Mean ROUGE: 29.41 (rank 9) || SummScore rank: 15
Beam #4	Summary	Think about what kind of help you want. You could say, "I need your help with something but I'm scared I'll get in trouble." If you are asking for help, but also trying not to get someone else in trouble, you could leave out names, but still be honest about what is going on.
	Scores	Mean ROUGE: 28.40 (rank 10) || SummScore rank: 20
Beam #5	Summary	Think about what kind of help you want. You could say, "I need your help with something but I'm scared I'll get in trouble." If you are asking for help, but also trying not to get someone else in trouble, you could leave out names, but still be honest about what is going on. For example, "My friend is thinking about cheating on a test and I need advice on how to stop them from doing something so stupid."
	Scores	Mean ROUGE: 27.36 (rank 12) || SummScore rank: 11
Beam #6	Summary	Think about what kind of help you want. You could say, "I need your help with something but I'm scared I'll get in trouble." If you are asking for help, but also trying not to get someone else in trouble, you could leave out names, but still be honest about what is going on.You could also send your teacher an email or message letting them know you would like to ask their help with a personal issue.
	Scores	Mean ROUGE: 31.47 (rank 6) || SummScore rank: 9
Beam #7	Summary	Think about what kind of help you want. You could say, "I need your help with something but I'm scared I'll get in trouble." If you are asking for help, but also trying not to get someone else in trouble, you could leave out names, but still be honest about what is going on.If you are afraid to approach your teacher, put a note on their chair.
	Scores	Mean ROUGE: 30.92 (rank 7) || SummScore rank: 17
Beam #8	Summary	You could say, "I need your help with something but I'm scared I'll get in trouble." If you are asking for help, but also trying not to get someone else in trouble, you could leave out names, but still be honest about what is going on. You could also send your teacher an email or message letting them know you would like to ask their help with a personal issue.
	Scores	Mean ROUGE: 28.06 (rank 11) || SummScore rank: 13
...		
Beam #19	Summary	For example, you could say, "When you have time later today, could we talk about a problem I'm having?" If you're afraid to approach your teacher, put a note on their chair. Try saying, "I need your help with a problem, but I don't know what kind of help I need.", Also, being honest will help your teacher figure out the right way to help you. The more truthful information your teacher has about what is going on, the more they will be able to help
	Scores	Mean ROUGE: **35.08** (rank 1) || SummScore rank: **2**
Beam #20	Summary	Think about what kind of help you want. You could say, "I need your help with something but I'm scared I'll get in trouble." If you are asking for help, but also trying not to get someone else in trouble, you could leave out names, but still be honest about what is going on.If you're afraid to approach your teacher, put a note on their chair. It will also give your teacher more time to help you because they aren't trying to teach at the same time
	Scores	Mean ROUGE: **34.51** (rank 2) || SummScore rank: **1** (**SummScore output**)
Reference		Negotiators from Ireland's two biggest parties are to meet for a final effort to form a minority government.

Table 30: Self-trained PEGASUS with beam search on WikiHow.

SAMSum: re-ranking from the unsupervised PEGASUS		
Source		Mike: dude, wendy has grown prettier Dave:<laughing><laughing><thumb_up> Mike:<love><love> Dave: i know right? Mike: yeah, since she came from Houston, she looks like an angel Dave: i'll have to hit on her soon Mike: haha, stay off, i hear Jerry is her lover Dave: really<sad> Mike: yeah Dave: since when Mike: haha, i dont know, but you can push your luck Dave: haha, i will Mike:<blink> good luck Dave: yeah, ill need it
Beam #1	Summary	Mike: dude, wendy has grown prettier Dave: Mike: Dave: i know right?
	Scores	Mean ROUGE: **19.73** (rank 3) ‖ SummScore rank: **19**
Beam #2	Summary	Mike: guy, wendy has grown prettier Dave: Mike: Dave: i know right?
	Scores	Mean ROUGE: 19.73 (rank 3) ‖ SummScore rank: 20
Beam #3	Summary	Mike: Yeah, since she came from Houston, she looks like an angel Dave: i'll have to hit on her soon Mike: haha, stay off, i hear Jerry is her lover Dave: really Mike: yeah Dave: since when Mike:haha, i dont know, but you can push
	Scores	Mean ROUGE: 22.31 (rank 2) ‖ SummScore rank: 2
Beam #4	Summary	Mike: uh, since she came from Houston, she looks like an angel Dave: i'll have to hit on her soon Mike: haha, stay off, i hear Jerry is her lover Dave: really Mike: yeah Dave: since when Mike:haha, i dont know, but you can push
	Scores	Mean ROUGE: 22.31 (rank 2) ‖ SummScore rank: 11
Beam #5	Summary	Mike: sure, since she came from Houston, she looks like an angel Dave: i'll have to hit on her soon Mike: haha, stay off, i hear Jerry is her lover Dave: really Mike: yeah Dave: since when Mike:haha, i dont know, but you can push
	Scores	Mean ROUGE: 22.31 (rank 2) ‖ SummScore rank: 13
Beam #6	Summary	Mike:yeah, since she came from Houston, she looks like an angel Dave: i'll have to hit on her soon Mike: haha, stay off, i hear Jerry is her lover Dave: really Mike: yeah Dave: since when Mike:haha, i dont know, but you can push
	Scores	Mean ROUGE: 22.31 (rank 2) ‖ SummScore rank: 6
Beam #7	Summary	Mike: yes, since she came from Houston, she looks like an angel Dave: i'll have to hit on her soon Mike: haha, stay off, i hear Jerry is her lover Dave: really Mike: yeah Dave: since when Mike:haha, i dont know, but you can push
	Scores	Mean ROUGE: 22.31 (rank 2) ‖ SummScore rank: 10
Beam #8	Summary	Mike: Yeah, since she came from Houston, she looks like an angel Dave: i'll have to hit on her soon Mike: haha, stay off, i hear Jerry is her lover Dave: really Mike: yeah Dave: since when Mike: hahaha, i dont know, but you can push
	Scores	Mean ROUGE: **22.31** (rank 2) ‖ SummScore rank: **1 (SummScore output)**
Beam #9	Summary	Mike: uh, yeah, since she came from Houston, she looks like an angel Dave: i'll have to hit on her soon Mike: haha, stay off, i hear Jerry is her lover Dave: really Mike: yeah Dave: since when Mike is haha, i dont know, but you
	Scores	Mean ROUGE: 22.31 (rank 2) ‖ SummScore rank: 18
Beam #10	Summary	Mike: obviously, since she came from Houston, she looks like an angel Dave: i'll have to hit on her soon Mike: haha, stay off, i hear Jerry is her lover Dave: really Mike: yeah Dave: since when Mike:haha, i dont know, but you can push
	Scores	Mean ROUGE: 22.31 (rank 2) ‖ SummScore rank: 15
Beam #11	Summary	Mike: uh, since she came from Houston, she looks like an angel Dave: i'll have to hit on her soon Mike: haha, stay off, i hear Jerry is her lover Dave: really Mike: yeah Dave: since when Mike: hahaha, i dont know, but you can push
	Scores	Mean ROUGE: 22.31 (rank 2) ‖ SummScore rank: 9
Beam #12	Summary	Mike: sure, since she came from Houston, she looks like an angel Dave: i'll have to hit on her soon Mike: haha, stay off, i hear Jerry is her lover Dave: really Mike: yeah Dave: since when Mike: hahaha, i dont know, but you can push
	Scores	Mean ROUGE: 22.31 (rank 2) ‖ SummScore rank: 12
Beam #13	Summary	Mike: uh, yeah, since she came from Houston, she looks like an angel Dave: i'll have to hit on her soon Mike: haha, stay off, i hear Jerry is her lover Dave: really Mike: yeah Dave: since when Mike:haha, i dont know, but you
	Scores	Mean ROUGE: **22.64** (rank 1) ‖ SummScore rank: **5**
Beam #14	Summary	Mike: yes, since she came from Houston, she looks like an angel Dave: i'll have to hit on her soon Mike: haha, stay off, i hear Jerry is her lover Dave: really Mike: yeah Dave: since when Mike: hahaha, i dont know, but you can push
	Scores	Mean ROUGE: 22.31 (rank 2) ‖ SummScore rank: 8
Beam #15	Summary	Mike: no, since she came from Houston, she looks like an angel Dave: i'll have to hit on her soon Mike: haha, stay off, i hear Jerry is her lover Dave: really Mike: yeah Dave: since when Mike:haha, i dont know, but you can push
	Scores	Mean ROUGE: 22.31 (rank 2) ‖ SummScore rank: 17
Beam #16	Summary	Mike:yeah, since she came from Houston, she looks like an angel Dave: i'll have to hit on her soon Mike: haha, stay off, i hear Jerry is her lover Dave: really Mike: yeah Dave: since when Mike: hahaha, i dont know, but you can push
	Scores	Mean ROUGE: 22.31 (rank 2) ‖ SummScore rank: 4
Beam #17	Summary	Mike: cool, since she came from Houston, she looks like an angel Dave: i'll have to hit on her soon Mike: haha, stay off, i hear Jerry is her lover Dave: really Mike: yeah Dave: since when Mike:haha, i dont know, but you can push
	Scores	Mean ROUGE: 22.31 (rank 2) ‖ SummScore rank: 14
Beam #18	Summary	Mike:Yeah, since she came from Houston, she looks like an angel Dave: i'll have to hit on her soon Mike: haha, stay off, i hear Jerry is her lover Dave: really Mike: yeah Dave: since when Mike:haha, i dont know, but you can push
	Scores	Mean ROUGE: 22.31 (rank 2) ‖ SummScore rank: 7
Beam #19	Summary	Mike: uh, yeah, since she came from Houston, she looks like an angel Dave: i'll have to hit on her soon Mike: haha, stay off, i hear Jerry is her lover Dave: really Mike: yeah Dave: since when Mike: hahaha, i dont know, but you
	Scores	Mean ROUGE: **22.64** (rank 1) ‖ SummScore rank: **3**
Beam #20	Summary	Mike: uh, since she came from Houston, she looks like an angel Dave: i'll have to hit on her soon Mike: haha, stay off, i hear Jerry is her lover Dave: really Mike: yeah Dave: since when Mike:haha, i don't know, but you
	Scores	Mean ROUGE: **22.64** (rank 1) ‖ SummScore rank: **16**
Reference		Mike and Dave notice Wendy got prettier. Dave wants to hit on her, but she's with Jerry. He'll try anyway.

Table 31: SummScore re-ranking applied to the unsupervised PEGASUS with beam search on SAMSum.

SAMSum: re-ranking from ChatGPT		
Source		Adam: I have a juicy gossip for you guys ;) Borys: About whom? Adam: So apparently, Iga planned a weekend getaway with her boyfriend, but they had to cancel because he couldn't talk his group into changing the date of the presentation :D Katia: And she is pissed? Adam: What do you think? That is why they aren't sitting together since Monday ;) Katia: Who would have expected such a gossip girl from you :D
Candidate #1	Summary	Adam shares a gossip about Iga's cancelled weekend getaway with her boyfriend due to his group's presentation, which has caused tension between them. Katia is surprised by Adam's gossiping behavior.
	Scores	Mean ROUGE: **19.80** (rank 16) ‖ SummScore rank: **7**
Candidate #2	Summary	Adam shares juicy gossip about Iga's cancelled weekend getaway with her boyfriend due to a presentation, which caused her to be upset and avoid sitting with her boyfriend since Monday. Katia jokes about Adam's unexpected role as a gossip girl.
	Scores	Mean ROUGE: 23.92 (rank 13) ‖ SummScore rank: 5
Candidate #3	Summary	Adam shares juicy gossip about Iga's cancelled weekend getaway with her boyfriend due to his group's presentation, causing tension between them, and surprising his friends with his gossiping skills.
	Scores	Mean ROUGE: 20.38 (rank 15) ‖ SummScore rank: 20
Candidate #4	Summary	Adam shares gossip about Iga's cancelled weekend getaway with her boyfriend due to his group's presentation, causing tension between them and leading to them not sitting together. Katia is surprised by Adam's gossiping tendencies.
	Scores	Mean ROUGE: 26.79 (rank 11) ‖ SummScore rank: 11
Candidate #5	Summary	Adam shares gossip about Iga's cancelled weekend getaway with her boyfriend, causing her to be angry and not sit with him since Monday, surprising his friends.
	Scores	Mean ROUGE: 29.91 (rank 7) ‖ SummScore rank: 15
Candidate #6	Summary	Adam shares juicy gossip about Iga's cancelled weekend getaway with her boyfriend due to his group's presentation date, which caused tension between them. Katia is surprised by Adam's gossiping skills.
	Scores	Mean ROUGE: 19.80 (rank 16) ‖ SummScore rank: 6
Candidate #7	Summary	Adam shares gossip about Iga's cancelled weekend getaway with her boyfriend due to his group presentation, causing tension between them and resulting in them not sitting together since Monday. Katia is surprised by Adam's gossiping.
	Scores	Mean ROUGE: 31.28 (rank 6) ‖ SummScore rank: 10
Candidate #8	Summary	Adam shares gossip about Iga's cancelled weekend getaway with her boyfriend due to his group's presentation, causing tension between them and resulting in them not sitting together since Monday, surprising his friends.
	Scores	Mean ROUGE: 32.59 (rank 3) ‖ SummScore rank: 19
Candidate #9	Summary	Adam shares gossip with his friends about Iga's cancelled weekend getaway with her boyfriend, which has caused tension between them. His friends are surprised by his gossiping.
	Scores	Mean ROUGE: 18.28 (rank 18) ‖ SummScore rank: 3
Candidate #10	Summary	Adam shares gossip about Iga's canceled weekend getaway with her boyfriend due to his group's presentation date, causing tension between them and resulting in them not sitting together since Monday. Katia teases Adam about his unexpected role as a gossip girl.
	Scores	Mean ROUGE: 31.42 (rank 5) ‖ SummScore rank: 8
Candidate #11	Summary	Adam shares juicy gossip that Iga's weekend getaway with her boyfriend was cancelled and she is angry about it, causing her to not sit with him since Monday. Katia is surprised by Adam's gossiping.
	Scores	Mean ROUGE: 28.05 (rank 9) ‖ SummScore rank: 2
Candidate #12	Summary	Adam shares juicy gossip about Iga's cancelled weekend getaway with her boyfriend due to a presentation, causing tension between them and leading to them not sitting together. Katia is surprised by Adam's gossiping tendencies.
	Scores	Mean ROUGE: 27.15 (rank 10) ‖ SummScore rank: 12
Candidate #13	Summary	Adam shares juicy gossip about Iga's cancelled weekend getaway with her boyfriend due to his work presentation, causing tension in their relationship and leading to them not sitting together. Katia is surprised by Adam's gossiping tendencies.
	Scores	Mean ROUGE: 25.56 (rank 12) ‖ SummScore rank: 15
Candidate #14	Summary	Adam shares juicy gossip that Iga's weekend getaway with her boyfriend was cancelled due to a presentation date, leading to Iga being upset and not sitting with her boyfriend.
	Scores	Mean ROUGE: 31.49 (rank 4) ‖ SummScore rank: 9
Candidate #15	Summary	Adam shares juicy gossip with his friends about Iga's cancelled weekend getaway with her boyfriend due to work, causing tension between them. His friends are surprised by Adam's gossiping tendencies.
	Scores	Mean ROUGE: 17.23 (rank 19) ‖ SummScore rank: 13
Candidate #16	Summary	Adam shares gossip about Iga's cancelled weekend getaway with her boyfriend, causing tension between them, and surprises Katia with his gossiping.
	Scores	Mean ROUGE: 19.00 (rank 17) ‖ SummScore rank: 18
Candidate #17	Summary	Adam shares juicy gossip about Iga's cancelled weekend getaway with her boyfriend due to his work presentation, causing tension between the couple and leading to them not sitting together since Monday.
	Scores	Mean ROUGE: **33.53** (rank 1) ‖ SummScore rank: **17**
Candidate #18	Summary	Adam shares juicy gossip about Iga's cancelled weekend getaway with her boyfriend due to his group's presentation, which has caused tension between them and they are not sitting together since Monday. Katia is surprised by Adam's gossiping skills.
	Scores	Mean ROUGE: **33.09** (rank 2) ‖ SummScore rank: **1** (**SummScore output**)
Candidate #19	Summary	Adam shares juicy gossip that Iga's weekend getaway with her boyfriend got canceled because of his group's presentation and she is now not sitting with him. Katia is surprised by Adam's gossiping skills.
	Scores	Mean ROUGE: 23.52 (rank 14) ‖ SummScore rank: 4
Candidate #20	Summary	Adam shares gossip about Iga's cancelled weekend getaway with her boyfriend due to his presentation, causing tension between them and resulting in them not sitting together since Monday, surprising Katia who didn't expect such gossip from Adam.
	Scores	Mean ROUGE: 28.73 (rank 8) ‖ SummScore rank: 16
Reference		Iga and her boyfriend are not sitting together since Monday, because they had to cancel their weekend getaway. It made Iga very angry. Iga's boyfriend failed to convince a group to change the date of the presentation.

Table 32: SummScore re-ranking applied to ChatGPT with top-p sampling on SAMSum.

SAMSum: re-ranking from the PEGASUS trained on XSum		
Source		Mia: God, he keeps talking about the stupid wall Jennifer: I think he's just senile Karine: unfortunately it's much more Karine: this populist strategy to scare people with a minority Karine: and then present himself as the only savior Peter: Karine is right, I think he's not as senile and stupid as we would like to believe Mia: so it's even worse - he's pure evil Peter: possibly yes Karine: I really think so, because it's not only about politics anymore Karine: he is destroying the Earth Karine: trying to shake the world's order Jennifer: but he's also quite stupid Jennifer: and the dangerous thing is his self confidence Jennifer: he's persistence Jennifer: and ignorance Karine: but we have a chance to change it Karine: maybe Trump will be over soon? Mia: we will see Mia: but we should do anything to win the next elections Karine: everything!
Beam #1	Summary	In our series of letters from British schoolchildren, Mia, Jennifer and Karine give their views on US President Donald Trump.
	Scores	Mean ROUGE: **23.16** (rank 5) ‖ SummScore rank: **18**
Beam #2	Summary	In our series of letters from British children, Mia, Jennifer and Karine give their views on US President Donald Trump.
	Scores	Mean ROUGE: 23.16 (rank 5) ‖ SummScore rank: 17
Beam #3	Summary	All images are copyrighted.
	Scores	Mean ROUGE: 00.00 (rank 7) ‖ SummScore rank: 20
Beam #4	Summary	In our series of letters from British children, Mia, Jennifer and Karine tell us what they think about US President Donald Trump.
	Scores	Mean ROUGE: 21.65 (rank 6) ‖ SummScore rank: 6
Beam #5	Summary	In our series of letters from British children, Mia, Jennifer, Karine and Peter tell us what they think about Donald Trump.
	Scores	Mean ROUGE: 36.19 (rank 2) ‖ SummScore rank: 5
Beam #6	Summary	In our series of letters from British children, Mia, Jennifer, Karine and Peter tell us what they think about US President Donald Trump.
	Scores	Mean ROUGE: 33.89 (rank 3) ‖ SummScore rank: 5
Beam #7	Summary	In our series of letters from British children, Mia, Jennifer, Karine and Peter tell us what they think of Donald Trump.
	Scores	Mean ROUGE: 36.19 (rank 2) ‖ SummScore rank: 10
Beam #8	Summary	In our series of letters from British children, Mia, Jennifer, Karine and Peter tell us what they think of US President Donald Trump.
	Scores	Mean ROUGE: 33.89 (rank 3) ‖ SummScore rank: **11**
Beam #9	Summary	In our series of letters from British children, Mia, Jennifer, Karine and Peter give their views on US President Donald Trump.
	Scores	Mean ROUGE: 36.19 (rank 2) ‖ SummScore rank: 16
Beam #10	Summary	In our series of letters from British children, Mia, Jennifer and Karine tell us what they think about Donald Trump.
	Scores	Mean ROUGE: 23.16 (rank 5) ‖ SummScore rank: 3
Beam #11	Summary	In our series of letters from British schoolchildren, Mia, Jennifer and Karine give their views on Donald Trump.
	Scores	Mean ROUGE: 24.89 (rank 4) ‖ SummScore rank: 19
Beam #12	Summary	In our series of letters from British children, Mia, Jennifer and Karine tell us what they think of Donald Trump.
	Scores	Mean ROUGE: 23.16 (rank 5) ‖ SummScore rank: 9
Beam #13	Summary	In our series of letters from British schoolchildren, Mia, Jennifer and Karine tell us what they think of Donald Trump.
	Scores	Mean ROUGE: 23.16 (rank 5) ‖ SummScore rank: 13
Beam #14	Summary	In our series of letters from British schoolchildren, Mia, Jennifer and Karine tell us what they think about Donald Trump.
	Scores	Mean ROUGE: 23.16 (rank 5) ‖ SummScore rank: 8
Beam #15	Summary	In our series of letters from British children, Mia, Jennifer and Karine give their views on Donald Trump.
	Scores	Mean ROUGE: 24.89 (rank 4) ‖ SummScore rank: 15
Beam #16	Summary	In our series of letters from British children, Mia, Jennifer, Karine and Peter give their views on Donald Trump.
	Scores	Mean ROUGE: **38.83** (rank 1) ‖ SummScore rank: **14**
Beam #17	Summary	In our series of letters from British children, Mia, Jennifer, Karine and Peter tell us what they think of Donald Trump's presidency.
	Scores	Mean ROUGE: 33.89 (rank 3) ‖ SummScore rank: 7
Beam #18	Summary	In our series of letters from British children, Mia, Jennifer, Karine and Peter tell us what they think about Donald Trump's presidency.
	Scores	Mean ROUGE: 33.89 (rank 3) ‖ SummScore rank: 2
Beam #19	Summary	n our series of letters from British children, Mia, Jennifer, Karine and Peter give their views on US President Donald Trump's presidency.
	Scores	Mean ROUGE: 33.89 (rank 3) ‖ SummScore rank: 12
Beam #20	Summary	In our series of letters from British children, Mia, Jennifer, Karine and Peter tell us why they think Donald Trump is a fool.
	Scores	Mean ROUGE: **33.89** (rank 3) ‖ SummScore rank: **1 (SummScore output)**
Reference		Mia, Jennifer, Karine and Peter do not like Trump.

Table 33: SummScore re-ranking applied to the PEGASUS transferred from XSum with beam search on SAMSum.

SAMSum: re-ranking from the self-trained PEGASUS		
Source		Louis: Hey, hows your day? :D Cheryl: Okaaay… I guess Louis: Aha, someone's feeling a bit down, am I right? Cheryl: yea, sort of… Louis: Go on, tell me what happened Cheryl: I…just had an argument with my mom Louis: Jesus what again Cheryl: I forgot to close the window when I was leaving home! Louis: And that's it? Cheryl: No, not only… Ya know, wouldn't be that bad, but I got angry, started screaming and everything ;/ Louis: not a good idea, babe Cheryl: I knoooow Louis: Was it really bad? Cheryl: I suppose yea, she kicked me out xd Louis: WHAT Cheryl: I mean I don't have to move right now, but she gave me time till the end of the year Louis: I'm sorry… Cheryl: Naah, don't be, I believe it's for good. I couldn't stand her anyway xD
Beam #1	Summary	Cheryl: No, not only... Ya know, wouldn't be that bad, but I got angry, started screaming and everything ;/ Louis: not a good idea, babe Cheryl: I knoooow Louis: Was it really bad?
	Scores	Mean ROUGE: **12.15** (rank 8) ‖ SummScore rank: **8**
Beam #2	Summary	Cheryl: no, not only... Ya know, wouldn't be that bad, but I got angry, started screaming and everything ;/ Louis: not a good idea, babe Cheryl: I knoooow Louis: Was it really bad?
	Scores	Mean ROUGE: 12.15 (rank 8) ‖ SummScore rank: 10
Beam #3	Summary	Cheryl: Ya know, wouldn't be that bad, but I got angry, started screaming and everything ;/ Louis: not a good idea, babe Cheryl: I knoooow Louis: Was it really bad?
	Scores	Mean ROUGE: 12.73 (rank 7) ‖ SummScore rank: 18
Beam #4	Summary	Cheryl: No, not only... Ya know, wouldn't be that bad, but I got angry, started screaming and everything ;/ Louis: not a good idea, babe Cheryl: I knoooow Louis: Was it really bad? Cheryl:
	Scores	Mean ROUGE: 11.97 (rank 9) ‖ SummScore rank: 9
Beam #5	Summary	Cheryl: No, not only... Ya know, wouldn't be that bad, but I got angry, started screaming and everything ;/ Louis: not a good idea, babe Cheryl: I knoooow Louis: Was it really bad? Cheryl
	Scores	Mean ROUGE: 11.97 (rank 9) ‖ SummScore rank: 11
Beam #6	Summary	Cheryl: No, not only... Ya know, wouldn't be that bad, but I got angry, started screaming and everything ;/ Louis: not a good idea, babe Cheryl: i knoooow Louis: Was it really bad?
	Scores	Mean ROUGE: 12.15 (rank 8) ‖ SummScore rank: 16
Beam #7	Summary	Cheryl: No, not only... Ya know, wouldn't be that bad, but I got angry, started screaming and everything ;/ Louis: not a good idea, babe Cheryl:I knoooow Louis: Was it really bad?
	Scores	Mean ROUGE: 12.15 (rank 8) ‖ SummScore rank: 15
Beam #8	Summary	Cheryl: :D Cheryl: Okaaay... I guess Louis: Aha, someone's feeling a bit down, am I right? Cheryl: yea, sort of... Louis: Go on, tell me what happened Cheryl: I...just had an argument with my mom
	Scores	Mean ROUGE: 17.23 (rank 5) ‖ SummScore rank: 5
Beam #9	Summary	Cheryl: No, not only... Ya know, wouldn't be that bad, but I got angry, started screaming and everything ;/ Louis: not a good idea, babe Cheryl: I knoooow ; Louis: Was it really bad?
	Scores	Mean ROUGE: 12.15 (rank 8) ‖ SummScore rank: 12
Beam #10	Summary	Cheryl: yea, sort of... Louis: Go on, tell me what happened Cheryl: I...just had an argument with my mom Louis: Jesus what again Cheryl:I forgot to close the window when I was leaving home!
	Scores	Mean ROUGE: **29.19** (rank 1) ‖ SummScore rank: **17**
Beam #11	Summary	Cheryl: yea, sort of... Louis: Go on, tell me what happened Cheryl: I...just had an argument with my mom Louis: Jesus what again Cheryl:I forgot to close the window when I was leaving home! Louis: And that's it?
	Scores	Mean ROUGE: 29.00 (rank 2) ‖ SummScore rank: 7
Beam #12	Summary	:D Cheryl: Okaaay... I guess Louis: Aha, someone's feeling a bit down, am I right? Cheryl: yea, sort of... Louis: Go on, tell me what happened Cheryl: I...just had an argument with my mom
	Scores	Mean ROUGE: **17.48** (rank 4) ‖ SummScore rank: **1 (SummScore output)**
Beam #13	Summary	:D Cheryl: Okaaay... I guess Louis: Aha, someone's feeling a bit down, am I right? Cheryl: yea, sort of... Louis: Go on, tell me what happened Cheryl: I...just had an argument with my mom Louis
	Scores	Mean ROUGE: 17.23 (rank 5) ‖ SummScore rank: 2
Beam #14	Summary	Cheryl: yea, sort of... Louis: Go on, tell me what happened Cheryl:I...just had an argument with my mom Louis: Jesus what again Cheryl: I forgot to close the window when I was leaving home! Louis: And that's it?
	Scores	Mean ROUGE: 29.00 (rank 2) ‖ SummScore rank: 6
Beam #15	Summary	Cheryl: No, not only... Ya know, wouldn't be that bad, but I got angry, started screaming and everything ;/ Louis: not a good idea, babe Cheryl: I knoooow Louis: Was it really bad? Louis: Go on.
	Scores	Mean ROUGE: 11.62 (rank 10) ‖ SummScore rank: 14
Beam #16	Summary	Cheryl: :D Cheryl: Okaaay... I guess Louis: Aha, someone's feeling a bit down, am I right? Louis: Go on, tell me what happened Cheryl: I...ju st had an argument with my mom Louis: Jesus what again Cheryl:
	Scores	Mean ROUGE: 16.04 (rank 6) ‖ SummScore rank: 4
Beam #17	Summary	:D Cheryl: Okaaay... I guess Louis: Aha, someone's feeling a bit down, am I right? Cheryl: yea, sort of... Louis: Go on, tell me what happened Cheryl: I...just had an argument with my mom Louis:
	Scores	Mean ROUGE: 17.23 (rank 5) ‖ SummScore rank: 3
Beam #18	Summary	Cheryl: Naah, don't be, I believe it's for good. I couldn't stand her anyway xD
	Scores	Mean ROUGE: 5.23 (rank 11) ‖ SummScore rank: 19
Beam #19	Summary	Cheryl: Ok, I suppose yea, she kicked me out xd Louis: WHAT Cheryl: I mean I don't have to move right now, but she gave me time till the end of the year Louis: I'm sorry... Cheryl: Naah, don'
	Scores	Mean ROUGE: 28.10 (rank 3) ‖ SummScore rank: 13
Beam #20	Summary	Cheryl: Hi Louis Louis: Hi Cheryl, how are you?
	Scores	Mean ROUGE: 3.17 (rank 12) ‖ SummScore rank: 20
Reference		Cheryl had an argument with her mom. She forgot to close the window, got angry and started a fight. Her mom gave her time till the end of the year to move out.

Table 34: Self-trained PEGASUS with beam search on SAMSum.

ACL 2023 Responsible NLP Checklist

A For every submission:

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

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

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

☒ A4. Have you used AI writing assistants when working on this paper?
Not relevant.

B ☑ Did you use or create scientific artifacts?

Sections 4.1, 4.2, 4.3

☑ B1. Did you cite the creators of artifacts you used?
Sections 4.1

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ 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)?
Not applicable. Left blank.

☐ 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.?
Not applicable. Left blank.

☑ 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.
Section 4.1

C ☑ Did you run computational experiments?

Sections 4.2, 4.3, 4.4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 3.3

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 3.3

☑ 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?
Sections 4.2, 4.3, 4.4

☑ 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.)?
Section 4.1

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Section 4.5

☑ 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.?
Section 4.5

☑ 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)?
Section 4.5

☐ 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?
Section 4.5