Focus Attention: Promoting Faithfulness and Diversity in Summarization

Rahul Aralikatte[*]
University of Copenhagen
rahul@di.ku.dk

Shashi Narayan
Google Research
shashinarayan@google.com

Joshua Maynez
Google Research
joshuahm@google.com

Sascha Rothe
Google Research
rothe@google.com

Ryan McDonald[*]
ASAPP
ryanmcd@asapp.com

Abstract

Professional summaries are written with document-level information, such as the theme of the document, in mind. This is in contrast with most seq2seq decoders which simultaneously learn to focus on salient content, while deciding what to generate, at each decoding step. With the motivation to narrow this gap, we introduce Focus Attention Mechanism, a simple yet effective method to encourage decoders to proactively generate tokens that are similar or topical to the input document. Further, we propose a Focus Sampling method to enable generation of diverse summaries, an area currently understudied in summarization. When evaluated on the BBC extreme summarization task, two state-of-the-art models augmented with Focus Attention generate summaries that are closer to the target and more faithful to their input documents, outperforming their vanilla counterparts on ROUGE and multiple faithfulness measures. We also empirically demonstrate that Focus Sampling is more effective in generating diverse and faithful summaries than top-k or nucleus sampling-based decoding methods.

1 Introduction

Document summarization — producing the shorter version of a document while preserving salient information (Mani, 2001; Nenkova and McKeown, 2011) — is challenging even for humans. Today, systems can generate summaries with a high level of fluency and coherence. This is due to recent advances such as sequence-to-sequence architectures (seq2seq) with attention and copy mechanism (Hochreiter and Schmidhuber, 1997; Bahdanau et al., 2015; Gu et al., 2016), fully attention-based Transformer architectures (Vaswani et al., 2017), and large pretrained language models (Devlin et al., 2019; Radford et al., 2018; Yang et al., 2019; Liu et al., 2019; Dong et al., 2019a; Song et al., 2019; Lewis et al., 2019; Rothe et al., 2020; Raffel et al., 2019; Zhang et al., 2019).

A

GOLD: Australia has expelled an Israeli diplomat saying Israel was behind the forging of Australian passports linked to the murder of a Hamas operative in Dubai.

PEGASUS: Australia has expelled an Israeli diplomat after concluding that forged Australian passports used in the killing of a Hamas militant in Dubai were issued by Israel.

Our PEGFAME model: The Australian government has expelled an Israeli diplomat over the use of forged Australian passports in the killing of a Hamas militant in Dubai.

B

PEGASUS with Top-k Sampling

Israel has summoned the Australian ambassador to complain after the Australian government said forged passports used in the killing of a Hamas operative in Dubai belonged to Netanyahu's foreign ministry.

The Australian government has ordered Israel to withdraw an officer over the use of forged Australian passports used by the 2013 murder of a Lebanese opposition figure in Dubai.

PEGASUS with Nucleus Sampling

Israel hasracuse withdrawn an envoy after the Australian government said it concluded that Israeli agents used forged passports used to kill a Dubai Bendigo businessman.

The Australian government has recalled an Israeli diplomat over accusation that fake Australian passports used 436 kilometres (300 miles) from Canberra in the death of a Hamas militant were stolen by Israeli agents.

C

Our PEGFAME model with novel Focus Sampling

Australia has expelled an Israeli diplomatic staff after accusing the country's security agency, the Israeli military's intelligence agency, of being responsible for the use of Australian visas used in the killing of a Palestinian.

The Australian government has expelled an Israeli diplomatic staff after it said the country was responsible for the use of Australian visas used in the killing of a Palestinian in the Middle East.

Figure 1: Block A shows the best predictions from PEGASUS and our PEGFAME (PEGASUS with FAME) model, along with the GOLD summary for an XSUM article. Block B presents diverse summaries generated from PEGASUS using top-k and nucleus sampling. Block C shows diverse summaries generated using our PEGFAME model with Focus sampling. The text in orange is not supported by the input article.

However, in terms of summary quality, many challenges remain. For example, generating summaries that are faithful to the input is an unsolved problem (Kryscinski et al., 2020; Maynez et al., 2020; Gabriel et al., 2020). Furthermore, there can be multiple equally good summaries per source docu-

[*]Work done when authors were interning/working at Google.

*Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics
and the 11th International Joint Conference on Natural Language Processing*, pages 6078–6095
August 1–6, 2021. ©2021 Association for Computational Linguistics

ment. Neural generation models fail to account for this and tend to generate outputs with low diversity due to standard likelihood training, approximate decoding objectives, and lack of high quality multi-reference datasets (Fan et al., 2018; Kulikov et al., 2019; Freitag et al., 2020; Choi et al., 2020). Not much attention has been given to generation of diverse, yet faithful summaries – two goals are often challenging to achieve simultaneously (Hashimoto et al., 2019); a model can produce diverse outputs through sampling (Fan et al., 2018; Holtzman et al., 2020), but at the cost of quality.

In this paper we introduce a Focus Attention MEchanism (or FAME) to transformer-based seq2seq architectures. FAME is inspired by how humans write summaries. Specifically, FAME aims to perform source-side planning to focus the summary on supported and topical content. FAME achieves this through a novel technique which augments standard contextual representations with a dynamic source-conditioned vocabulary biasing layer. We present the following experimental findings:

FAME promotes summaries faithful to the source When evaluated on the BBC extreme summarization task (XSUM; Narayan et al., 2018), experiments with two state-of-the-art summarizers – ROBERTAS2S (Rothe et al., 2020) and PEGASUS (Zhang et al., 2019) – show that both models generate summaries that are more faithful to their input documents when augmented with FAME, in comparison with their vanilla counterparts.[1] Faithfulness is measured through a variety of previously proposed metrics. In addition, we leverage the manually annotated document-summary pairs for faithfulness from Maynez et al. (2020) and train a scorer which serves as an efficient proxy for expensive human evaluations. We call this metric *BERTFaithful*.

FAME enables diverse summaries FAME, by design, supports *Focus Sampling* – a technique that is more effective in sampling topically relevant tokens to generate diverse, yet topically consistent and faithful outputs, than other sampling methods (Fan et al., 2018; Holtzman et al., 2020). Figure 1 illustrates how focus sampling generates better summaries than other sampling methods. We demonstrate the effectiveness of our new Focus Sampling technique using a variety of existing diversity and faithfulness measures. Empirically, we find that optimizing for high diversity often comes at the cost of faithfulness. Thus FAME provides a mechanism for trading-off high faithfulness with better diversity in summarization.

2 Related Work

Task-Specific Architectural Priors Several works enhance seq2seq architectures with task-specific priors. Pointer-generator style models (See et al., 2017; Xu et al., 2020) can accurately generate mostly extractive summaries by copying words from the source text via pointing. Text editing models (Malmi et al., 2019; Dong et al., 2019b; Mallinson et al., 2020) cast text generation as a sequence tagging problem with carefully selected edit operations required for the task. Others focus on improving content selection to better constrain the model to likely input phrases (Gehrmann et al., 2018) or by improving the representation of relevant input tokens (Zhou et al., 2017). Instead of directly modeling such priors, FAME learns the theme of the document through dynamic vocabulary biasing. Thus, FAME can be seen as a generalization of Pointer-generator or text-editing models via soft vocabulary learning. In fact, our FAME models achieve state-of-the-art on text-editing tasks (Appendix C).

Topic-Aware Generation Models The idea of capturing document-level semantic information has been widely explored in the summarization community. Barzilay and Elhadad (1997) use WordNet (Fellbaum, 1998) to model a text's content relative to a topic based on lexical chains. Lin and Hovy (2000) propose to learn topic signatures for summarizing documents. Recently, document-level topic information has been used for improving neural language models (Mikolov and Zweig, 2012; Ghosh et al., 2016; Dieng et al., 2017; Karmaker Santu et al., 2019), neural response generators (Xing et al., 2017; Dziri et al., 2019), and not surprisingly, neural summarizers (Narayan et al., 2018; Ailem et al., 2019; Wang et al., 2020c). Both, Narayan et al. (2018) and Ailem et al. (2019), use a pretrained Latent Dirichlet Allocation (LDA; Blei et al., 2003) model, whereas, Wang et al. (2020c) use Poisson factor analysis (Zhou et al., 2012), to synthesize topic vectors for the input. Instead, we dynamically learn a target-induced topic distribution for the input under the assumption that the human-written

[1] In the paper we focus on assessing FAME on XSUM. But other summarization and text editing results can be found in Appendix B and C.

summary is a good proxy for the input document.

Faithful Generation Models Cao et al. (2017) force faithful generation by conditioning on both source text and extracted fact descriptions from the source text. Song et al. (2020) propose to jointly generate a sentence and its syntactic dependency parse to induce grammaticality and faithfulness. Tian et al. (2019) learn a confidence score to ensure that the model attends to the source whenever necessary. Wang et al. (2020d) introduce new input-output matching and embedding similarity losses to alleviate hallucination issues. Yet, the task of generating text that is consistent with the input remains an open problem (Gabriel et al., 2020).

Diverse Generation Models There has been a surge of interest in making language models generate more diverse and human-like outputs. Vijayakumar et al. (2018) and Kulikov et al. (2019) diversify beam search, using a task-specific scoring function, or constrain beam hypotheses to be sufficiently different. Others avoid text degeneration by truncating the unreliable tail of the probability distribution at each decoding step, either by sampling from the top-k tokens (*Top-k Sampling*; Fan et al., 2018) or by sampling from a dynamic nucleus of tokens with the bulk of the probability mass (*Nucleus Sampling*; Holtzman et al., 2020). Others modify the training objective to make the distribution sparse (Martins et al., 2020) or assign lower probability to unlikely generations (Welleck et al., 2019a).

For conditional text generation, most work focuses on generating diverse questions (Narayan et al., 2016; Dong et al., 2017; Sultan et al., 2020; Wang et al., 2020b) or paraphrases (Li et al., 2016b; Dai et al., 2017; Xu et al., 2018; Cao and Wan, 2020). Following Gehrmann et al. (2018), Cho et al. (2019) use a mixture of experts to sample different binary masks on the source sequence for diverse content selection for summarization.
Our focus sampling is similar to top-k and nucleus sampling methods; in that it truncates the tail of the probability distribution. However, instead of truncating it at each decoding step, it biases the decoder proactively to generate output from a set of tokens which are topically-relevant to the input.

3 Summarization with Focus Attention

Given an input document $X_{1:n}$, we aim to generate its summary $Y_{1:m}$, where n and m are input and output sequence lengths. We address this problem using seq2seq architectures with Transformer encoder and decoder, augmented with FAME, as depicted in Figure 2. FAME learns a distribution $\boldsymbol{t}_{x_i}$ for each input token x_i over the vocabulary, measuring similarity of x_i (in context) to the tokens in the vocabulary. The vocabulary distributions, $\boldsymbol{t}_{x_i}$, for all x_i are combined to form a dynamic vocabulary bias that is added to the decoder logits. This mechanism enhances the conditioning on the input source and encourages the decoder to generate tokens that are topically similar to the input.

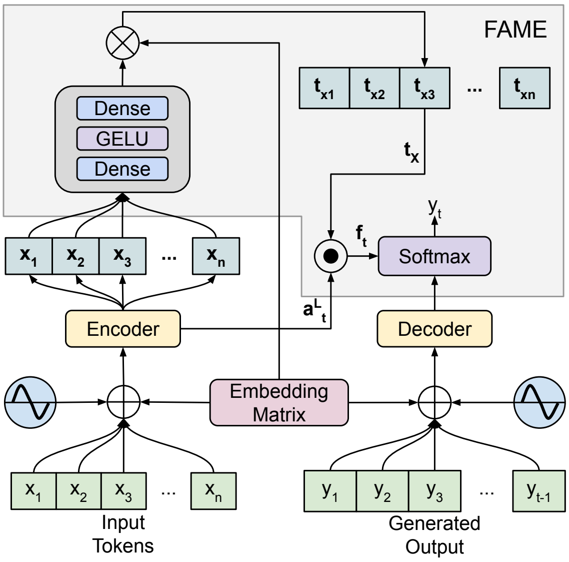

Figure 2: A Transformer-based encoder-decoder architecture with FAME.

Transformer-based seq2seq Model The encoder uses BERT Transformer layers with multi-headed self-attention to encode X to a vector sequence $\boldsymbol{X} = \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, with $\boldsymbol{x}_i \in \mathbb{R}^h$, where h is the size of hidden representation. The decoder uses an identical architecture, except that at decoding step t, layer l adds a conditional representation $\boldsymbol{y}_t^l \in \mathbb{R}^h$ for the token y_t by attending to the output representation $\boldsymbol{Y}_{1:t-1}^{l-1} = \boldsymbol{y}_1^{l-1}, \ldots, \boldsymbol{y}_{t-1}^{l-1}$ generated so far through self-attention and by attending to the input contextual representation $\boldsymbol{X}$ through encoder-decoder attention. The probability of predicting the next token y_t from a vocabulary V is:

$$p(y_t|Y_{1:t-1}, X; \theta) = \mathrm{softmax}(\boldsymbol{E}\boldsymbol{y}_t^L), \quad (1)$$

where, $\boldsymbol{y}_t^L$ is the representation from the final decoder layer L, $\boldsymbol{E} \in \mathbb{R}^{|V| \times h}$ the embedding matrix and θ the model parameters. Parameters are trained by minimizing cross-entropy at each decoding step:

$$L_{\text{MLE}}(\theta) = -\frac{1}{m}\sum_{i=1}^{m} \log p(\hat{y}_t|\hat{Y}_{1:t-1}, X; \theta),$$

where, $\hat{Y}_{1:m}$ is the human-written summary.

Focus Attention MEchansim (FAME) It is challenging for a decoder to obtain all relevant information from the conditional representation $\boldsymbol{y}_t^L$ to learn the vocabulary output logits such that predictions y_t are consistent with the input. Other modeling factors, specifically the decoder language model, can overwhelm model predictions. FAME (Figure 2) addresses this by introducing a short-circuit from the source to the vocabulary output logits via a source-conditioned bias on vocabulary items.

We take the encoder representation $\boldsymbol{X} = \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ and learn a *Token-level Vocabulary Distribution* $\boldsymbol{t}_{x_i} = \text{gelu}(\boldsymbol{x}_i \boldsymbol{W}_1)\boldsymbol{W}_2\boldsymbol{E} \in \mathbb{R}^{|V|}$, for each token x_i in the input sequence $\boldsymbol{X}$. $\boldsymbol{t}_{x_i}$ measures the contextual similarity of the input token x_i to the tokens in the vocabulary; $\boldsymbol{W}_1 \in \mathbb{R}^{h \times h'}$ and $\boldsymbol{W}_2 \in \mathbb{R}^{h' \times h}$ are parameters of newly introduced dense layers, h' is the intermediate filter size. We define a *Source-conditioned Vocabulary Distribution* as $\boldsymbol{t}_X = 1/n \sum_{i=1}^{n} \boldsymbol{t}_{x_i} \in \mathbb{R}^{|V|}$ as an average of token-level vocabulary distributions for tokens present in the input sequence X, capturing the similarity of X to the tokens in the vocabulary.

Let $\boldsymbol{a}_t^L \in \mathbb{R}^n$ be the encoder-decoder attention distribution over the source tokens for the output token y_t and the final decoder layer L. We use $\boldsymbol{a}_t^L$ to produce a weighted sum of the token-level vocabulary distributions to compute a dynamic vocabulary bias, or *Focus Bias* $\boldsymbol{f}_t = \sum_{i=1}^{n} \boldsymbol{a}_{t,i}^L \boldsymbol{t}_{x_i} \in \mathbb{R}^{|V|}$ at decoding step t. We modify the probability of predicting the next token y_t from a vocabulary V as:

$$p(y_t|Y_{1:t-1}, X; \theta) = \text{softmax}(\boldsymbol{y}_t^L \boldsymbol{E} + \boldsymbol{f}_t) \quad (2)$$

We call this *Focused Probability Distribution*, and it modifies the output logits dynamically to put more focus on those tokens in the vocabulary which are similar to the attended tokens in X. The focus bias introduces a human-inspired control to the model where we do not generate the output in a fully abstractive manner (as in Eq. (1)), but we proactively generate output tokens that are similar to the input tokens (as in Eq. (2)).

Summary-induced Topic Focused Distribution We aim to guide our focus bias $\boldsymbol{f}_t$ to be a better representative of the topical content relevant for the task. We achieve this by using the human-written summary $\hat{Y}$ as a proxy for the topical content of the input and impose the following prior on the source-conditioned vocabulary distribution $\boldsymbol{t}_X$:

$$L_{\text{Topic}}(\theta) = -\frac{1}{|V|}\sum_{i=1}^{|V|}([v_i \in \hat{Y}] \log(\sigma(\boldsymbol{t}_{X,i})) + [v_i \notin \hat{Y}] \log(1 - \sigma(\boldsymbol{t}_{X,i})))(3)$$

We further refine Eq. (3) by replacing $\hat{Y}$ with $\hat{Y}_c = \hat{Y} - F$, where F is a set of $|F|$ most frequent tokens in the vocabulary,[2] to improve focus on content words. Our final loss function is then

$$L = \lambda L_{\text{MLE}} + (1-\lambda) L_{\text{Topic}}, \quad (4)$$

where, λ is an hyper parameter.[3]

By enforcing $\boldsymbol{t}_X$ to be a topic distribution for the input X, we encourage the focus bias $\boldsymbol{f}_t$ to promote topically relevant tokens, and subsequently generate topically consistent outputs. Importantly, our focus bias with target-induced topic distribution is task-agnostic and less vulnerable to reference divergence issues (Dhingra et al., 2019; Maynez et al., 2020), and can learn any property embodied in the target relevant for the task. For example, depending on the task, $\boldsymbol{f}_t$ can learn to favour input tokens (e.g., for mostly extractive summaries) or new tokens (e.g., for mostly abstractive summaries). This is in sharp contrast to models that introduce task-specific priors, e.g., the pointer-generator network (See et al., 2017) that can copy words from the source text, but does not do well on extreme summarization which is highly abstractive in nature (Narayan et al., 2018).

Focus Sampling: Promoting Diversity in Faithful Generation We introduce *Focus Sampling* with FAME to construct a subset $V_k \subseteq V$ by sampling k tokens from the topic distribution $\boldsymbol{t}_X$ (Focus$_{\text{sample},k}$). Then, we modify Eq. (2) as

$$p(y_t|Y_{1:t-1}, X; \theta) = \begin{cases} \text{softmax}(\boldsymbol{y}_t^L \boldsymbol{E} + f_t)_i & \text{if } v_i \in V_k \cup F \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

For document summarization, the subset V_k will capture topically salient tokens necessary to generate a summary; F is always added to V_k to ensure

[2] which are usually articles or other function words.

[3] λ is set to 0.5 for all experiments.

that the model has access to function words. By tuning the parameters of sampling, we can enforce the model to control the faithfulness or diversity of the outputs.

Focus sampling has similarities to top-k ($\text{Div}_{\text{top},k}$; Fan et al., 2018) and nucleus sampling ($\text{Div}_{\text{nucleus}}$; Holtzman et al., 2020); in that they all aim to promote diversity. At each decoding step, the top-k sampling diversifies the generation process by sampling a token from the top k tokens in the final output distribution. Similarly, nucleus sampling samples from a dynamic nucleus of tokens containing the vast majority (with a cumulative probability p) of the probability distribution. Both top-k and nucleus sampling shorten the tail of the output distribution at each decoding step, whereas focus sampling constrains the decoder to use a fixed and topically relevant vocabulary V_k. Unlike the other two techniques, $\text{Focus}_{\text{sample},k}$ can also benefit from standard beam search decoding, leading to superior generation that is not only diverse, but also consistent with the input document.

4 Experimental Setup

In this section we present our experimental setup to assess the ability of our FAME models to generate faithful summaries and to demonstrate that focus sampling is more effective in generating diverse and faithful summaries than other sampling-based decoding methods.

4.1 Extreme Summarization

We evaluate FAME models on extreme document summarization (XSUM; Narayan et al., 2018). The XSUM summaries, are extreme in that the documents are summarized into single-sentence summaries. These summaries demonstrate a high level of abstractiveness, and generating them automatically requires document-level inference, abstraction, and paraphrasing. Due to their extreme nature, XSUM summaries are ideal to evaluate FAME models' ability to capture the theme of the document.[4] We use on the original cased version consisting of 204,045/11,332/11,334 training/validation/test document-summary pairs. During training, the input documents are truncated to 512 tokens. The length of the summaries are limited to 64.

4.2 Pretrained Models with FAME

We introduce FAME to two popular seq2seq architectures: RoBERTa initialized seq2seq (ROBERTAS2S, Rothe et al., 2020) and PEGASUS (Zhang et al., 2019). We refer ROBERTAS2S models with FAME as ROBFAME and PEGASUS with FAME with PEGFAME.

We experiment with ROBERTAS2S-Large with shared encoder and decoder; it has 24 layers, a hidden size of 1024, filter size of 4096, 16 attention heads, and a vocabulary with 50K sentence pieces (Kudo and Richardson, 2018). ROBERTAS2S has around 455M parameters and ROBFAME has an additional 8M parameters.

The best-performing PEGASUS model from Zhang et al. (2019) is not directly comparable with ROBERTAS2S. It does not share the encoder and decoder, it only has 16 layers, a hidden size of 1024, filter size of 4096, 16 attention heads, with a total of 568M parameters, and it also uses a much larger vocabulary with 91K sentence pieces. Hence, we trained our own PEGASUS model. We use the same architecture as ROBERTAS2S and pretrain it on a mixture of C4 (Raffel et al., 2019) and HugeNews (Zhang et al., 2019) datasets with the original objective of generating salient GAP-sentences.

Our experiments focus on this newly trained PEGASUS model which has same number of parameters and vocabulary as ROBERTAS2S. But in contrast to ROBERTAS2S, the encoder-decoder attention in PEGASUS is pretrained. This allows us to analyse how focus attention affects pretrained (PEGASUS) vs randomly-initialized (ROBERTAS2S) encoder-decoder attentions.[5]

4.3 Evaluation Metrics

Lexical Overlap We report ROUGE F1 scores (Lin and Hovy, 2003) against reference summaries; in particular, we report on ROUGE-1 and ROUGE-2 for informativeness and ROUGE-L for fluency.[6]

Semantic Similarity We report *BERTScore* (Zhang et al., 2020) which computes the contextual similarity between a candidate and its reference summary.

[4] We further experiment with long-form story highlight generation (CNN/DM; Hermann et al., 2015) and two text editing tasks: Sentence Fusion (Geva et al., 2019) and Sentence Splitting (Botha et al., 2018). Their results can be found in Appendix B and C. Our FAME models achieve SOTA on both text-editing tasks.

[5] See Appendix A for implementation details and hyperparameter settings.

[6] We lowercased candidate and reference summaries and used pyrouge with parameters "-a -c 95 -m -n 4 -w 1.2."

Models	Lexical Overlap (w/ ref)			Sem. Sim.	Faithfulness				others		
							BERTFaithful				R1(P%)
	R1	R2	RL	BERTSc.	ent.	Feqa	%	conf.	Len.	Rep.(↓)	With doc.
ROBERTAS2S	41.45	18.79	33.90	80.6	39.1	19.8	21.5	0.216	21.2	24.2	71.1
ROBFAME	42.15	19.68	34.81	80.8	41.3	21.2	22.7	0.226	20.8	20.7	72.5
PEGASUS	44.85	22.26	37.03	81.7	43.6	24.5	27.0	0.263	21.1	6.0	73.8
PEGFAME	**45.31**	**22.75**	**37.46**	**81.9**	**44.8**	**24.8**	**27.3**	**0.269**	20.8	**5.3**	**74.3**

Table 1: Abstractive Summarization results on XSUM test set comparing FAME models with their baselines. For all our models, we use standard beam decoding with a beam size of 4 to generate the single best summary for a document. Focus sampling is not used here. See Section 4.3 for details on the evaluation metrics reported. Best number for each metric is **boldfaced**.

Faithfulness ROUGE and BERTScore do not correlate well with faithfulness of the generated summaries (Maynez et al., 2020). Human evaluation is traditionally considered as the gold standard for measuring faithfulness. But recent research has shown that even human evaluation has shortcomings (Schoch et al., 2020). Moreover, it is prohibitively expensive. This has led to the proposal of meta-evaluation metrics for various generation tasks (Durmus et al., 2020; Kryściński et al., 2019; Sellam et al., 2020; Rei et al., 2020).

We evaluate FAME models on semantic inference metrics such as textual entailment (Pasunuru and Bansal, 2018; Welleck et al., 2019b; Falke et al., 2019; Kryscinski et al., 2019) and question answering (Arumae and Liu, 2019; Wang et al., 2020a). In particular, we report the probability of a summary entailing (*ent.*) its input document (Maynez et al., 2020) and QA-based *Feqa* scores (Durmus et al., 2020). For ent. scores, we train an entailment classifier by fine-tuning a BERT-Large pretrained model (Devlin et al., 2019) on the Multi-NLI dataset (Williams et al., 2018). For Feqa, we use a fine-tuned BART (Lewis et al., 2019) language model for question generation to generate questions from the summaries, and a BERT-base model fine-tuned on SQuAD (Rajpurkar et al., 2018) to answer the generated questions with input document as context.[7]

In addition to *ent.* and *Feqa*, we train a scorer leveraging manually annotated document-summary pairs for faithfulness, as a surrogate for human evaluation and call this metric *BERTFaithful*.[8] In particular, we finetune a BERT-Base classifier on 500 manually annotated document and gold summary pairs for the XSum dataset from Maynez et al. (2020) to predict whether a summary is faithful to the input document or not.[9] We report the percentage of summaries that were faithful ($\frac{1}{N}\sum_i \mathbb{1}[p_i(\text{faithful}) > 0.5]$) and the model's confidence to generate faithful summaries ($\frac{1}{N}\sum_i p_i(\text{faithful})$); N is the total number of examples in the test set.

Diversity We report the number of times (out of n), a model is able to generate a completely new summary (*Unique*), and *Distinct-N* (Li et al., 2016a), measuring the lexical diversity in the generated summaries. Distinct-N is estimated as the number of distinct n-grams of order n divided by the total number of n-grams of the same order, in all generated summaries.

Finally, we also report the average length of summaries (*Len.*), repetition errors (*Rep.*, estimated as the percentage of summaries with at least one repetition of rare or content words), and ROUGE-1 precision against the input document (*R1, P%*), to better understand their quality.

5 Results

FAME Summaries are More Fluent, Informative and Faithful. Table 1 presents results comparing our FAME models, ROBFAME and PEGFAME, against their counterparts ROBERTAS2S

[7] We used the Feqa code available here: https://github.com/esdurmus/feqa/.

[8] A very similar scorer was used in the GEM benchmark (Gehrmann et al., 2021) to identify and extract the subset with faithful reference summaries from the XSum dataset (Narayan et al., 2018).

[9] Out of 500, 90% of the document-summary pairs were used for training and the rest 50 document-summary pairs were used for validation. We used the validation set to estimate Spearman's correlation coefficients of different metrics with the human assessment for faithfulness. We found that both entailment scores (*ent.*) and *BERTFaithful* are moderately correlated with faithfulness with correlation coefficients of 0.4387 and 0.3889, respectively. As such, we believe that BERTFaithful works as an efficient proxy for expensive human evaluation for faithfulness for XSum summaries. More work is needed to understand if BERTFaithful generalizes to other datasets.

Metrics	Unique	Dist.-N			ROUGE			ent.	BERTSc.
		1	2	3	R1	R2	RL		
RoBERTaS2S ($\text{Div}_{\text{top},k}$)	9.98	2.5	25.0	57.7	33.6	12.0	26.5	21.8	76.9
RoBERTaS2S ($\text{Div}_{\text{nucleus}}$)	9.99	**4.1**	30.1	62.2	32.4	11.4	25.6	19.7	75.7
RobFame ($\text{Div}_{\text{top},k}$)	9.99	2.3	25.0	58.1	32.7	11.3	25.7	20.3	76.6
RobFame ($\text{Div}_{\text{nucleus}}$)	9.99	**4.1**	**30.7**	**63.2**	31.3	10.6	24.7	18.0	75.4
RobFame ($\text{Focus}_{\text{sample},k}$)	1.61	3.5	22.4	43.9	**38.0**	**15.7**	**31.0**	**34.3**	**78.6**
RobFame ($\text{Focus}_{\text{sample},k}$, $\text{Div}_{\text{top},k}$)	9.99	2.1	20.3	51.8	31.8	10.2	24.7	24.3	75.4
RobFame ($\text{Focus}_{\text{sample},k}$, $\text{Div}_{\text{nucleus}}$)	9.98	1.9	18.4	48.2	32.9	11.1	25.8	25.9	76.1
Pegasus ($\text{Div}_{\text{top},k}$)	9.98	1.9	23.2	55.3	36.6	14.3	28.8	27.7	78.4
Pegasus ($\text{Div}_{\text{nucleus}}$)	9.99	**3.8**	**30.5**	**63.1**	34.1	12.8	26.9	22.7	76.5
PegFame ($\text{Div}_{\text{top},k}$)	9.98	1.9	23.2	55.5	36.7	14.5	29.0	28.5	**78.5**
PegFame ($\text{Div}_{\text{nucleus}}$)	9.99	**3.8**	30.4	**63.1**	34.2	12.8	27.0	23.2	76.6
PegFame ($\text{Focus}_{\text{sample},k}$)	2.77	2.4	16.5	34.2	**37.5**	**15.4**	**30.3**	**33.6**	77.9
PegFame ($\text{Focus}_{\text{sample},k}$, $\text{Div}_{\text{top},k}$)	8.99	2.8	23.0	54.7	31.5	10.3	24.4	22.8	74.7
PegFame ($\text{Focus}_{\text{sample},k}$, $\text{Div}_{\text{nucleus}}$)	9.98	2.6	20.8	50.9	32.5	11.0	25.3	24.8	75.3

Table 2: Assessment of diversity, relevance and faithfulness with focus sampling on the XSum test set.

and Pegasus, respectively. Both Fame models clearly outperform their vanilla counterparts in terms of generating summaries that are more fluent (see RL and Rep.), more informative (see R1, R2 and BERTSc.) and more faithful (see ent., Feqa and BERTFaithful). Among all four models, PegFame summaries are most fluent, informative and faithful.

We further did pairwise comparisons for all measures in Table 1 and found that all differences are statistically significant except for BERTScore and faithfulness measures between Pegasus and PegFame.[10] These assessments demonstrate that Fame models aid both RoBERTaS2S and Pegasus in generating fluent, faithful and relevant summaries, but are more effective in RoBERTaS2S than in Pegasus for extreme summarization.

Generating Diverse and Faithful Summaries with Focus Sampling. Table 2 presents results assessing focus sampling ($\text{Focus}_{\text{sample},k}$), top-$k$ sampling ($\text{Div}_{\text{top},k}$) and nucleus sampling ($\text{Div}_{\text{nucleus}}$), for their abilities to generate diverse and faithful summaries. For $\text{Focus}_{\text{sample},k}$, we choose $k = 10,000$. We follow Holtzman et al. (2020) and choose $k = 640$ and the nucleus probability $p = 0.95$, for $\text{Div}_{\text{top},k}$ and $\text{Div}_{\text{nucleus}}$, respectively. For $\text{Focus}_{\text{sample},k}$, we decode with a beam size of 4. We also report $\text{Focus}_{\text{sample},k}$ with $\text{Div}_{\text{top},k}$ and $\text{Div}_{\text{nucleus}}$ to assess if they can benefit one-another. In each setting we sample 10 summaries for each input document. For all metrics, we report the average over all 10 samples.[11]

Both $\text{Div}_{\text{top},k}$ and $\text{Div}_{\text{nucleus}}$ almost always generate a new summary. In comparison $\text{Focus}_{\text{sample},k}$ generates 1.61 and 2.77 unique summaries using RobFame and PegFame models, respectively. $\text{Div}_{\text{nucleus}}$ tends to generate the most distinct unigrams, bigrams, and trigrams. Interestingly, $\text{Focus}_{\text{sample},k}$ summaries have a more diverse collection of unigrams than in $\text{Div}_{\text{top},k}$ summaries (3.5% vs 2.3% for RobFame and 2.4% vs 1.9% for PegFame).

The high diversity in $\text{Div}_{\text{top},k}$ and $\text{Div}_{\text{nucleus}}$ comes at the cost of faithfulness; summaries generated with these sampling techniques have poor entailment scores. $\text{Focus}_{\text{sample},k}$, on the other hand, generates summaries which entail documents the most. It also has the highest ROUGE scores across the board. Some of the generated examples can be seen in Figure 1. More predictions from other models can be found in Appendix E. Augmenting $\text{Div}_{\text{top},k}$ and $\text{Div}_{\text{nucleus}}$ with $\text{Focus}_{\text{sample},k}$ is not desirable because, though it increases diversity in terms of uniqueness and Distinct-3 scores, faithfulness suffers again.

Comparing results in Table 2 to the results in Table 1, it is clear that diversity comes at the cost of quality (e.g., RL/ent. scores for RobFame and RobFame-$\text{Focus}_{\text{sample},k}$ are 34.81/41.3 and 31.0/34.3, respectively). However, $\text{Focus}_{\text{sample},k}$ is superior to both $\text{Div}_{\text{top},k}$ and $\text{Div}_{\text{nucleus}}$ in gen-

[10] All significance tests in this work are pairwise comparisons (one-way ANOVA with posthoc Tukey HSD tests; $p < 0.01$).

[11] Feqa and BERTFaithful scores are dropped due to time constraints.

erating better quality summaries.

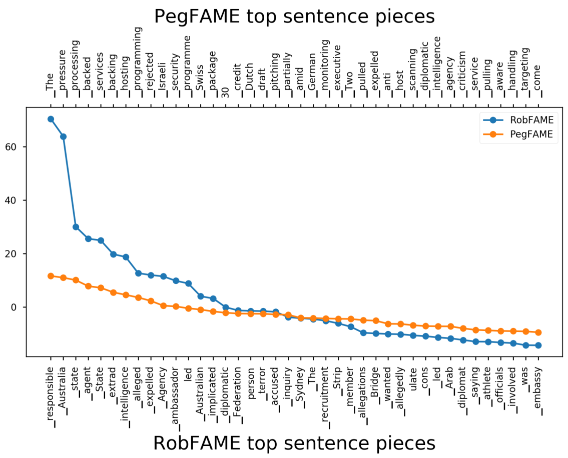

Figure 3: Top 40 sentence pieces and their logits from topic distribution t_X in RobFame and PegFame for the XSum article discussed in Figure 1.

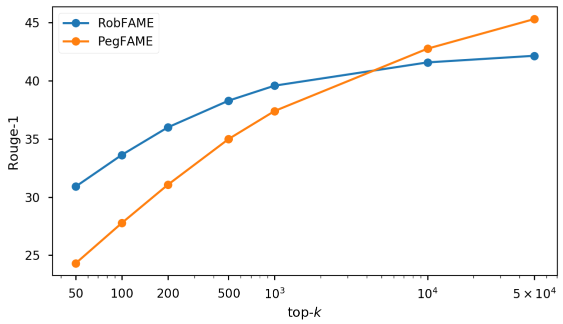

Figure 4: ROUGE-1 F1 scores of RobFame and PegFame models with different top-k vocabularies (Eq. (5)) on the XSum test set. Similar patters are observed for ROUGE-2 and ROUGE-L scores.

Focus Attention and Sampling Work Differently in RobFame and PegFame. Since both encoder-decoder and focus attention parameters of RobFame are randomly initialized, they learn to compliment each other and learn a peaky topic distribution. On the other hand, since PegFame's encoder-decoder attention is pre-trained, there is a push-pull effect between it and focus attention. This results in a smoother topic distribution, as seen in Figure 3.[12]

Although we see that both models' token sets capture the target intent well, the peaky distribution of RobFame enables more accurate predictions than that of PegFame, in a controlled generation setting. A comparison is presented in Figure 4 where we show how ROUGE-1 scores vary when we use only top-k tokens from t_X for generation.[13] We observe that RobFame consistently outperforms PegFame with the lower values of $k \in \{50, 100, 200, 500, 1000\}$.

Further, we observe that RobFame generates fewer unique summaries (1.61 vs 2.77) but has higher Distinct-N scores (3.5/22.4/43.9 vs 2.4/16.5/34.2) than PegFame, with $\text{Focus}_{\text{sample},k}$ in Table 2. This can be again be attributed to how Fame works differently in RobFame and PegFame. When V_k is sampled from RobFame's peaky distribution, the beam search decoding often tends to generate similar summaries (leading to a lower Uniqueness score) as the sampled V_ks do not diverge by much from each other. But when it does diverge, the decoder tends to generate completely new summaries (leading to higher Distinct-N scores).

Currently, we set $k = 10,000$ for our focus sampling experiments following our observations in Figure 4. Future work will focus on how to better leverage trade-off between diversity and faithfulness by controlling the peakiness of the topic distribution t_X.

Models	R1	R2	RL
Lead	16.30	1.61	11.95
PtGen (See et al., 2017)	29.70	9.21	23.24
ConvS2S (Narayan et al., 2018)	31.89	11.54	25.75
MMN (Kim et al., 2019)	32.00	12.10	26.00
MASS (Song et al., 2019)	39.75	17.24	31.95
BART (Lewis et al., 2019)	45.14	22.27	37.25
Pegasus (Zhang et al., 2019)	**47.21**	**24.56**	**39.25**
RobertaS2S (Rothe et al., 2020)	41.45	18.79	33.90
RobFame (w/o Eq. (3))	41.27	18.86	33.90
RobFame	42.15	19.68	34.81
Oracle	**72.22**	**42.22**	**53.89**
Pegasus (ours)	44.85	22.26	37.03
PegFame (w/o Eq. (3))	44.54	22.00	36.83
PegFame	45.31	22.75	37.46
Oracle	**82.39**	**60.61**	**69.19**

Table 3: Ablations and SOTA comparisons on XSum dataset. The **underlined bold** results are from the best performing models from literature and the **bold** results are the best performing Fame models.

Ablations and SOTA Comparisons We emphasize that Fame or focus sampling does not aim to improve on state-of-the-results in terms of ROUGE, but to generate more faithful or diverse summaries

[12] This difference in topic distributions is consistent across the whole test set. We compute the peakiness score of a topic distribution as the slope of the line connecting logits of the top-1st token to the top-100th token. The average peakiness scores across the XSum testset for RobFame and PegFame are 1.25 (51°) and 0.45 (24.3°), respectively.

[13] Additional results and model predictions for these experiments can be found in Appendix D.

while maintaining their quality. For completeness, we compare our ROBFAME and PEGFAME models to their ablations and other state-of-the-art models on XSUM in Table 3.

We report ROUGE scores for FAME in the ideal scenario (ORACLE) where it focuses on all the correct tokens in the input, i.e., the topic distribution $\boldsymbol{t}_X$ is identical to the distribution observed in the reference summary. These models generate summaries with very high ROUGE scores when the model is given the correct tokens to focus on. The gap between the ORACLE and FAME scores suggests that there is still a lot of work to be done in this space. Focus attention without any topical supervision (models w/o Eq. (3)) is not significantly better than the baselines. But ROBFAME and PEGFAME (trained with joint supervision in Eq. (4)) significantly outperform ROBERTAS2S and PEGASUS, respectively.

Our best model PEGFAME performs better than PtGen (See et al., 2017), ConvS2S (Narayan et al., 2018), MMN (Kim et al., 2019), MASS (Song et al., 2019) and BART (Lewis et al., 2019), but worse when the original PEGASUS (Zhang et al., 2019). This can be expected as the number of parameters in PEGFAME is far less than that in the original PEGASUS.

6 Conclusion

We introduced FAME, a new attention mechanism which dynamically biases the decoder to proactively generate tokens that are topically similar to the input. FAME enhances the faithfulness of existing state-of-the-art abstract summarization models while improving their overall ROUGE scores. Finally, our newly introduced focus sampling technique is a better alternative to top-k or nucleus sampling to generate diverse set of faithful summaries.

Acknowledgements

We thank Sebastian Gehrmann, Slav Petrov, the reviewers, and the action editor for their invaluable feedback.

Ethical Considerations

The nature of text generation leads to multiple ethical considerations when applied to applications. The main failure mode is that the model can learn to mimic target properties in the training data that are not desirable.

Faithfulness and Factuality Since models create new text, there is the danger that they may neither be faithful to the source material nor factual. This can be exacerbated when the data itself has highly abstractive targets, which require the model to generate words not seen in the source material during training. This often leads the model to generate content inconsistent with the source material (Kryscinski et al., 2020; Maynez et al., 2020; Gabriel et al., 2020).

Trustworthy Data If the data itself is not trustworthy (comes from suspect or malicious sources) the model itself will naturally become untrustworthy as it will ultimately learn the language and topics of the training data. For instance, if the training data is about Obama birther conspiracies, and the model is asked to generate information about the early life of Obama, there is a risk that such false claims will be predicted by the model.

Bias in Data Similarly, biases in the data around gender, race, etc., risk being propagated in the model predictions, which is common for most NLP tasks. This is especially true when the models are trained from non-contemporary data that do not represent current norms and practices (Blodgett et al., 2020).

The above considerations are non-malicious, in that the model is merely learning to behave as its underlying source material. If users of such models are not aware of these issues and do not account for them, e.g., with better data selection, evaluation, etc., then the generated text can be damaging.

Generation models can also be misused in malicious ways. These include generating fake news, spam, and other text meant to mislead large parts of the general population.

References

Melissa Ailem, Bowen Zhang, and Fei Sha. 2019. Topic augmented generator for abstractive summarization. *CoRR*, abs/1908.07026.

Kristjan Arumae and Fei Liu. 2019. Guiding extractive summarization with question-answering rewards. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2566–2577, Minneapolis, Minnesota.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Regina Barzilay and Michael Elhadad. 1997. Using lexical chains for text summarization. In *Intelligent Scalable Text Summarization*.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Jan A. Botha, Manaal Faruqui, John Alex, Jason Baldridge, and Dipanjan Das. 2018. Learning to split and rephrase from Wikipedia edit history. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 732–737, Brussels, Belgium. Association for Computational Linguistics.

Yue Cao and Xiaojun Wan. 2020. DivGAN: Towards diverse paraphrase generation via diversified generative adversarial network. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2411–2421, Online. Association for Computational Linguistics.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2017. Faithful to the original: Fact aware neural abstractive summarization.

Jaemin Cho, Minjoon Seo, and Hannaneh Hajishirzi. 2019. Mixture content selection for diverse sequence generation. 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 3121–3131, Hong Kong, China. Association for Computational Linguistics.

Byung-Ju Choi, Jimin Hong, David Park, and Sang Wan Lee. 2020. F^2-softmax: Diversifying neural text generation via frequency factorized softmax. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9167–9182, Online. Association for Computational Linguistics.

Bo Dai, Dahua Lin, Raquel Urtasun, and Sanja Fidler. 2017. Towards diverse and natural image descriptions via a conditional GAN. *CoRR*, abs/1703.06029.

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*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4884–4895, Florence, Italy.

Adji B. Dieng, Chong Wang, Jianfeng Gao, and John Paisley. 2017. TopicRNN: A recurrent neural network with long-range semantic dependency. In *Proceedings of the 5th International Conference on Learning Representations*, Toulon, France.

Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. 2017. Learning to paraphrase for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 875–886, Copenhagen, Denmark. Association for Computational Linguistics.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019a. Unified language model pre-training for natural language understanding and generation. In H. Wallach, H. Larochelle, A. Beygelzimer, F. Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 13042–13054. Curran Associates, Inc.

Yue Dong, Zichao Li, Mehdi Rezagholizadeh, and Jackie Chi Kit Cheung. 2019b. EditNTS: An neural programmer-interpreter model for sentence simplification through explicit editing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3393–3402, Florence, Italy. Association for Computational Linguistics.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar Zaiane. 2019. Augmenting neural response generation with context-aware topical attention. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 18–31, Florence, Italy. Association for Computational Linguistics.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy.

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

Christiane Fellbaum, editor. 1998. *WordNet: an electronic lexical database*. MIT Press.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics.

Saadia Gabriel, Asli Celikyilmaz, Rahul Jha, Yejin Choi, and Jianfeng Gao. 2020. Go figure! a meta evaluation of factuality in summarization.

Sebastian Gehrmann, Tosin P. Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Aremu Anuoluwapo, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh D. Dhole, Wanyu Du, Esin Durmus, Ondrej Dusek, Chris Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Rubungo Andre Niyongabo, Salomey Osei, Ankur P. Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. *CoRR*, abs/2102.01672.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium. Association for Computational Linguistics.

Mor Geva, Eric Malmi, Idan Szpektor, and Jonathan Berant. 2019. DiscoFuse: A large-scale dataset for discourse-based sentence fusion. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3443–3455, Minneapolis, Minnesota. Association for Computational Linguistics.

Shalini Ghosh, Oriol Vinyals, Brian Strope, Scott Roy, Tom Dean, and Larry Heck. 2016. Contextual LSTM (CLSTM) models for large scale NLP tasks. *CoRR*, abs/1602.06291.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1640, Berlin, Germany. Association for Computational Linguistics.

Tatsunori Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. 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 1689–1701, Minneapolis, Minnesota. Association for Computational Linguistics.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1693–1701. Curran Associates, Inc.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Shubhra Kanti Karmaker Santu, Kalyan Veeramachaneni, and Chengxiang Zhai. 2019. TILM: Neural language models with evolving topical influence. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 778–788, Hong Kong, China. Association for Computational Linguistics.

Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. 2019. Abstractive summarization of Reddit posts with multi-level memory networks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2519–2531, Minneapolis, Minnesota. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization. *CoRR*, abs/1910.12840.

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, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Ilia Kulikov, Alexander Miller, Kyunghyun Cho, and Jason Weston. 2019. Importance of search and evaluation strategies in neural dialogue modeling. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 76–87, Tokyo, Japan. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016b. A simple, fast diverse decoding algorithm for neural generation. *CoRR*, abs/1611.08562.

Chin-Yew Lin and Eduard Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*.

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

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

Jonathan Mallinson, Aliaksei Severyn, Eric Malmi, and Guillermo Garrido. 2020. Felix: Flexible text editing through tagging and insertion.

Eric Malmi, Sebastian Krause, Sascha Rothe, Daniil Mirylenka, and Aliaksei Severyn. 2019. Encode, tag, realize: High-precision text editing. 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 5054–5065, Hong Kong, China. Association for Computational Linguistics.

Inderjeet Mani. 2001. *Automatic summarization*, volume 3. John Benjamins Publishing.

Pedro Henrique Martins, Zita Marinho, and André F. T. Martins. 2020. Sparse text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4252–4273, Online. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Tomas Mikolov and Geoffrey Zweig. 2012. Context dependent recurrent neural network language model. In *Proceedings of the Spoken Language Technology Workshop*, pages 234–239. IEEE.

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

Shashi Narayan, Siva Reddy, and Shay B. Cohen. 2016. Paraphrase generation from latent-variable PCFGs for semantic parsing. In *Proceedings of the 9th International Natural Language Generation conference*, pages 153–162, Edinburgh, UK. Association for Computational Linguistics.

Ani Nenkova and Kathleen McKeown. 2011. Automatic summarization. *Foundations and Trends in Information Retrieval*, 5(2–3):103–233.

Ramakanth Pasunuru and Mohit Bansal. 2018. Multireward reinforced summarization with saliency and entailment. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 646–653. Association for Computational Linguistics.

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

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

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. *CoRR*, abs/1901.07291.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

Stephanie Schoch, Diyi Yang, and Yangfeng Ji. 2020. "this is a problem, don't you agree?" framing and bias in human evaluation for natural language generation. In *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*, pages 10–16, Online (Dublin, Ireland). Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 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.

Kaiqiang Song, Logan Lebanoff, Q. Guo, Xipeng Qiu, X. Xue, Chen Li, Dong Yu, and Fei Liu. 2020. Joint parsing and generation for abstractive summarization. In *AAAI*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: Masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 5926–5936. PMLR.

Md Arafat Sultan, Shubham Chandel, Ramón Fernandez Astudillo, and Vittorio Castelli. 2020. On the importance of diversity in question generation for QA. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5651–5656, Online. Association for Computational Linguistics.

Ran Tian, Shashi Narayan, Thibault Sellam, and Ankur P. Parikh. 2019. Sticking to the facts: Confident decoding for faithful data-to-text generation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

Ashwin K. Vijayakumar, Michael Cogswell, Ramprasaath R. Selvaraju, Qing Sun, Stefan Lee, David J. Crandall, and Dhruv Batra. 2018. Diverse beam search for improved description of complex scenes. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 7371–7379. AAAI Press.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020a. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Zhen Wang, Siwei Rao, Jie Zhang, Zhen Qin, Guangjian Tian, and Jun Wang. 2020b. Diversify question generation with continuous content selectors and question type modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2134–2143, Online. Association for Computational Linguistics.

Zhengjue Wang, Zhibin Duan, Hao Zhang, Chaojie Wang, Long Tian, Bo Chen, and Mingyuan Zhou. 2020c. Friendly topic assistant for transformer based abstractive summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 485–497, Online. Association for Computational Linguistics.

Zhenyi Wang, Xiaoyang Wang, Bang An, Dong Yu, and Changyou Chen. 2020d. Towards faithful neural table-to-text generation with content-matching constraints.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2019a. Neural text generation with unlikelihood training. *CoRR*, abs/1908.04319.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019b. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741, Florence, Italy.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies, pages 1112–1122. Association for Computational Linguistics.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 3351–3357. AAAI Press.

Jingjing Xu, Xuancheng Ren, Junyang Lin, and Xu Sun. 2018. Diversity-promoting GAN: A cross-entropy based generative adversarial network for diversified text generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3940–3949, Brussels, Belgium. Association for Computational Linguistics.

Song Xu, Haoran Li, Peng Yuan, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. Self-attention guided copy mechanism for abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1355–1362, Online. Association for Computational Linguistics.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *Proceedings of the 8th International Conference on Learning Representations*, Virtual Conference, Formerly Addis Ababa Ethiopia.

Mingyuan Zhou, Lauren Hannah, David Dunson, and Lawrence Carin. 2012. Beta-negative binomial process and poisson factor analysis. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 1462–1471, La Palma, Canary Islands. PMLR.

Qingyu Zhou, Nan Yang, Furu Wei, and Ming Zhou. 2017. Selective encoding for abstractive sentence summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1095–1104, Vancouver, Canada. Association for Computational Linguistics.

A Implementation and Reproducibility Details

Following Rothe et al. (2020), the encoder and decoder of ROBERTAS2S and ROBFAME models are initialized with public RoBERTa checkpoints. The encoder and decoder parameters are shared in both cases. Only the encoder-decoder attention parameters are initialized randomly. For ROBFAME, the focus attention parameters are also randomly initialized. We experiment with large RoBERTa checkpoints with 24 layers, a hidden size of 1024, filter size of 4096, 16 attention heads, and a vocabulary with 50K sentence pieces (Kudo and Richardson, 2018). ROBERTAS2S has around 455M parameters and ROBFAME has 463M parameters, with an additional 8M parameters. Our PEGASUS and PEGFAME implementation also have the same configuration, except for the encoder-decoder attention parameters which are pretrained.

We used Cloud TPU v3 accelerators for training. All models are fine-tuned on the target task using Adam with a learning rate of 0.05. We use a linear learning rate warm up with 40k steps, normalized by the square root of the hidden size, and a square root decay. We do not perform any tuning on these hyperparameters. We use a global batch size of 128 document-summary pairs. We adapt to different number of training steps depending on the training data sizes. Models are trained for 400k and 200k steps for CNN/DM and XSUM respectively, saving check-points every 1000 steps. We choose the best model based on ROUGE-L performance on the respective validation set.

The vocabulary for functional tokens F is constructed by taking the most frequent sentence pieces in the training set. We tune $|F|$ using the respective validation sets; for XSUM, we choose $f = 500$ frequent sentence pieces and for CNN/DM, $f = 1000$. For all our experiments with the FAME models, the beam size is set to 4.

We use Cloud TPU v3 accelerators for computing entailment scores which takes about 20 minutes for the two datasets' test sets. Question generation and answering for Feqa are run on a NVIDIA V100 GPU, and it takes between 8-12 hours for one setting of each test set.

B Abstractive Summarization Results on CNN/DailyMail

The CNN/DM dataset (Hermann et al., 2015) consists of 287,227/13,368/11,490 training/validation/test document-summary pairs. The CNN/DM summaries are in the form of bullet-point story highlights and exhibit a high degree of extraction, requiring the models to learn to copy from the source documents. The XSUM summaries, on the other hand, are extreme, in that the documents are summarized into single-sentence summaries with a high level of abstractiveness. For comparison, the XSUM summaries show a much larger percentages of novel constructions than found in CNN/DM summaries (35.8/83.5/95.5/98.5 vs 16.8/54.3/72.4/80.4 novel 1/2/3/4-grams). We use the original cased version. During training, the input documents are truncated to 512 tokens and the length of the summaries are limited to 128 tokens.

Models	CNN/DM		
	R1	R2	RL
Lead	39.60	17.70	36.20
PtGen (See et al., 2017)	39.53	17.28	36.38
Bottom-Up (Gehrmann et al., 2018)	41.22	18.68	38.34
SAGCopy (Xu et al., 2020)	42.53	19.92	39.44
MASS (Song et al., 2019)	42.12	19.50	39.01
UniLM (Dong et al., 2019a)	43.33	20.21	40.51
BART (Lewis et al., 2019)	44.16	21.28	40.90
T5 (Raffel et al., 2019)	43.52	**21.55**	40.69
PEGASUS (C4, Zhang et al., 2019)	43.90	21.20	40.76
PEGASUS (HugeNews, Zhang et al., 2019)	44.17	21.47	41.11
ProphetNet (Qi et al., 2020)	**44.20**	21.17	**41.30**
ROBERTAS2S (Rothe et al., 2020)	39.88	18.66	37.22
ROBFAME (ours)	40.27	18.43	37.51
PEGASUS (ours)	42.62	20.38	39.61
PEGFAME (ours)	**42.95**	**20.79**	**39.90**

Table 4: Abstractive summarization results on CNN/DM datasets. The **underlined bold** results are from the best performing models from literature and the **bold** results are the best performing FAME models.

Table 4 and 5 present complete results for CNN/DM dataset. We see similar kind of improvements as observed in Table 1, except for ROUGE-2 for ROBFAME which is 0.23 points worse than the ROBERTAS2S baseline. Our best model PEGFAME performs better than both copy mechanism models: LSTM-based PtGen (See et al., 2017) and Transformer-based SAGCopy (Xu et al., 2020). PEGFAME performs worse when compared with T5 (Raffel et al., 2019), the original PEGASUS (Zhang et al., 2019) and ProphetNet (Qi et al., 2020). This can be expected as the number of parameters in PEGFAME is almost half of T5 or ProphetNet, and is 100M less than that in the original PEGASUS.

ROBFAME performs worse than ROBERTAS2S on both ent. and Feqa measures for CNN/DM, similar to ROUGE-2 in Table 4. We hypothesize that this is due to the extractive nature of the CNN/DM dataset and the fact that it is not able to copy to-

Models	Len.	Rep. %	R1(P%) With doc.	doc. → sum. ent. (↑)	¬ cont.	Feqa acc.	avg.(#Q)	BERTSc.
RobertaS2S	52.1	77.6	92.7	88.8	96.4	37.3	18.1	76.0
RobFame	55.5	79.6	92.5	87.3	96.3	35.2	19.3	76.1
Pegasus	58.1	69.4	95.0	90.9	97.5	40.3	21.0	76.8
PegFame	58.5	71.0	95.3	**91.0**	**97.6**	**41.1**	21.1	**76.9**

Table 5: Faithfulness and qualitative assessment of summaries on Cnn/Dm dataset.

kens from the input to the necessary extent as the encoder-decoder attention is not pre-trained. Moreover, Feqa scores for RobertaS2S and RobFame may not be fully comparable due to variation in their summary lengths and the number of Feqa questions generated; the RobFame summaries, on average, are 3 words longer and generate 1.2 more questions than that of RobertaS2S. Nevertheless, we don't see this kind of drop in ¬cont. scores (i.e., summary not contradicting, either entailed by or neutral to the document) and BERTScores.

C Text Editing Results

We also train the Fame models on two text editing tasks: (i) for sentence fusion – the problem of combining multiple sentences into a single coherent sentence – we used the "balanced Wikipedia" portion of the DiscoFuse dataset (Geva et al., 2019), and (ii) for split-and-rephrase – the reverse task of sentence fusion – we used the WikiSplit dataset (Botha et al., 2018), which consists of 1M examples of sentence splits extracted from the Wikipedia edit history. As the name suggests, both text editing tasks require a low degree of abstraction.

For both the tasks, we train the models for 300k steps with a global batch size of 256. The input and output are padded to a length of 128, which covers 100% of the training, evaluation and test data. The vocabulary for functional tokens F is constructed by taking the top 100 and 500 sentence pieces for DiscoFuse and WikiSplit respectively.

We report corpus-level BLEU[14], the exact match accuracy, and SARI scores (Xu et al., 2016)[15]. The results can be seen in Table 6. The vanilla Pegasus model already beats the current state-of-the-art on both DiscoFuse and WikiSplit. The PegFame model performs better, albeit by a small margin, on all metrics on DiscoFuse. On WikiSplit, it has a higher exact match accuracy while maintaining the SARI score and performs 0.1 BLEU worse than Pegasus.

DiscoFuse	Exact	SARI	BLEU
(Geva et al., 2019)	51.1	84.5	–
LaserTagger (Malmi et al., 2019)	53.8	85.5	–
Felix (Mallinson et al., 2020)	61.3	88.8	–
RobertaS2S (Rothe et al., 2020)	66.6	90.3	–
Pegasus (ours)	<u>67.4</u>	<u>90.5</u>	95.8
PegFame (ours)	**67.8**	**90.7**	**95.9**
WikiSplit	**Exact**	**SARI**	**BLEU**
(Botha et al., 2018)	14.3	61.5	76.4
LaseTagger (Malmi et al., 2019)	15.2	61.7	76.3
RobertaS2S (Rothe et al., 2020)	16.4	63.8	**77.4**
Pegasus (ours)	<u>16.6</u>	**64.1**	**77.4**
PegFame (ours)	**16.8**	**64.1**	77.3

Table 6: Text editing results on Discofuse and WikiSplit. The <u>underlined</u> scores beat the current state-of-the-art and the **bold** scores are the new state-of-the-art.

D Controlled Generation with focus attention using Top-k tokens

Table 7 presents results from our controlled summary generation experiments with top-k tokens from t_X using focus attention ($\text{Focus}_{\text{top},k}$) on the XSum test set. In Figures 3 and 4, we describe how RobFame consistently outperforms PegFame at lower values of $k \in \{50, 100, 200, 500, 1000\}$ due to their peaky and smooth t_X, respectively. While Figure 4 only plots rouge-1 F1 scores, Table 7 additionally reports rouge-2, rouge-L, entailment, Feqa, and BERTScores. Figure 6 presents predictions from models using $\text{Focus}_{\text{top},k}$ for the article presented in Figures 1 and 5.

E Diverse Summarization with $\text{Div}_{\text{top},k}$, $\text{Div}_{\text{nucleus}}$ and $\text{Focus}_{\text{sample},k}$

Figures 7 show the diverse summaries generated using $\text{Focus}_{\text{sample},k}$ for the article shown in Figure 5. The predictions from $\text{Div}_{\text{top},k}$ and $\text{Div}_{\text{nucleus}}$ are omitted due to the prescribed limit on the number of pages allowed for the Appendix. Please find them on the arXiv version at `https://arxiv.org/abs/2105.11921`.

[14] We use NLTK v3.2.2 with case sensitive scoring to estimate BLEU scores.

[15] SARI is a lexical similarity metric which compares the model's output to multiple references and the input in order to assess the model's ability to add, delete, and keep an n-gram. It's implementation is available at: `https://github.com/tensorflow/tensor2tensor/blob/master/tensor2tensor/utils/sari_hook.py`.

Metrics	ROUGE			ent.	Feqa	BERTScore
	R1	R2	RL			
RobertaS2S	41.45	18.79	33.90	39.1	19.8	80.6
RobFame	**42.15**	**19.68**	**34.81**	**41.3**	**21.2**	**80.8**
RobFame (Focus$_{\text{top},k=50}$)	30.90	10.60	24.85	27.1	10.6	74.2
RobFame (Focus$_{\text{top},k=100}$)	33.62	12.39	27.14	30.3	12.4	74.2
RobFame (Focus$_{\text{top},k=200}$)	35.99	14.12	29.23	32.4	13.9	77.3
RobFame (Focus$_{\text{top},k=500}$)	38.29	16.04	31.30	35.8	15.9	78.6
RobFame (Focus$_{\text{top},k=1000}$)	39.58	17.18	32.49	37.3	17.3	79.3
RobFame (Focus$_{\text{top},k=10000}$)	41.58	19.13	34.30	40.7	20.2	80.5
Pegasus	44.85	22.26	37.03	43.6	24.5	81.7
PegFame	**45.31**	**22.75**	**37.46**	**44.8**	**24.8**	**81.9**
PegFame (Focus$_{\text{top},k=50}$)	24.30	7.52	19.32	20.8	8.0	68.8
PegFame (Focus$_{\text{top},k=100}$)	27.77	9.26	22.09	24.1	9.3	71.3
PegFame (Focus$_{\text{top},k=200}$)	31.05	11.14	24.82	27.0	10.8	73.6
PegFame (Focus$_{\text{top},k=500}$)	34.99	13.65	28.19	31.0	13.0	76.2
PegFame (Focus$_{\text{top},k=1000}$)	37.40	15.30	30.16	33.6	14.9	75.9
PegFame (Focus$_{\text{top},k=10000}$)	42.76	19.89	34.97	40.2	20.1	80.5

Table 7: Assessment of controlled summary generation with focus sampling Focus$_{\text{top},k}$ on the XSum test set. We experiment with limiting Fame models to different sizes of vocabulary V_k using the topic distribution t_X; in particular, we experiment with $k = \{50, 100, 200, 500, 1000, 10000\}$. We also report numbers for RobertaS2S, RobFame, Pegasus and PegFame, using the whole vocabulary of size 50k. The **bold** results in each block are the best performing RobertaS2S-based and Pegasus-based models.

Gold	Australia has expelled an Israeli diplomat saying Israel was behind the forging of Australian passports linked to the murder of a Hamas operative in Dubai.
Article	Australia's foreign minister said these were "not the actions of a friend". The UK took similar action in March, after concluding that Israel was responsible for the use of forged UK passports in the plot. The Israeli foreign ministry said Australia's decision was disappointing. Ministry spokesman Yigal Palmor said it was "not in line with the importance and the quality of the relationship between our countries". 'Sorrow not anger' At least four forged Australian passports were used in the killing of Mahmoud al-Mabhouh in Dubai in January. The originals belonged to Australians living in Israel. The Australian government said a police investigation had left it in no doubt that the Israeli authorities were behind "the abuse and counterfeiting of the passports". As a result Foreign Minister Stephen Smith asked Israel to withdraw a diplomat, whom he did not identify. "The decision to ask Israel to remove from Australia one of its officers at the Israeli embassy in Canberra is not something which fills the Australian government with any joy," he said. "On the contrary, the decision was made much more in sorrow than in anger." Passports from France, Ireland, Germany and Britain were used in the operation, and in March, the British government expelled an Israeli diplomat from London. The Israeli government has said there is no proof that it was behind the killing, although Dubai officials have said they are 99.9% sure that agents from Mossad were responsible.
RobertaS2S	Australia has asked Australia to withdraw an Israeli diplomat from its embassy in Canberra after an alleged plot to kill a Abu Dhabi militant in Dubai.
RobFame	Australia has asked Israel to withdraw one of its diplomats from its embassy in Canberra after it admitted it used forged passports.
Pegasus	Australia has expelled an Israeli diplomat after concluding that forged Australian passports used in the killing of a Hamas militant in Dubai were issued by Israel.
PegFame	The Australian government has expelled an Israeli diplomat over the use of forged Australian passports in the killing of a Hamas militant in Dubai.

Figure 5: A 2010 BBC article from the XSum testset, its human written summary and model predictions from RobertaS2S, and Pegasus, with and without Fame. The text in orange is not supported by the input article.

RobFame (Focus$_{\text{top},k=50}$)	Australia has said it will not be expelled an ambassador from Australia following the alleged s agent for the so-called Arab Arab State.
RobFame (Focus$_{\text{top},k=100}$)	Australia has said it will not be expelled an ambassador from Australia following the killing of a terror agent in the Arab world.
RobFame (Focus$_{\text{top},k=200}$)	Australia has said it will not be expelled an ambassador from Australia following the killing of an Australian terror suspect in the Arab world.
RobFame (Focus$_{\text{top},k=500}$)	Australia has asked Israel to end its diplomatic investigation into an alleged plot to murder an Australian terror suspect.
RobFame (Focus$_{\text{top},k=1000}$)	Australia has asked Israel to strip an ambassador from its embassy following the death of an Arab man in Dubai.
RobFame (Focus$_{\text{top},k=10000}$)	Australia has asked Israel to withdraw one of its diplomats from its embassy in Canberra following the death of a terror suspect.
PegFame (Focus$_{\text{top},k=50}$)	The Israeli government has been expelled from the country after it was found that the country's security agency, the Israeli intelligence agency, was to be to be found to have used a number of the country's out-of-country p when it was used in the Emirates car-j best.
PegFame (Focus$_{\text{top},k=100}$)	The Israeli government has been expelled from the country after it was found that the country's security agency, the Israeli intelligence agency, had used the country's visas in the Emirates terror.
PegFame (Focus$_{\text{top},k=200}$)	The Australian government has expelled an Israeli diplomats after it found that the country's security agency, the Israeli intelligence agency, had used the country's visas in the Emirates terror attack.
PegFame (Focus$_{\text{top},k=500}$)	The Australian government has expelled an Israeli diplomatic staff after accusing the country's security agency, the Israeli intelligence agency, of using a number of Australian visas in the Emirates terror attack.
PegFame (Focus$_{\text{top},k=1000}$)	Australia has expelled an Israeli diplomatic staff after accusing the country's security agency, the Israeli military's intelligence agency, of being responsible for the use of Australian visas used in the killing of a Palestinian.
PegFame (Focus$_{\text{top},k=10000}$)	Australia has expelled an Israeli diplomat over the use of forged Australian passports in the killing of a Hamas militant in Dubai.

Figure 6: Model predictions with focus sampling Focus$_{\text{top},k}$, a controlled generation setting. The text in orange is not supported by the input article. We note that with smaller values of k, both RobertaS2S-based and Pegasus-based models tend to hallucinate more often.

RobFame (Focus$_{\text{sample},k}$)
Australia has asked Israel to strip one of its diplomats from its embassy following the death of an Arab man in Dubai. Australia has asked Israel to end its diplomatic investigation into an alleged plot to murder an Australian terror suspect. Australia has asked Israel to strip one of its diplomats from its embassy in Australia over the death of a terror suspect.
PegFame (Focus$_{\text{sample},k}$)
The Australian government has expelled an Israeli diplomatic staff after accusing it of using a number of Australian visas in the killing of a Palestinian in a car bombing. The Australian government has expelled an Israeli diplomatic staff after it said the country was responsible for the use of Australian visas used in the killing of a Palestinian in a car bombing. Australia has expelled an Israeli diplomatic staff after accusing the country's security agency, the Israeli military's intelligence agency, of being responsible for the use of Australian visas used in the killing of a Palestinian. Australia has expelled an Israeli diplomatic mission after accusing the country's security agency, the Israeli military's intelligence agency, of being responsible for the use of Australian visas used in the killing of a Palestinian in the Arab city of Emirates. The Australian government has expelled an Israeli diplomatic staff after it said the country was responsible for the use of Australian visas used in the killing of a Palestinian in the Middle East.

Figure 7: Fame model predictions with Focus$_{\text{sample},k}$ ($k = 10000$). The text in orange is not supported by the input article.