RESURRECTING SUBMODULARITY FOR NEURAL TEXT GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Submodularity is desirable for a variety of objectives in content selection where the current neural encoder-decoder framework is inadequate. However, it has so far not been explored in the neural encoder-decoder system for text generation. In this work, we define diminishing attentions with submodular functions and in turn, prove the submodularity of the effective neural coverage. The greedy algorithm approximating the solution to the submodular maximization problem is not suited to attention score optimization in auto-regressive generation. Therefore instead of following how submodular function has been widely used, we propose a simplified yet principled solution. The resulting attention module offers an architecturally simple and empirically effective method to improve the coverage of neural text generation. We run experiments on three directed text generation tasks with different levels of recovering rate, across two modalities, three different neural model architectures and two training strategy variations. The results and analyses demonstrate that our method generalizes well across these settings, produces texts of good quality and outperforms state-of-the-art baselines.

1 INTRODUCTION

Monotone nondecreasing submodular objectives have been shown to be ideal for content selection and alignment in *extractive* text summarization and *statistical* machine translation, respectively (Lin & Bilmes, 2010; 2011a;b). Indeed, it can be shown that many popular extractive summarization methods (Carbonell & Goldstein, 1998; Berg-Kirkpatrick et al., 2011) optimize a submodular objective. Despite their appropriateness, submodular functions for content selection have so far been ignored in neural text generation models.

The neural encoder-decoder framework trained in an end-to-end manner maintains the state-of-the-art (SoTA) in a class of directed text generation tasks aimed at recovering the source message either to the full or a compressed version of it. A major shortcoming of such architectures in dealing with text generation is that they could keep covering some parts in the source while ignoring the other important concepts, thus resulting in less comprehensive coverage. Various mechansims for improving the neural coverage have been shown to be effective (See et al., 2017; Tu et al., 2016; Wu et al., 2016). However, they either require extra parameters and loss to furnish the model with better learning capacity or place a specific bound on the sum of attention scores.

In this work, we define a class of novel attention mechanisms called *diminishing attentions* with submodular functions and in turn, prove the submodularity of the effective neural coverage. The submodular maximization problem is generally approximated by greedy selection. However, it is not suited to optimizing attention scores in auto-regressive generation systems. We therefore put forward a simplified yet principled and empirically effective solution. By imposing *submodularity* on the coverage enforced by the decoder states on the encoder states, our diminishing attention method enhances the model's awareness of previous steps, leading to more comprehensive overall coverage of the source and maintaining a focus on the most important content when the goal is to generate a compressed version of the source (*e.g.,* text summarization). We further enhance our basic diminishing attention and propose *dynamic diminishing attention* to enable dynamically adapted coverage. Our results highlight the benefits of submodular coverage. Our diminishing attention mechanisms achieve SoTA results on three diverse directed text generation tasks, abstractive sum-

marization, neural machine translation (NMT) and image-paragraph generation spanning across two modalities, three neural architectures and two training strategy variations.

2 BACKGROUND

2.1 SUBMODULAR FUNCTIONS

Let $\mathbb{V} = \{v_1, \ldots, v_n\}$ denote a set of n objects and $f : 2^{\mathbb{V}} \rightarrow \mathbb{R}$ is a set-function that returns a real value for any subset $\mathbb{S} \subseteq \mathbb{V}$. We also assume $f(\phi) = 0$. Our goal is to find the subset:

$$\mathbb{S}^* = \arg\max_{\mathbb{S} \subseteq \mathbb{V}} f(\mathbb{S}) \qquad \text{s.t. } |\mathbb{S}^*| \leq m \tag{1}$$

where m is the budget; *e.g.,* for summarization, m is the maximum summary length allowed. Note that $f : 2^{\mathbb{V}} \rightarrow \mathbb{R}$ can also be expressed as $f : \{0, 1\}^n \rightarrow \mathbb{R}$, where a subset $\mathbb{S} \subseteq \mathbb{V}$ is represented as a one-hot vector of length n, that is, $\mathbb{S} = (\mathbb{1}(v_1 \in \mathbb{S}), \ldots, \mathbb{1}(v_n \in \mathbb{S}))$ with $\mathbb{1}$ being the indicator function that returns 1 if the argument is true otherwise 0. In general, solving Equation 1 is NP-hard. Even when f is *monotone submodular* (defined below), it is still NP-complete.

Definition 2.1. f is submodular if $f(\mathbb{S} + v) - f(\mathbb{S}) \geq f(\mathbb{T} + v) - f(\mathbb{T})$ for all $\mathbb{S} \subseteq \mathbb{T} \subseteq \mathbb{V}, v \notin \mathbb{S}$.

This property is also known as *diminishing returns*, which says that the information gain given by a candidate object (*e.g.,* a word or sentence) is larger when there are fewer objects already selected (as summary). The function f is *monotone nondecreasing* if for all $\mathbb{S} \subseteq \mathbb{T}$, $f(\mathbb{S}) \leq f(\mathbb{T})$. In this paper, we will simply refer to monotone nondecreasing submodular functions as submodular functions.

Submodular functions can be considered as the discrete analogue of concave functions in that $f(\boldsymbol{\theta}) : \mathbb{R}^n \rightarrow \mathbb{R}$ is concave if the derivative $f'(\boldsymbol{\theta})$ is non-increasing in $\boldsymbol{\theta}$, and $f(\mathbb{S}) : \{0, 1\}^n \rightarrow \mathbb{R}$ is submodular if for all i the discrete derivative, $\partial_i f(\mathbb{S}) = f(\mathbb{S} + v_i) - f(\mathbb{S})$ is non-increasing in $\mathbb{S}$. Furthermore, if $g : \mathbb{R}_+ \rightarrow \mathbb{R}$ is concave, then the composition $f^{'}(\mathbb{S}) = g(f(\mathbb{S})) : 2^{\mathbb{V}} \rightarrow \mathbb{R}$ is also submodular. The convex combination of two submodular functions is also submodular.

2.2 NEURAL COVERAGE

Neural coverage of one encoder state can be defined as the sum of the attention scores that it receives over the first until the previous decoding step (Wu et al., 2016; Tu et al., 2016). Formally, the coverage of encoder state i at decoding step t is $c_i^t = \sum_{t'=0}^{t-1} a_i^{t'}$, where $a_i^{t'}$ are the attention scores. In abstractive summarization, See et al. (2017) use coverage to keep track of what has been generated so far by assigning trainable parameters to the coverage and using it to guide the attention module in making decisions. They also introduce a coverage loss to discourage the network from repeatedly attending to the same parts, thereby avoiding repetition in the generated summary. In NMT, Wu et al. (2016) apply a coverage penalty during decoding which restricts the coverage of an input token from exceeding 1. Tu et al. (2016) maintain a coverage vector which is updated with the recurrence unit and fed into the attention model. The major differences between our method and the previous methods are that we do not require extra parameters or extra losses to furnish the network with better learning capacity, and we do not place a specific bound on the sum of attention scores. Moreover, the effective neural coverage of our method is submodular.

3 METHOD

In this section, we present *submodular coverage* and our *diminishing attentions* for the neural encoder-decoder model, and we show the *effective coverage* based on the diminishing attentions.

3.1 SUBMODULAR COVERAGE

In the general encoder-decoder framework, the input is represented as a set of latent states (concepts) from an encoder, and the decoder constructs the output autoregressively by generating one token at a time. While generating a token, the decoder computes an attention distribution over the encoded latent states, which represents the relevance of the corresponding input to the output token.

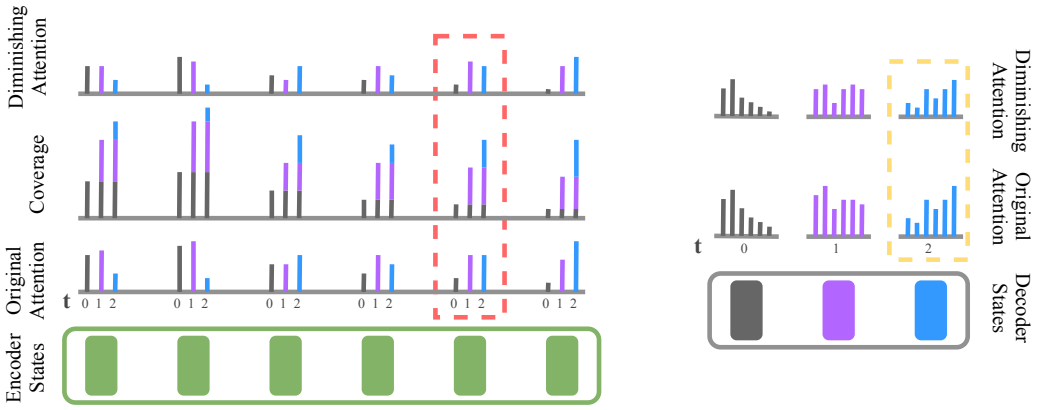

Figure 1: Illustration of diminishing attention in an encoder-decoder model over the encoder states. The brown, purple and blue bars are for the first, second and third decoding step respectively. The left figure shows coverage, original attention and diminishing attention for decoding steps $t = 0$ to 2. The **red dashed block** shows how diminishing attention is related to original attention and coverage for the same single encoder state. For example, at decoding step 2, even though original attention has the same value as that at step 1, the diminishing attention is smaller since the encoder state has been covered more from steps 0 to 1. The right figure shows a summary of the original and diminishing attentions on the encoder states of the left figure in the first, second and third decoding steps. The **yellow dashed block** shows at decoding step 2, the effective attentions for encoder states with higher coverage (*e.g.,* the first encoder state) have diminished more than those with lower coverage (*e.g.,* the last encoder state).

Following previous work (Wu et al., 2016; See et al., 2017), we quantify the degree of coverage of an encoder state as the sum of the set of attentions that the decoder puts on the state in the course of generating the output sequence. Let us consider adding a new token w into two outputs $\mathbb{S}$ and $\mathbb{S}'$, where the concepts covered by $\mathbb{S}'$ is a subset of those covered by $\mathbb{S}$. Intuitively, the information gain from adding w to $\mathbb{S}'$ should be higher than adding it to $\mathbb{S}$, as the new concepts carried by w might have already been covered by those that are in $\mathbb{S}$ but not in $\mathbb{S}'$. This is indeed the *diminishing return* property.

We thus put forward our hypothesis on a desirable property of the neural coverage function that it should be submodular. The greedy algorithm proposed by Nemhauser et al. (1978) approximates the solution to the submodular maximization problem (Eq. 1) with an optimality of 0.63 or higher. For that purpose, the attention scores should be added to the coverage in a greedy manner. However, greedy search among all the states is not possible when the decoder states are generated autoregressively, one at a time. We therefore propose a simplified and principled solution as detailed below.

Let $\mathbb{A}_i = \{a_i^0, \ldots, a_i^t\}$ denote the set of attention scores that an encoder state i receives from the first ($t = 0$) till the current decoding step t, and $F : 2^{\mathbb{A}_i} \rightarrow \mathbb{R}$ be a set function that maps these scores to a score which we define as *submodular coverage* at the current step t.[1]

Definition 3.1. Submodular coverage:

$$F(\mathbb{A}_i^t) = g(\sum_{t'=0}^{t} a_i^{t'}) + b \tag{2}$$

where g is a concave and non-decreasing function (*e.g.,* $\log(x+1)$, $\sqrt{x+1}$), and b is a constant and equal to $-g(0)$. F is monotone submodular because it imposes a concave function on the modular or additive coverage function $f = \sum_{t'=0}^{t} a_i^{t'}$ (see the composition property mentioned in §2.1).

[1]For convenience of developing our method, we define c_i^t as the sum of attentions from the first until the current decoding step instead of the previous decoding step.

3.2 DIMINISHING ATTENTION

By subtracting the submodular coverage between the current step and the previous step, we model diminishing attention scores based on the original attention. Formally, *diminishing attention* (DimAttn) is defined as

$$\textbf{DimAttn}_i^t = F(\mathbb{A}_i^t) - F(\mathbb{A}_i^{t-1}), \tag{3}$$

which models *diminishing return* directly, and will be used as the attention weight corresponding to the encoder state i to produce the context vector at decoding step t to predict the next token.

Thus the effective attention scores are optimized with a submodular function. The diminishing return property of F in Eq. 3 realizes the effect that if an encoder state i receives the same amount of attention at two different decoding steps t and t' such that $t' > t$, the effective attention would diminish more at t' because the coverage at t' is larger. Furthermore, because g is concave, when two encoder states i and j have different amounts of coverage at step $t-1$, and they receive the same attention score at step t, the state with a larger coverage from previous steps would receive a smaller effective attention. We visualize these two properties of diminishing attention in Figure 1.

Effective coverage. The *effective coverage* of an encoder state is the sum of effective attention scores that it receives from the first till the current decoding step.

Theorem 3.1. *The effective coverage with diminishing attention is submodular.*

Proof. Let the effective coverage of an encoder state i at decoding step t be ec_i^t, then we can show

$$ec_i^t = \sum_{t'=0}^{t} \text{DimAttn}_i^{t'} = \sum_{t'=0}^{t} (F(\mathbb{A}_i^{t'}) - F(\mathbb{A}_i^{t'-1})) = F(\mathbb{A}_i^t) - F(\emptyset) = F(\mathbb{A}_i^t) \tag{4}$$

where all the terms in between get cancelled. Since $F(\mathbb{A}_i^t)$ is submodular, ec_i^t is also submodular. □

To emphasize, the effective coverage that each encoder state acquires from the decoder at every decoding step is equal to the submodular coverage defined in Eq. 2, while coverage is apparently modular with attention. Additionally, since g is monotone non-decreasing, it is guaranteed that although the coverage has been changed, the encoder states which receive the largest coverage with the original attention still receive the largest effective coverage with the diminishing attention.

3.3 DYNAMIC DIMINISHING ATTENTION

Using a single submodular coverage function alone may not yield the most appropriate diminishing return effect of the coverage for each encoder state in the decoding process due to the lack of flexibility. More ideally, the model should be capable of further adopting varied degrees of diminishing effect as the decoding proceeds.

Let $F_1(\mathbb{A}_i^t) = g_1\left(\sum_{t'=0}^{t} a_i^{t'}\right) + b_1$ and $F_2(\mathbb{A}_i^t) = g_2\left(\sum_{t'=0}^{t} a_i^{t'}\right) + b_2$ be two different submodular coverage functions. We assume g_1 has a smaller first-order derivative than g_2, thus given the same A_i^t, the diminishing effect of the submodular coverage F_1 would be stronger than that of F_2.

If an encoder state has received a particularly large attention at a certain step, the weight of the more aggressive diminishing function should increase. We thus compute the probability of applying a more aggressive diminishing function at step t as $P_i^t = \max_t(\mathbb{A}_i^{t-1})$. We use it to dynamically control the relative weights of the diminishing functions. The **dynamic diminishing attention** (DyDimAttn) is thus defined as:

$$\textbf{DyDimAttn}_i^t = P_i^t[F_1(\mathbb{A}_i^t) - F_1(\mathbb{A}_i^{t-1})] + (1 - P_i^t)[F_2(\mathbb{A}_i^t) - F_2(\mathbb{A}_i^{t-1})] \tag{5}$$

which is a convex combination of two diminishing attentions, where the diminishing attention which diminishes faster is weighted with P_i^t and the other weighted with $(1 - P_i^t)$.

Since P_i^t keeps changing, the proof of effective coverage of diminishing attention (Eq. 4) is not suited for dynamic diminishing attention. Thus we prove the submodularity of the effective coverage of dynamic diminishing attention with the definition of submodular functions.

Theorem 3.2. *The **effective coverage** of dynamic diminishing attention is submodular.*

Proof. The effective coverage of an encoder state i at step t with dynamic diminishing attention is

$$ec_i^t = \sum_{t'=0}^{t} \text{DyDimAttn}_i^{t'} \tag{6}$$

Coverage increases as the set A_i^t gets larger over steps 1 to t and it is obvious that P_i^t is monotone non-decreasing over steps 1 to t. The return of adding the same amount of original attention score to the set A_i^t is smaller at a later step as the weight over the concave function which diminishes faster (P_i^t) becomes larger over steps 1 to t. Thus by definition 2.1, the effective coverage of dynamic diminishing attention is submodular. □

4 EXPERIMENTS

From the perspective of coverage, text summarization aims to recover a compressed version of the source document, concentrating on the most important concepts; image-paragraph generation aims to recover descriptions of the image regions while ignoring minor details; and MT aims to recover the full of the source, articulating every detail. In this section, we show that diminishing attentions improve the performance of these three tasks with different levels of recovering rate. Other than the method-specific ones, we use the same hyper-parameters as the baseline for most settings. See Appendix A for more implementation details.

4.1 ABSTRACTIVE TEXT SUMMARIZATION

Abstractive summarization involves generating novel phrases to cover the most important information of the input document in a human-like fashion. State-of-the-art pretraining-based abstractive summarization models (Lewis et al., 2019; Yan et al., 2020) suffer from the problem of having repetitive phrases in the output, which has been addressed by blocking duplicated trigrams during inference (Paulus et al., 2018).

Setup. We use two benchmark news summarization datasets following standard splits: CNN/DM (Hermann et al., 2015; Nallapati et al., 2016) and NYT50 (Durrett et al., 2016).

CNN/DM. On CNN/DM, we first evaluate our method based on the LSTM based Pointer-Generator (PG) model (See et al., 2017) which we fine-tune with our diminishing attentions. Following the original setting, the source article is truncated to 400 tokens in training the baseline PG models. Inclusion of more input tokens does not give additional gain to the baselines, whereas exposing the models to more input tokens was beneficial on the validation set when diminishing attentions were employed. We truncate the source article to 600 tokens for training with DimAttn and 800 tokens for DyDimAttn. We include a comparison in the same 400-token setup in Appendix B.1.

On CNN/DM, we also evaluate our attentions within the recently proposed SoTA model BART (Lewis et al., 2019), where we replace the last 7 layers of encoder-decoder cross attention with our diminishing attentions and finetune it. Given the same set of hyper-parameters, it takes around 3.81 seconds for training each batch with the original attention and around 3.97 seconds for training each batch with the diminishing attention on 1 RTX2080 GPU. This means that training the diminishing attention takes only around 0.42% extra amount of time for training one batch.

NYT50. On NYT50, we evaluate based on the SoTA BERT-based Transformer model (Liu & Lapata, 2019a) by replacing the encoder-decoder cross attention at the last layer of the decoder with our attentions and fine-tuning it.

We evaluate the performance with F1 ROUGE (Lin, 2004) and **MoverScore** (Zhao et al., 2019), which is Earth Mover distance based on BERT (Devlin et al., 2019) contextual embeddings.

Results. *ROUGE scores.* We show the results of PG-based models and BART-based models on CNNDM in Table 1. Our method is effective on both LSTM and BART-based models, showing its generalizability across network architectures. In the third block, we show results of recent state-of-the-art models on CNNDM and our model based on BARTSum outperforms all of them.

MoverScore. In Tables 1 and 2, we have also shown MoverScore results for the models on the respective datasets. The consistent improvements in MoverScore show that models equipped with

Table 1: ROUGE F1 score and MoverScore (1-gram and 2-grams) results on CNN/DM. We also report the size of pretraining data (P-Data) and parameters (Params) of each model.

	ROUGE Scores			MoverScore			
Model	**R-1**	**R-2**	**R-L**	**1-gr**	**2-gr**	**P-Data**	**Params**
LEAD-3	40.00	17.50	36.28	–	–		
LSTM-based							
PG	36.69	15.92	33.63	12.46	19.37	0	27M
PG + Cov.	39.08	17.09	35.92	17.55	24.17	0	27M + 512
PG + Dim	40.01	17.74	36.94	17.56	24.16	0	27M
PG + DyDim	**40.13**	**17.94**	**37.21**	**17.77**	**24.38**	0	27M
Large-Size Pretrained Models							
ERNIE-GEN (Xiao et al., 2020)	44.02	21.17	41.26	–	–	16G	340M
$\text{PEGASUS}_{\text{C4}}$ (Zhang et al., 2019)	43.90	21.20	40.76	–	–	750G	568M
$\text{PEGASUS}_{\text{HugeNews}}$ (Zhang et al., 2019)	44.17	21.47	41.11	–	–	3.8T	568M
ProphetNet (Yan et al., 2020)	44.20	21.17	41.30	–	–	160G	400M
BARTSum (Lewis et al., 2019)	44.16	21.28	40.90	22.34	28.47	160G	400M
BARTSum + Dim	44.86	**21.76**	41.62	23.05	29.04	160G	400M
BARTSum + DyDim	**44.92**	21.70	**41.66**	**23.12**	**29.13**	160G	400M

Table 2: ROUGE Scores and MoverScores results on the NYT50 summarization dataset.

	ROUGE Scores			MoverScore	
Model	**R-1**	**R-2**	**R-L**	**1-gr**	**2-gr**
LEAD-3	24.52	12.78	21.75	–	–
BertSum	48.33	31.03	44.85	28.16	34.09
+ Dim	49.29	**31.72**	45.78	29.10	34.87
+ DyDim	**49.46**	31.59	**45.94**	**29.24**	**35.00**

Table 3: CIDEr results based on two training regimes (Cross-Entropy and Self-Critical) on the Stanford Image-Paragraph dataset.

Model	**Cross-Entropy**	**Self-Critical**
Baseline	22.68	30.63
+Dim	25.47	33.15
+DyDim	**25.49**	**33.28**

diminishing attentions are capable of generating outputs more semantically similar to the gold summary than the baselines. This indicates that our method is more effective in capturing the overall meaning of the source article than the baselines.

4.2 Image-paragraph generation

Image-paragraph generation aims to generate a coherent paragraph to describe different aspects of an input image. The widely-used Stanford Image Paragraph dataset (Krause et al., 2017) has been known to be too small in size for the model to learn the language structure and pattern. Previous work (Melas-Kyriazi et al., 2018) has shown that the generated paragraphs usually contain many repetitive phrases and sentences while covering the source poorly. We follow (Melas-Kyriazi et al., 2018) in using the Top-Down model from (Anderson et al., 2018) with a CNN pretrained for object detection and a 1-layer LSTM as the language decoder with top-down attention applied over max-pooled features of 40-100 regions of interests (RoI), which we replace with our diminishing attentions. We run on the Stanford Image-Paragraph dataset using the standard splits. We fine-tune on two standard baselines (Melas-Kyriazi et al., 2018) with different training strategies including minimizing the cross entropy loss and optimizing the CIDEr (Vedantam et al., 2015) reward with self-critical reinforcement learning. From the results in Table 3, we observe 2.81 improvement in CIDEr (Vedantam et al., 2015) over the cross entropy training baseline, and 2.65 improvement over the self-critical training baseline, setting a new state-of-the-art.

4.3 Neural machine translation

We incorporate our attention mechanisms into the cross-attention at the last decoder layer of the Transformer-Big model (Ott et al., 2018) which consists of a 6-layer transformer and fine-tune the pretrained baseline model. We run on the WMT'14 English-German (En-De) and WMT'14 English-French (En-Fr) tasks following the settings of Ott et al. (2018). As shown in Table 4, we obtain 0.4 improvement on En-De and 0.3 improvement on En-Fr in standard tokenized BLEU, which to our

Table 5: Effective coverage entropy $\mathcal{H}$, repetition rate and BLEU score on subsets with short and long source sentences on English-German newstest2014.

	Entropy	**Repetition**		**BLEU**		
Model	$\mathcal{H}$	**uni-rep(%)**	**bi-rep(%)**	**short**	**long**	**overall**
Reference	–	4.46	0.08	–	–	–
Transformer-Big	2.08	5.87	0.18	28.9	29.5	29.3
Transformer-Big + Dim	2.37	5.82	0.16	29.1	29.8	29.7
Transformer-Big + DyDim	2.41	5.85	0.16	29.1	29.9	29.7

knowledge are state of the art without using extra monolingual data (Edunov et al., 2018; Zhu et al., 2020) or parse tree information (Nguyen et al., 2020). We also compute statistical significance for the difference in BLEU scores between our model and the Transformer using paired bootstrap resampling (Koehn, 2004). We conclude that the diminishing attention and dynamic diminishing attention are better than the baseline with at least 99.5% statistical confidence.

We conduct further analysis on the WMT En-De task. We first compare the entropy of the normalized effective coverage across all the encoder states at the end of inference, which is denoted as $\mathcal{H}$ (t is the final step and $\mathbb{A}_i^t$ contains all the attention scores of encoder state i).

$$\mathcal{H} = -\sum_i ec_i^t \log ec_i^t \qquad \text{where } ec_i^t = ec_i^t / \sum_i ec_i^t$$

and we take the average of $\mathcal{H}$ of all the test instances. From Table 5, we see that entropy of the effective coverage of our attentions are higher than that of the baseline. This indicates that the effective coverage distribution of our method is more even across the encoder states than that of the baseline, which suggests that more concepts of the source are covered and the coverage is improved.

Table 4: BLEU scores and statistical **conf**idence (Koehn, 2004) on WMT newstest2014 for English-German and English-French translation tasks.

Model	**En-De** (conf)	**En-Fr** (conf)
Transformer-Big	29.3	43.2
Wu et al. (2019) (reported)	29.7	43.2
Transformer Big +Dim	29.7 (99.5)	43.4 (99.5)
Transformer Big +DyDim	**29.7** (99.7)	**43.5** (100.0)

Next, we compare the uni- and bi-gram repetition rates in percentage computed with the duplicate n-grams in a summary and see that repetitions become lower with our attentions. Finally, we sort the source sentences in the testset by length and split it into two halves – *short* and *long*. We observe that our method has more BLEU gains on the longer half. Intuitively, longer source sentences are in more need of even effective coverage to ensure each and every detail of the source is translated.

5 Analysis

We provide more analysis of our method taking summarization on CNN/DM as a case study.

5.1 Quantitative and qualitative analysis

We empirically analyze that submodularity imposed on the coverage enables our models to generate summaries with better coverage of the source document from two aspects: *layout bias* and *repetition ratio*. We also compare our method with trigram-blocking (Paulus et al., 2018).

Table 6: **N-gram overlaps** with lead-3.

Model	**R-1**	**R-2**	**R-L**
Reference	40.00	17.50	36.19
PG + Cov.	54.40	42.54	52.25
PG + Dim	**53.02**	**39.97**	**50.75**
PG + DyDim	53.08	40.25	50.93

Layout bias. *Layout bias* is a common issue in news datasets where the leading section of an article contains the most important information (Kryscinski et al., 2019), and encoder-decoder models are prone to remembering this pattern and ignoring other important content in the rest of the article (Kedzie et al., 2018). Truncating the documents to 400 tokens caters to this bias. By increasing the maximum encoding steps to 600 for DimAttn and 800 for DyDimAttn (§4.1), we feed the model more information and allow it to automatically learn to extract important information from a longer source. Table 6 shows that our models have less n-gram overlaps with lead-3 sentences compared to the baseline without compromising the ROUGE scores. This indicates that diminishing attentions enable more comprehensive coverage of the source while maintaining a focus on the most important content.

Repetition *Trigram-blocking* is widely used for eliminating redundancy in summarization (Liu & Lapata, 2019a; Gehrmann et al., 2018). However, blocking alone does not guarantee high quality as

it is not learned. In our analysis of the output from the PG + Cov baseline, we observe that although adopting trigram-blocking results in less repetition and higher ROUGE, the generated summaries are excessively extractive while our method leads to less repetition and more abstractiveness. Our method also improves the abstractiveness of the BERT-based model (see Appendix B.3 and Appendix B.4 for examples and statistics).

5.2 HUMAN EVALUATION

We conducted a user study on Amazon Mechanical Turk for the BART-based models. We randomly sampled 500 examples from the CNN/DM test set where each example was distributed to 3 US workers. Each worker was asked to evaluate *representativeness*, *readability* and *factual correctness* of the system summaries. We provided the following definition of representativeness and readability as guidelines to the workers: representativeness refers to how well the summary covers the most significant concepts in the source, more specifically, the summary should cover the important concepts and maintain conciseness at the same time; readability is defined as *grammaticality and coherence* where the annotators evaluate the text quality, *i.e.,* being fluent, logical, consistent and uunderstandable. Annotators were presented with two randomly ordered summaries and asked to pick the better one (win) or equally good (tie) in terms of representativeness and readability and evaluate if the summaries are factually correct (more details about the study are in Appendix B.5). We show the human agreement percentage and p-value of sign test to assess whether the differences between our model and the baseline were significant. From the results in Table 7, we notice that our model has significantly better representativeness, reinforcing our hypothesis that the coverage of abstractive summarization should be submodular. We also found that our method increases readability and maintains the factual correctness of the baseline.

Table 7: Human evaluation results on Representativeness, Readability and Factual Correctness. Human Agr. is the percentage agreement.

Model	**Repr. Win**	**Read. Win**	**Fac.**
BARTSum	35.9%	42.5%	96.3%
DyDim	**57.2%**	**45.1%**	**96.9%**
Tie	6.87%	12.4%	–
Human Agr.	64.1%	62.1%	90.0%
p-value (sign test)	1e-6	0.0014	–

5.3 DISCUSSION

With regard to neural text generation models aimed at recovering the source, the notion of coverage is more imperative when a model generates repetitive concepts while trying to recover the source. This could possibly coincide with that the decoder degenerates when it has yet to be sufficiently trained. For example, in image-paragraph generation, the decoder language model may not be well-trained because of the insufficient data. We also notice that models produce highly repetitive outputs at the early stage of training. However, diminishing attentions also effectively improve the performance of large-scale models including the BARTSum Transformer model for summarization with 400M parameters and the Transformer-Big model for machine translation with 220M parameters. This shows that even with the power of transfer learning brought by the giant pre-trained encoder-decoder model for BARTSum, or the huge size of parameters and dataset for Transformer-Big, the existing encoder-decoder attention mechanism may not incorporate the most appropriate inductive bias for coverage in these tasks, while our method, by making a simple architectural change to the attention mechanism, effectively improves their performance without adding new parameters. We hypothesize this is because that diminishing attentions explicitly model the submodularity of the neural coverage, which has been shown a natural fit for content selection (Lin & Bilmes, 2011a;b).

6 CONCLUSION

We have defined a class of *diminishing attentions* and in turn, proved the submodularity of the effective neural coverage. Submodularity is desirable for coverage. To address the problem that greedy selection cannot be utilized over attention scores in the neural framework, we propose a simplfied solution. Experimental results and a series of analyses on three tasks across two modalities, five datasets and three neural architectures demonstrate that our method produces text outputs of good quality, outperforms comparable baselines and achieves state-of-the-art performance.

REFERENCES

P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6077–6086, 2018.

Taylor Berg-Kirkpatrick, Dan Gillick, and Dan Klein. Jointly learning to extract and compress. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 481–490, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P11-1049.

Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pp. 335–336, New York, NY, USA, 1998. ACM. ISBN 1-58113-015-5. doi: 10.1145/290941.291025. URL http://doi.acm.org/10.1145/290941.291025.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

Greg Durrett, Taylor Berg-Kirkpatrick, and Dan Klein. Learning-based single-document summarization with compression and anaphoricity constraints. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1998–2008, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1188. URL https://www.aclweb.org/anthology/P16-1188.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 489–500, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1045. URL https://www.aclweb.org/anthology/D18-1045.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4098–4109, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1443. URL https://www.aclweb.org/anthology/D18-1443.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 28*, pp. 1693–1701. Curran Associates, Inc., 2015. URL http://papers.nips.cc/paper/5945-teaching-machines-to-read-and-comprehend.pdf.

Chris Kedzie, Kathleen McKeown, and Hal Daumé III. Content selection in deep learning models of summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1818–1828, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1208. URL https://www.aclweb.org/anthology/D18-1208.

Philipp Koehn. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 388–395, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W04-3250.

J. Krause, J. Johnson, R. Krishna, and L. Fei-Fei. A hierarchical approach for generating descriptive image paragraphs. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3337–3345, 2017.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 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)*, pp. 540–551, Hong Kong, China, November 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D19-1051.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Proc. ACL workshop on Text Summarization Branches Out*, pp. 10, 2004. URL http://research.microsoft.com/~cyl/download/papers/WAS2004.pdf.

Hui Lin and Jeff Bilmes. Multi-document summarization via budgeted maximization of submodular functions. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 912–920, Los Angeles, California, June 2010. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N10-1134.

Hui Lin and Jeff Bilmes. A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 510–520, Portland, Oregon, USA, June 2011a. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P11-1052.

Hui Lin and Jeff Bilmes. Word alignment via submodular maximization over matroids. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 170–175, Portland, Oregon, USA, June 2011b. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P11-2030.

Yang Liu and Mirella Lapata. Text summarization with pretrained encoders. 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)*, pp. 3728–3738, Hong Kong, China, November 2019a. Association for Computational Linguistics. doi: 10.18653/v1/D19-1387. URL https://www.aclweb.org/anthology/D19-1387.

Yang Liu and Mirella Lapata. Hierarchical transformers for multi-document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5070–5081, Florence, Italy, July 2019b. Association for Computational Linguistics. doi: 10.18653/v1/P19-1500. URL https://www.aclweb.org/anthology/P19-1500.

Luke Melas-Kyriazi, Alexander Rush, and George Han. Training for diversity in image paragraph captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 757–761, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1084. URL https://www.aclweb.org/anthology/D18-1084.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar GulÌ‡lçehre, and Bing Xiang. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pp. 280–290, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/K16-1028. URL https://www.aclweb.org/anthology/K16-1028.

G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions–i. *Math. Program.*, 14(1):265–294, December 1978. ISSN 0025-5610. doi: 10.1007/BF01588971. URL https://doi.org/10.1007/BF01588971.

Xuan-Phi Nguyen, Shafiq Joty, Steven Hoi, and Richard Socher. Tree-structured attention with hierarchical accumulation. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJxK5pEYvr.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation (WMT)*, 2018.

Romain Paulus, Caiming Xiong, and Richard Socher. A deep reinforced model for abstractive summarization. In *ICLR*, 2018. URL https://openreview.net/pdf?id=HkAClQgA-.

Abigail See, Peter J. Liu, and Christopher D. Manning. 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)*, pp. 1073–1083, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1099. URL https://www.aclweb.org/anthology/P17-1099.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. Modeling coverage for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 76–85, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1008. URL https://www.aclweb.org/anthology/P16-1008.

R. Vedantam, C. L. Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4566–4575, 2015.

Felix Wu, Angela Fan, Alexei Baevski, Yann Dauphin, and Michael Auli. Pay less attention with lightweight and dynamic convolutions. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SkVhlh09tX.

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

Dongling Xiao, Han Zhang, Yukun Li, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-gen: An enhanced multi-flow pre-training and fine-tuning framework for natural language generation, 2020.

Yu Yan, Weizhen Qi, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training, 2020.

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

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 563–578, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1053. URL https://www.aclweb.org/anthology/D19-1053.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tieyan Liu. Incorporating bert into neural machine translation. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Hyl7ygStwB.

Table 8: Summarization dataset statistics.

Dataset	CNN/DM	NYT
Size	312,085	104,286
Train/val/test	287,227/13,368/11,490	96,834/4,000/3,452

A IMPLEMENTATION DETAILS

We use NVIDIA RTX 2080Ti for training PG-based and BERT-based summarization models and the En-De neural machine translation (NMT) models, NVIDIA Tesla V100 for training the BART-based summarization models and En-Fr NMT models, and NVIDIA GTX1080Ti for training image-paragraph generation models. All models are trained with a single GPU.

A.1 ABSTRACTIVE SUMMARIZATION

Datasets We use the entity non-anonymized version of CNN/DM, and use the same data preprocessing on CNN/DM and NYT as the baseline models including the PG-based models[2], BERT-based models[3] and BART-based models[4]. Our train/validation/test split of NYT is the same as those of (Liu & Lapata, 2019a). See Table 8 for the dataset statistics.

PG-based models We train the PG models — PG and PG with Coverage (PG + Cov.), as our baselines using the settings from See et al. (2017). We use $g = g_1 = \log(x + 1)$ for diminishing attention and $g_2 = \sqrt{x + 1}$ for dynamic diminishing attention [5]. For a fair comparison, we train a PG model without their coverage mechanism, and apply our proposed diminishing attentions (DimAttn or DyDimAttn) from 230k iterations onward and train for 10k iterations for DimAttn and 15k iterations for DyDimAttn. We use the Adagrad optimizer with a learning rate of 0.15 for training PG and the same learning rate of 0.15 for training the coverage mechanism with a batch size of 32. We found that increasing the beam size from 4 to 6 leads to significant improvements to our model while it does not give any improvement to the baselines. We use a length normalization factor of 1.5 for our method and apply trigram-blocking (Paulus et al., 2018) during inference.

BERT-based models The BERT-based models(Liu & Lapata, 2019a) are trained with an initial learning rate of 2e-3 for encoder and 0.2 for decoder using the Adam optimizer. We fine-tune the models of Liu & Lapata (2019a) for 10k updates for diminishing attention and 15k updates for dynamic diminishing attention using the same hyper-parameters as (Liu & Lapata, 2019b). We use $g = g_1 = 2.2$ and same as g_2 as PG-based models. We use a beam size of 5 and length penalty of 1.

BART-based models We finetune the BART-large model (Lewis et al., 2019) with 406M parameters for 5 epochs using Adam optimizer with a similar learning rate schedule as Lewis et al. (2019) - a learning rate of 3e-5 and 500 warm-up steps. Maximum token per GPU is set to 1024 and we accummulate gradients for 32 times for one update. We use a length normalization factor of 0.0 and a beam size of 4 and apply trigram-blocking for inference. We selected the exponent of g in a range of $\{0.55, 0.65, 0.75\}$. We use the same exponent for g_1 as g and select $g2$ in a range of $\{0.5, 0.6, 0.7\}$. We use $g = g_1 = (x + 1)^{0.65}$ for diminishing attention and $g_2 = (x + 1)^{0.6}$ for dynamic diminishing attention.

A.2 MACHINE TRANSLATION

The WMT14' train sets contain about 4.5 million instances for the En-De task and 35 million instances for the En-Fr task. We use newstest2013 as the validation set, and newstest2014 as the testing set. We use Adam optimizer with a learning rate of 0.0005 to fine-tune the baseline models

[2] https://github.com/abisee/cnn-dailymail

[3] https://github.com/nlpyang/PreSumm

[4] https://github.com/pytorch/fairseq/blob/master/examples/bart/README.summarization.md

[5] The derivative of $\sqrt{x+1}$ becomes larger than that of $\log_a(x+1)$ when $x > 3 \log a$, which rarely happens in summarization.

Table 9: Ablation study. All the models have the same settings as the PG baseline, *i.e.,* the maximum encoding steps are all set to 400 (thus the difference in ROUGE scores from the first block in Table 1 in the main paper.

Model	**R-1**	**R-2**	**R-L**
PG + Cov.	39.08	17.09	35.92
PG + DimAttn	39.30	17.48	36.31
+ Length Norm.	39.64	17.51	36.61
+ Trigram-Blocking	39.92	17.64	36.88
PG + Cov.	39.08	17.09	35.92
PG + DyDimAttn	39.70	17.80	36.67
+ Length Norm.	39.92	17.80	36.89
+ Trigram-Blocking	40.13	17.87	37.09

for 300 updates for En-De and 3500 updates for En-Fr with 16 times of gradient accumulation for each update. On En-De, we use a beam size of 4 and length penalty factor of 0.7 for our diminishing attentions. On En-Fr, we use a beam size of 4 and length normalization factor of 0.6. We use the same g, g_1 and g_2 as the BART-based models.

A.3 Image-paragraph generation

We use the standard split of 14,575/2,487/2,489(train/val/test) for the Stanford Image-Paragraph dataset (Krause et al., 2017). We use Adam optimizer with a learning rate of 5e-4 and batch size of 20 for fine-tuning the cross-entropy baseline for 20 epochs and a learning rate of 6.7e-06 and batch size of 30 for fine-tuning the self-critical training baseline for 40 epochs. We decay the learning rate every 5 epochs and use a learning rate decay rate of 0.85. We select the log base of g in a range of $\{1.9, 2.1, 2.3, 2.7\}$ and use the same log base for g_1 as g, and select g_2 in a range of $\{1.75, 1.95, 2.15, 2.35\}$. We use $g = \log_{1.9}(x+1)$ for diminishing attention and $g_1 = \log_{1.9}(x+1)$ and $g_2 = \log_{1.95}(x+1)$ for dynamic diminishing attention for both cross-entropy and self-critical training. We follow the baselines in all the other hyperparameter settings.

B Analysis

B.1 Ablation study on PG-based CNN/DM models

We conduct an ablation study to analyze the impact of each component in our model. Table 9 shows the improvements for diminishing attentions and other components as they are added one at a time. By adding DimAttn or DyDimAttn only, our model outperforms the baselines. Length normalization and trigram-blocking further improve the ROUGE scores.

B.2 N-gram repetition ratio of PG-based CNN/DM models

We measure the *repetition ratio* by calculating the duplicate n-grams in a summary. Figure 2 shows the repetition ratio of summaries generated by PG baselines, our model and gold summaries. Our method yields significantly less repetition in terms of unigrams and bigrams compared to the vanilla PG. It outperforms the PG+Cov. model as well. trigram repetition is completely eliminated as trigram-blocking is applied.

B.3 Examples of summaries

We show two examples of summaries generated by the PG + Cov. model with trigram-blocking and our PG + DyDimAttn model in Table 10 and Table 11. We also show two examples of summaries generated by BARTSum and BARTSum + DyDim model in Table 12 and Table 13.

Table 10: The first example from the CNN/DM test set showing the outputs of PG + Cov. + Trigram-blocking and our model.

Article
Ultimately , Bristol City were never destined to become the first football league club to win promotion this season at Deepdale, even though they are inching ever closer. Swindon's win over Peterborough meant Steve Cotterill will have to wait until Tuesday to finish the job of returning to the championship, but almost as importantly they made sure Preston wouldn't make any ground on them in the top two. Three points at Bradford on Tuesday will do the trick - only that standing in their way now. *(...)*
Reference
Second-placed preston hosted league one leaders Bristol City. Jermaine Beckford fired the home side into the lead in the 59th minute. Aaron Wilbraham equalised for Bristol City four minutes later.
PG + Cov. + Trigram-blocking
Swindon's win over Peterborough meant Steve Cotterill will have to wait until Tuesday to finish the job of returning to the championship, but almost as importantly they made sure Preston wouldn't make any ground on them in the top two. City had dispensed of their rivals on Tuesday night, hammering them 3-0, and took great delight in their triumph.
PG + DyDimAttn
Bristol City beat Preston 3-0 in the premier league on Tuesday night. Aaron Wilbraham opened the scoring for the hosts in the second half. The result means Steve Cotterill will have to wait until Tuesday to finish the job.

Table 11: The second example from the CNN/DM test set showing the outputs of PG + Cov. + Trigram-blocking and our model. Highlighted spans are the phrases whose lengths are equal to or longer than 3 tokens and are copied verbatim from the source document.

Article
(...) Those years will always hang heavy around Wenger's neck. Yes, yes, there's the FA cup currently sitting there in the trophy cabinet alongside the shiny shield they won in august. And yes, yes, they are clear favourites to win it again. An FA cup isn't the definition of success for the great clubs of europe. Only the Wenger apologists can't see that. Olivier Giroud celebrates Arsenal's third goal by Alexis Sanchez in their win over Liverpool on Saturday. Arsene Wenger's gunners now have 15 wins from 17 games since new year's day, but it's come too late. So are gooners seduced by this end of season run they're currently on? It's impressive, no question. 15 wins in 17 games since their loss new year's day. But this is classic Arsene Wenger isn't it? After that defeat at Southampton on January 1 the gunners were sixth in the table , 13 points behind leaders Chelsea. *(...)*
Reference
Arsenal have a chance at the premier league title if Chelsea throw it away. The gunners are on a terrific run of 15 wins in 17 since January 2 's loss. The two defeats in this spell were at Tottenham and at home to Monaco. These were the two key games all arsenal fans desperately wanted to win. Read : Arsenal have exactly the same record in league as last season. Click here for all the latest Arsenal news.
PG + Cov. + Trigram-blocking
Arsene Wenger's gunners now have 15 wins from 17 games since new year's day, but it's come too late. Arsene Wenger isn't it ? After that defeat at Southampton on january 1 the gunners were sixth in the table, 13 points behind leaders chelsea. The FA cup currently sitting there in the trophy cabinet alongside the shiny shield they won in august.
PG + DyDimAttn
Arsenal were sixth in the table, 13 points behind leaders chelsea. Arsene Wenger's side have 15 wins from 17 games since new year's day. But this is not the definition of success for the great clubs of europe. Click here for Manchester United transfer news.

Table 12: The first example from the CNN/DM test set showing the outputs of the BARTSum and BARTSum + DyDim.

Article
The NSW State Emergency Service (SES) are warning the public that scammers are making calls falsely claiming to fundraise, abusing people's goodwill as NSW is ravaged by wild weather. The public have been alerted to the heartless scheme via an important notice issued on the NSW SES Facebook page. 'The NSW SES does NOT request donations from the public,' NSW SES assure the public. 'We will never call and ask for money. 'If you are receiving calls of this nature, please be advised it is a scam.' The public have been alerted to the heartless scheme via an important notice issued on the NSW SES Facebook page. NSW SES have received more than 6500 requests for help since the storms began on Monday, with flash flooding, trees down and power outages across the Sydney, Newcastle and Hunter regions.*(...)* People have reacted with disgust to the attempt to con people out of their money, condemning those responsible for trying to benefit from a crisis. *(...)*
Reference
NSW SES warns scammers are phoning people claiming to fundraise. The State Emergency Service say they never call and ask for money. People have responded with disgust at the heartless. Con artists are taking advantage of people's goodwill as NSW is facing severe weather conditions. NSW SES have received more than 6500 requests for help since Monday.
BARTSum
The NSW State Emergency Service (SES) are warning the public that scammers are making calls falsely claiming to fundraise. The public have been alerted to the heartless scheme via an important notice issued on the NSW SES Facebook page. People have reacted with disgust to the attempt to con people out of their money, condemning those responsible for trying to benefit from a crisis. 'Oh my god! Who are these absolute mongrels that take advantage of these sorts of situations!' wrote Belinda Weston.
BARTSum + DyDimAttn
The NSW State Emergency Service (SES) are warning the public that scammers are making calls falsely claiming to fundraise, abusing people's goodwill as NSW is ravaged by wild weather. The public have been alerted to the heartless scheme via an important notice issued on the NSW SES Facebook page. People have reacted with disgust to the attempt to con people out of their money. More than 6500 requests for help since the storms began on Monday, with flash flooding, trees down and power outages across the Sydney, Newcastle and Hunter regions.

Table 13: The second example from the CNN/DM test set showing the outputs of the BARTSum and BARTSum + DyDim.

Article
(CNN)"Jake the dog and Finn the human. The fun will never end. Adventure Time." So begins the dreamy theme song intro to the strangely addictive Cartoon Network TV show that's centered around psychedelic characters like the Ice King, Marceline the Vampire Queen and, of course, Jake and Finn. Now, mega-fans of the hit show can experience "Adventure Time" in the skies. Thai Smile, a subsidiary of Thailand flag carrier Thai Airways, on Thursday unveiled colorful new livery featuring Jake, Finn and the beloved Princess Bubblegum sprawled across an Airbus A320 at Bangkok's Suvarnabhumi International Airport. The interior of the plane also has an Adventure Time theme, with overhead bins, head rests and even air sickness bags covered in the faces of characters from the show. Airlines show off their new flying colors. The Adventure Time plane is the result of a partnership between Thai Airways subsidiary Thai Smile and Cartoon Network Amazone, a new water park near the Thai resort city of Pattaya featuring attractions based on shows that appear on the Turner Broadcasting System channel. Turner Broadcasting is a parent company of CNN. Check out these cool airline liveries. The inaugural Thai Smile Adventure Time flight takes place on April 4, heading from Bangkok to Phuket. *(...)*
Reference
Thai Airways subsidiary Thai Smile features Cartoon Network paint job on A320 jet. Overhead bins, head rests and air sick bags feature characters from Cartoon Network.
BARTSum
Thai Smile unveils colorful new livery featuring Jake, Finn and the beloved Princess Bubblegum. The interior of the plane also has an Adventure Time theme, with overhead bins, head rests and even air sickness bags covered in the faces of characters. The inaugural Thai Smile Adventure Time flight takes place on April 4, heading from Bangkok to Phuket.
BARTSum + DyDimAttn
Thai Airways subsidiary Thai Smile unveils "Adventure Time" livery for new Airbus A320. The interior of the plane also has an Adventure Time theme. The plane is the result of a partnership between Thai Smile and Cartoon Network Amazone. The inaugural Thai Smile Adventure Time flight takes place on April 4.

Figure 2: **N-gram repetition** ratio of the model outputs for the PG baselines, diminishing attentions and reference summaries.

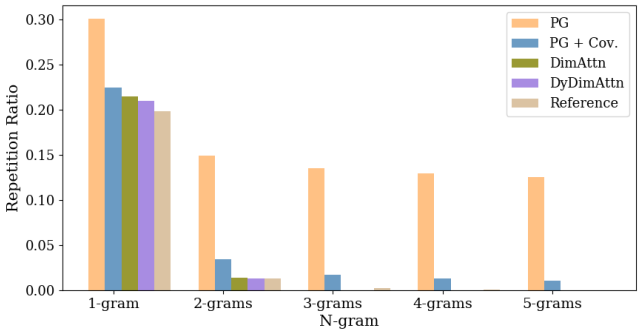

Table 14: Novel n-gram percentage on CNN/DM (PG-based models).

Model	**1-gr**	**2-gr**	**3-gr**	**4-gr**	**5-gr**
Reference	13.6	49.0	67.7	76.9	82.3
PG + Cov.	**0.3**	4.0	7.1	9.7	12.2
PG + DyDimAttn	**0.3**	**4.3**	**7.5**	**10.2**	**12.8**

Table 15: Novel n-gram percentage on NYT (BERT-based models).

Model	**1-gr**	**2-gr**	**3-gr**	**4-gr**	**5-gr**
Reference	21.7	51.9	66.8	75.4	81.3
BertSum	4.9	22.3	36.4	47.3	56.0
BertSum + DyDimAttn	**5.0**	**22.6**	**36.9**	**47.8**	**56.6**

B.4 Abstractiveness ratio

In Table 14, we present the *abstractiveness* of the summaries generated by PG-based models. The baseline model equipped with dynamic diminishing attention[6] has a higher abstractiveness against the baseline.

We present the abstractiveness of the summaries generated by the BERT-based models on NYT in Table 15.

The same consecutive tokens in the source document tend to be generated by the summarization models. This could be because that the source sequence is perfectly fluent and is favored by the language model. However, the diminishing attentions imposed by the submodular coverage inform the model to have more appropriate generation for better coverage, which is not necessarily exactly the same as the source sequence.

B.5 Human evaluation details

We show the interface used for human evaluation in Figure 3.

[6]We omit the abstractiveness of PG + DimAttn as there is a less significant effect.

Figure 3: Human evaluation interface.