

LenAtten: An Effective Length Controlling Unit For Text Summarization

Zhongyi Yu^{*,1}, Zhenghao Wu^{*,1}, Hao Zheng¹,
Zhe XuanYuan^{2,3}, Jefferson Fong¹, Weifeng Su^{†,1,3}

¹Computer Science and Technology Programme, ²Data Science Programme
Division of Science and Technology

³Guangdong Key Lab of AI and Multi-Modal Data Processing
BNU-HKBU United International College, Zhuhai, China

{zhongyicst, zhecwu, zhslzwd97}@gmail.com
{zhexuanyuan, jeffersonfong, wfsu}@uic.edu.cn

Abstract

Fixed length summarization aims at generating summaries with a preset number of words or characters. Most recent researches incorporate length information with word embeddings as the input to the recurrent decoding unit, causing a compromise between length controllability and summary quality. In this work, we present an effective length controlling unit Length Attention (LenAtten) to break this trade-off. Experimental results show that LenAtten not only brings improvements in length controllability and ROUGE scores but also has great generalization ability. In the task of generating a summary with the target length, our model is 732 times better than the best-performing length controllable summarizer in length controllability on the CNN/Daily Mail dataset. ¹

1 Introduction

Automatic text summarization aims at generating a short and coherent summary from one or multiple documents while preserving the main ideas of the original documents. Building upon the conventional summarization task, fixed length text summarization (FLS) demands extra focus on controlling the length of output summaries. Specifically, it requires generating summaries with a preset number of characters or words.

FLS is a rising research topic required in many scenarios. For example, in order to get universal user experiences on multiple platforms and devices, titles and abstracts for news articles are expected to have different numbers of characters. Instead of manually rewriting summaries, FLS can automatically generate required summaries by simply

* Equal Contribution

† Corresponding author

¹Code are publicly available at: <https://github.com/X-AISIG/LenAtten>

Source document

egyptian president hosni mubarak arrived here friday morning to discuss the latest developments of iraqi crisis with his turkish counterpart suleyman demirel .

Reference summary

egyptian president to discuss iraqi crisis with turkish counterpart

Model

Summary

PAULUS	egyptian president arrives in ankara
PAULUS+LA2 (GT)	mubarak arrives in ankara for talks on iraqi crisis with turkish pm
PAULUS+LA2 (30)	egyptian president arrives in ankara
PAULUS+LA2 (50)	egyptian president arrives in ankara for talks on iraq crisis
PAULUS+LA2 (70)	mubarak arrives in ankara for talks on iraqi crisis with turkish president demirel

Table 1: Output examples from the proposed method Length Attention (LA) on the Annotated English Gigaword dataset. Numbers in the parentheses represent different desired lengths. (GT) means the desired length is equal to the number of characters in the reference summary. PAULUS (Paulus et al., 2018).

inputting the desired output length. Besides, FLS can help news editors to reduce post-editing time (Makino et al., 2019) and further improve summary quality (Liu et al., 2018; Makino et al., 2019). Last but not least, as shown in Table 1, with FLS, users can get customizable summaries by setting different desired lengths.

Despite the benefits that could be brought, previous studies on FLS are very limited. Recent researches in FLS apply length information to either (i) the decoder (Kikuchi et al., 2016; Liu et al., 2018; Takase and Okazaki, 2019) or (ii) the optimization objective function (Makino et al., 2019). Though these systems are promising, they have to make a compromise between length controllability and summary quality. Kikuchi et al. (2016); Makino et al. (2019) can generate high-quality summaries, but perform inadequately at controlling length. Liu et al. (2018); Takase and Okazaki (2019) control the output length accurately, but these models suffer from producing summaries with low ROUGE scores.

In this paper, we present an effective length controlling unit, Length Attention (LenAtten). With LenAtten, summarizers can generate high-quality summaries with a preset number of characters, successfully breaks the trade-off between length controllability and summary quality.

Our contributions in this work are as follows: (1) A novel length controlling unit with great generalization capability is proposed to make summarizers generate high-quality summaries with a preset number of characters. (2) Experimental results show that LenAtten can break the trade-off between length controllability and summary quality. The length controllability of the proposed method is the new state-of-the-art on the examined datasets, to our knowledge.

2 Related Work

There are two types of approaches for text summarization: the extractive approach and the abstractive approach. Extractive approaches generate summaries by extracting words or sentences from the original text (Dorr et al., 2003; Nallapati et al., 2017; Liu and Lapata, 2019; Zhong et al., 2020), while abstractive approaches produce novel words or phrases (Rush et al., 2015; Chopra et al., 2016; Nallapati et al., 2016; Gu et al., 2016; See et al., 2017; Fan et al., 2018; Liu and Lapata, 2019).

Derived from the works in general text summarization, two approaches have been developed for FLS: (1) Incorporating length information into the decoder. LenInit proposed in Kikuchi et al. (2016) introduced length information into the initialization stage of a LSTM decoder. Liu et al. (2018) follows a similar approach as LenInit, but it’s based on a CNN sequence-to-sequence architecture. Other studies exploit length information in each decoding step. LenEmb introduced in Kikuchi et al. (2016) generates a learnable embedding for each target length, and uses it as an additional input to its decoder. Takase and Okazaki (2019) extended Transformer’s sinusoidal positional encoding (Vaswani et al., 2017) to make summarizers take account of stepwise remaining length during prediction. (2) Leveraging length information in global optimization methods. Makino et al. (2019) proposed a global optimization method named GOLC. GOLC incorporates length information with the minimum risk training (MRT) optimization method.

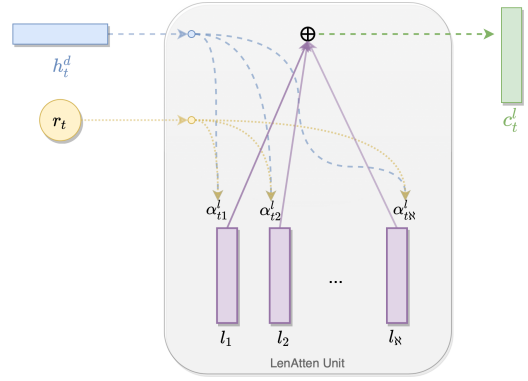


Figure 1: Illustration of the Length Attention Unit. Firstly, decoder hidden state (blue) and remaining length (yellow) are employed to compute the attention weights α^l . Then, the length context vector c_t^l (green) is produced by calculating the weighted sum between attention weights and pre-defined length embeddings (purple). Better viewed in color.

3 Our Approach: Length Attention

The motivation of LenAtten is to separate length information from the input of the recurrent decoding unit and to exploit proper length information based on the stepwise remaining length. As shown in Figure 1, at each decoding step, a length context vector is generated by calculating the weighted sum of a set of pre-defined embedding vectors l_* . Then, the length context vector is concatenated with the decoder hidden state and other attention vectors and fed to the input of the word prediction layer (details are shown in §4.2), so that summarizers can take the remaining length into account. The length context vector c_t^l at t -th decoding step is defined as follows:

$$c_t^l = \sum_{j=1}^{\aleph} \alpha_{tj}^l l_j \quad (1)$$

$$\alpha_{tj}^l = \frac{\exp(e_{tj}^l)}{\sum_{k=1}^{\aleph} \exp(e_{tk}^l)} \quad (2)$$

$$e_t^l = V_l^T \tanh(W_l h_t^d + w_r r_t + b_l), \quad (3)$$

where $e_t^l \in \mathbb{R}^{\aleph \times 1}$, α_{tj}^l is the length attention score on the j -th length embedding at the t -th decoding step, h_t^d is the decoder hidden state, and V_l, W_l, w_r, b_l are learnable parameters. r_t is a scalar representing the remaining length at the current decoding step and \aleph is a hyperparameter indicating the number of pre-defined length embeddings.

For length embeddings, we adopt the positional encoding proposed in Vaswani et al. (2017). We

keep the embeddings fixed to remove the bias brought by the length distribution of data. The j -th length embedding l_j is defined as follows:

$$l_j = \begin{cases} 0 & j = 1 \\ PE(j-1) & otherwise, \end{cases} \quad (4)$$

where $PE(\cdot)$ is the positional encoding.

At the t -th decoding step, the remaining length r_t is updated by subtracting the length of the previously generated token. For the first decoding step, r_1 is initialized with desired output length. Following equations are used when $t > 1$:

$$r_t = \begin{cases} 0 & r_{t-1} - L(y_{t-1}) \leq 0 \\ r_{t-1} - L(y_{t-1}) & otherwise \end{cases} \quad (5)$$

where $L(y_{t-1})$ returns the number of characters in the output word y_{t-1} .

4 Experiment

4.1 Experimental Settings

We evaluate LenAtten on the CNN/Daily Mail dataset (See et al., 2017) to compare it with previous studies. In addition, we test LenAtten with short articles and summaries on the Annotated English Gigaword dataset (Rush et al., 2015). By default, all models are trained with maximum likelihood estimation (MLE) on a NVIDIA TITAN RTX GPU.²

For evaluation metrics, we adopt the standard F1 score of ROUGE-1, ROUGE-2, and ROUGE-L (Lin, 2004) to evaluate summary quality. For evaluating models' ability to control the output sequence length, we follow (Makino et al., 2019) to compute (1) character-level length variance Var between reference summaries and generated summaries and (2) over-length ratio $\%_{over}$, which measures how many of the generated summaries are longer than their reference summaries. The length variance Var is computed as follows:

$$Var = 0.001 * \frac{1}{n} \sum_{i=1}^n |len(y_i) - len(y'_i)|^2 \quad (6)$$

where y_i is the reference summary, y'_i is the generated summary, and $len(\cdot)$ returns the number of characters in the given summary. For the FLS task, the length variance Var is expected to be zero as it indicates the lengths of output summaries are exactly the desired summary lengths.

²Detailed model configurations are provided in the Appendix.

4.2 Methods to be compared

We compare the proposed Length Attention unit with following methods:

LEAD-3 extracts the first three sentences of source articles as the summary.

PG is the standard pointer-generator network proposed in See et al. (2017).

MASS (Song et al., 2019) is a sequence to sequence pre-trained model based on Transformer.

LenAtten is also compared with length controllable summarization methods. For a fair comparison, we choose methods that also aim at generating summaries with a preset number of characters in a word-by-word manner.

LE is the LenEmb method proposed in Kikuchi et al. (2016).

GOLC is a global optimization method introduced in Makino et al. (2019).

We apply LenAtten to three summarization models:

S2S (RNN-based Seq2Seq Model) is a vanilla encoder-decoder summarizer. Specifically, we adopt a Bi-LSTM as the encoder and a unidirectional LSTM as the decoder. To integrate LenAtten, the length context vector c_t^l is added to the input of the word prediction layer to produce the vocabulary distribution P_{vocab} :

$$P_{vocab} = softmax(W[h_t^d || c_t^l || y_{t-1} || C] + b) \quad (7)$$

where W , b are learnable parameters, h_t^d is the decoder hidden state, y_{t-1} is the word embedding of the last generated token, " $||$ " is the vector concatenation operator. C is the last encoder hidden state, which is known as the fixed context vector.

PAULUS (Copying Mechanism) follows the design of Paulus et al. (2018), which incorporates two attention modules and the copying mechanism into a Seq2seq summarizer. To integrate LenAtten, the vocabulary distribution P_{vocab} is calculated using:

$$P_{vocab} = softmax(W[h_t^d || c_t^l || c_t^e || c_t^d] + b) \quad (8)$$

where c_t^e and c_t^d are the context vectors generated from the encoder and decoder attention units.

ATTENTION (Attention-based model) is implemented by removing copying mechanism from **PAULUS**. For the above-mentioned three models, we remove the length context vector c_t^l in the ablation study and keep other components unchanged.

CNN/DM					
Models	R-1 F	R-2 F	R-L F	Var (\downarrow)	%over
Baseline					
*LEAD-3	40.34	17.70	36.57	-	-
†MASS	41.38	19.11	38.42	-	-
‡PG	37.74	15.78	34.35	19.35	58.35
‡PG + LE(MLE)	37.45	15.31	34.28	4.5	19.11
‡PG + LE(GOLC)	38.27	16.22	34.99	5.13	6.70
S2S	19.38	3.58	14.35	89.74	7.00
ATTENTION	34.32	13.76	28.92	48.99	18.11
PAULUS	38.10	16.42	33.17	19.91	37.14
with Length Attention ($\aleph = 2$)					
S2S	21.09	3.93	16.79	0.0069	30.67
ATTENTION	36.53	14.21	32.63	0.0075	38.72
PAULUS	39.82	17.31	36.20	0.0070	57.75

AEG					
Models	R-1 F	R-2 F	R-L F	Var (\downarrow)	%over
Baseline					
S2S	36.99	16.03	33.01	0.3902	30.12
ATTENTION	42.55	21.54	38.72	0.3285	25.89
PAULUS	43.84	22.80	40.12	0.3058	16.01
PAULUS+LE	40.02	17.31	36.99	0.0500	0.4
with Length Attention ($\aleph = 2$)					
S2S	38.26	16.24	35.11	0.0043	35.45
ATTENTION	43.15	21.51	40.32	0.0044	50.15
PAULUS	43.92	22.80	41.16	0.0042	37.75

Table 2: Results on CNN/DM and AEG dataset. If not specified in the parentheses, the training objective function is MLE by default. Results retrieved from: * See et al. (2017); † Xu et al. (2020); ‡ Makino et al. (2019).

4.3 Experimental Results³

Reference Summary Lengths In this experiment, we evaluate our model by comparing it with previous works. The desired length is set as the number of characters in corresponding reference summaries. Table 2 shows that LenAtten has superior length controllability and higher ROUGE scores on both datasets. Specifically, the length variance (Var) of LenAtten is 732 times better than the best-performing length controllable method PG+LE(GOLC) in the CNN/DM dataset. Besides, adding LenAtten can boost ROUGE scores by 1-3 points. We believe the improvement in the ROUGE scores comes from the introduction of length information (i.e. the desired length information). The desired length information can be viewed as an inductive bias, which helps summarizers prefer some of the outputs over others. Under the same context, by conditioning on the desired output length, summarizers may prefer candidate summaries with output lengths similar to the desired length. Thus, summarizers can learn a better alignment with the

³For all the experiments, the number following LA (the LenAtten unit) represents the number of length embeddings (i.e. the value of \aleph).

Models	CNN/DM	AEG
S2S	5.450	3.624
S2S+LA2	5.360	3.393
ATTENTION	4.317	3.279
ATTENTION+LA2	4.278	3.074
PAULUS	3.478	3.085
PAULUS+LA2	3.391	2.899

Table 3: Test perplexity of models on the CNN/DM dataset and the AEG dataset.

reference summaries during training and outputs summaries with higher ROUGE scores in inference.

In addition, previous length controllable methods control the output lengths at the cost of damaging the ROUGE scores. The ROUGE scores of PG and PAULUS drop after adding LenEmb (i.e. PG + LE(MLE) and PAULUS + LE). In comparison, LenAtten not only performs better at reducing the length variance Var but also significantly improves ROUGE scores. This suggests that LenAtten can break the trade-off between summary quality and length controllability.

After integrating with LenAtten, the %over ratio of summarizers rises. This suggests that more of the generated summaries ended up being longer than the references. We believe this is because when the remaining length is small (e.g. 4 characters) but not 0, instead of stopping the generation process, summarizers with LA tend to generate more tokens to meet the length requirement. Since summarizers output a word at each inference step, they may select a word that’s longer than the remaining length. Thus, the generated summaries may end up being longer than the references.

Perplexity To figure out how LenAtten affects the performance of the language model, we examine the log-perplexity of models on the test sets. Perplexity is a commonly-used metric for evaluating language models. A lower perplexity score indicates better language model performance. In this experiment, the desired length is set to the reference summary length. As shown in Table 3, after adding LenAtten, log-perplexity drops consistently on both datasets. This suggests that the adding of LenAtten can boost language model performance.

Various Preset Lengths In this experiment, we test the generalization ability of LenAtten under various desired lengths. For the AEG dataset, most reference summaries contain 30-75 characters, and

AEG	30	50	75	100	120
R-1 F	37.89	42.67	41.13	39.03	37.45
R-2 F	18.87	21.01	18.99	17.16	16.00
R-L F	33.64	38.76	35.27	31.95	29.84
Var	0.0027	0.0024	0.0026	0.0030	0.0040
CNN/DM	100	200	400	800	1600
R-1 F	30.88	37.90	39.52	37.25	34.23
R-2 F	13.31	16.26	16.58	15.08	13.50
R-L F	23.17	32.95	33.99	29.34	24.54
Var	0.0067	0.0063	0.0058	0.0054	0.0051

Table 4: ROUGE scores and Length Variance Var of PAULUS+LA2 under different desired lengths.

few of them are more than 100 characters. For the CNN/DM dataset, most reference summaries are 100-750 characters. Thus, the desired length is set as 30, 50, 75, 100, and 120 for the AEG dataset and 100, 200, 400, 800, and 1600 for the CNN/DM dataset. We add the LenAtten unit to PAULUS and exploit full reference summaries to get ROUGE scores.

As shown in Table 4, on the AEG dataset, for frequently appeared lengths (30, 50, 75), and lengths that are exceptionally long (100, 120), LenAtten demonstrates great length controllability along with good ROUGE scores. Same conclusions can be drawn on the CNN/DM dataset. This shows that LenAtten has great generalization ability under various desired lengths.

Exploring Hyperparameter \aleph In this experiment, we analyze how different \aleph (the number of pre-defined length embeddings) affect the performance of LenAtten on the AEG dataset. Desired lengths are set to the lengths of reference summaries. Figure 2 shows the length controllability becomes better as the increase of \aleph , with no harm to the ROUGE-L scores.

5 Conclusions

In this paper, we present a novel length controlling unit, LenAtten, to help summarization models generate quality summaries with a preset number of characters. On the examined datasets, LenAtten outperforms length controllable summarization baselines steadily in terms of length controllability and demonstrates great generalization ability. LenAtten also breaks the trade-off between length controllability and summary quality. To our knowledge, in the task of generating summaries with target lengths, LenAtten is the new state-of-the-art on the CNN Daily Mail dataset.

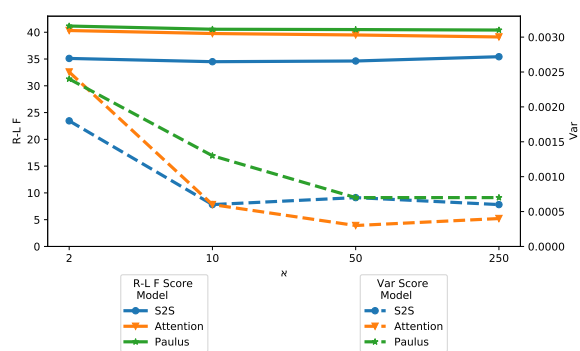


Figure 2: Examining hyperparameter \aleph on the AEG dataset. ROUGE-L F1 scores and Length Variance Var of different models under different \aleph are shown ($\aleph = 2, 10, 50, 250$).

Acknowledgment

We thank friends and colleagues at the UIC-AISIG as well as Minyong Li for their assistance with the study. We would like to thank anonymous reviewers for their helpful comments. This work is supported by the BNU-HKBU United International College research grant.

References

- Sumit Chopra, Michael Auli, and Alexander M. Rush. 2016. [Abstractive sentence summarization with attentive recurrent neural networks](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98, San Diego, California. Association for Computational Linguistics.
- Bonnie Dorr, David Zajic, and Richard Schwartz. 2003. [Hedge trimmer: A parse-and-trim approach to headline generation](#). In *Proceedings of the HLT-NAACL 03 Text Summarization Workshop*, pages 1–8.
- Angela Fan, David Grangier, and Michael Auli. 2018. [Controllable abstractive summarization](#). In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 45–54, Melbourne, Australia. Association for Computational Linguistics.
- 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.
- Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. 2016. [Controlling output length in neural encoder-decoders](#). In

- Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1328–1338, Austin, Texas. Association for Computational Linguistics.
- Chin-Yew Lin. 2004. **ROUGE: A package for automatic evaluation of summaries**. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Yang Liu and Mirella Lapata. 2019. **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)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.
- Yizhu Liu, Zhiyi Luo, and Kenny Zhu. 2018. **Controlling length in abstractive summarization using a convolutional neural network**. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4110–4119, Brussels, Belgium. Association for Computational Linguistics.
- Takuya Makino, Tomoya Iwakura, Hiroya Takamura, and Manabu Okumura. 2019. **Global optimization under length constraint for neural text summarization**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1039–1048, Florence, Italy. Association for Computational Linguistics.
- Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. **Summarunner: A recurrent neural network based sequence model for extractive summarization of documents**. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).
- Ramesh Nallapati, Bowen Zhou, Cícero Nogueira dos Santos, Çağlar Gülçehre, and Bing Xiang. 2016. **Abstractive text summarization using sequence-to-sequence rnns and beyond**. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*, pages 280–290. ACL.
- Romain Paulus, Caiming Xiong, and Richard Socher. 2018. **A deep reinforced model for abstractive summarization**. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.
- Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. **A neural attention model for abstractive sentence summarization**. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.
- Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. **Get to the point: Summarization with pointer-generator networks**. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.
- Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tiejun Liu. 2019. **MASS: masked sequence to sequence pre-training for language generation**. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5926–5936. PMLR.
- Sho Takase and Naoaki Okazaki. 2019. **Positional encoding to control output sequence length**. 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 3999–4004, Minneapolis, Minnesota. Association for Computational Linguistics.
- 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: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.
- 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.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. **Bertscore: Evaluating text generation with BERT**. *CoRR*, abs/1904.09675.
- Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. **Extractive summarization as text matching**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208, Online. Association for Computational Linguistics.

A Appendix

A.1 Experimental settings

Model Configuration In our experiments, the dimension of LSTM hidden state is set as 256, and the vocabulary size is 100,000. Word embeddings are fixed 300-dimensional GloVe vectors⁴. If a word is not covered in the GloVe, a random 300-dimensional vector is used. During training, Adam optimizer is applied with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$ and learning rate $\alpha = 0.001$. Besides, we set a 25% probability of choosing the previously generated token instead of the ground-truth token as y_{t-1} to reduce exposure bias. At test time, summaries are produced using beam search with beam size 4. We use a fully python implemented library⁵ to obtain the ROUGE score.

Dataset Distribution We plot the length distribution of reference summaries in the AEG (Figure 3) and the CNN/DM (Figure 4) dataset.

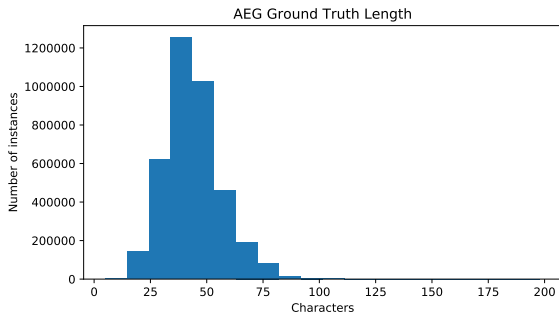


Figure 3: Length distribution of reference summaries on the Annotated English Gigaword dataset. Summaries with 30 to 75 characters cover the majority cases.

A.2 Additional Experiments

Semantic Similarity Another automatic evaluation metric *BertScore* (Zhang et al., 2019) recall score is used to measure the semantic similarity between system outputs and reference summaries. As shown in Figure 5, models with Length Attention module (LA2) outperform baselines (FREE) on both datasets.

Human Evaluation Correctness (CORR), completeness (COMP), and fluency (FLUE) of system outputs are assessed through 2 human evaluations. We randomly select 10 samples from each dataset.

⁴<http://nlp.stanford.edu/data/glove.6B.zip>

⁵<https://github.com/pltrdy/rouge>

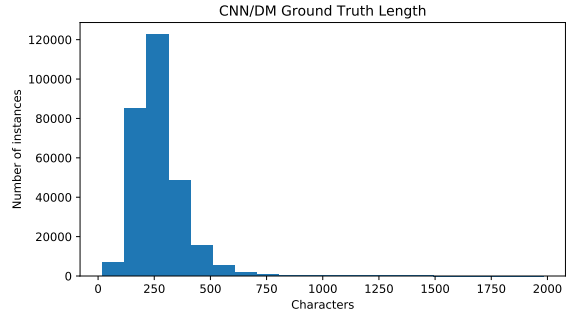


Figure 4: Length distribution of reference summaries on the CNN/Daily Mail dataset. Summaries exceed 2000 characters are ignored, since they only cover 0.009% of the dataset.

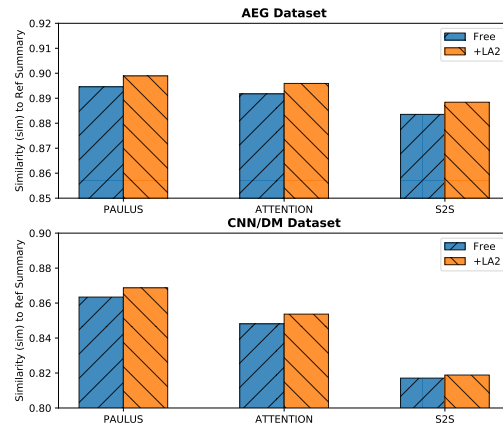


Figure 5: Semantic similarity between model outputs and reference summaries. Desired length is set as reference summary length.

30 skilled English speakers are presented with the original article and two summaries. One of the summaries is from the model without LenAtten, and the other one is from the same model plus LenAtten (e.g., S2S and S2S+LA2). The evaluation process is well-designed to prevent participants from knowing the source of the presented summaries. Models without LenAtten generate summaries without length restriction, and models with LenAtten are required to output summaries with desired lengths. There are 467 feedbacks collected for the first experiment and 160 for the second.

In the first experiment (Table 5), participants are asked to choose a better one from two given summaries. The desired length is set as the length of the reference summary.

In the second experiment (Table 6), we only examine PAULUS and PAULUS+LA2. The desired

	COMP	CORR	FLUE
AEG			
S2S (Free)	27.1%	44.4%	42.9%
S2S+LA2 (GT)	72.9%	55.6%	57.1%
ATTENTION (Free)	33.3%	44.2%	56.9%
ATTENTION+LA2 (GT)	66.7%	55.8%	43.1%
PAULUS (Free)	47.2%	53.3%	61.4%
PAULUS+LA2 (GT)	52.8%	46.7%	38.6%
CNN/DM			
S2S (Free)	22.5%	22.9%	53.8%
S2S+LA2 (GT)	77.5%	77.1%	46.2%
ATTENTION (Free)	21.9%	27.5%	36.7%
ATTENTION+LA2 (GT)	78.1%	72.5%	63.3%
PAULUS (Free)	53%	47.4%	48.1%
PAULUS+LA2 (GT)	47%	52.6%	51.9%

Table 5: Results of the first Human Evaluation. “(Free)”: model generates summaries freely. “(GT)”: model generates summaries with the desired length set as the length of the reference summary.

	COMP	CORR	FLUE
AEG			
PAULUS (Free, avg_len=57.46)	3.328	3.407	3.605
PAULUS+LA2 (30)	3.000	3.250	3.392
PAULUS+LA2 (50)	3.750	3.500	3.687
PAULUS+LA2 (70)	3.875	3.750	3.562
CNN/DM			
PAULUS (Free, avg_len=285.5)	3.250	3.345	3.273
PAULUS+LA2 (150)	3.125	3.375	3.166
PAULUS+LA2 (250)	3.451	3.483	3.419
PAULUS+LA2 (350)	3.827	3.482	3.172

Table 6: Results of the second human evaluation. “(Free, avg_len)”: model generates summaries freely. The average length of the generated summaries is also listed.

length is set as (30, 50, 70) on the AEG dataset and (150, 250, 350) on the CNN/DM dataset. Participants need to rate each summary from 0 to 5. In order to guarantee the accuracy and credibility of results, each article is presented once to each participant.

As shown in Table 5, models with LenAtten have better completeness and correctness scores on both datasets, along with a few improvements on the fluency. In the second experiment, Table 6 shows that (1) the completeness and correctness scores increase as the desired length increases. This trend is reasonable, since more information should be included in the final summary, as the summary length gets longer. This also suggests that, as the desired length gets longer, models with LenAtten can generate meaningful words instead of simply repeating one or two words. (2) When comparing the results of PAULUS (Free, 57.46; 285.5) and PAULUS+LA2 (50; 250), PAULUS+LA2 outperforms the PAULUS (Free) on all metrics. In other

Source document - A			
the indian union government thursday decided to increase customs duty on sugar to ## percent to curb cheap imports of the commodity , said a senior finance ministry official here .			
Reference summary - A			
india increases sugar import duty			
	Summary	R-1 R	Diff
Model	india to increase customs duty on sugar	60.00	4
Model + LA2	india to increase tariffs on sugar	40.00	0

Source document - B			
defending champion albert costa of spain reached the last eight in the french open here on monday , beating local favorite ##nd seed arnaud clement of france in straight sets .			
Reference summary - B			
costa enters last eight in french open			
	Summary	R-1 R	Diff
Model	costa into last eight in french open	85.71	-2
Model + LA2	costa into french open quarter-finals	42.85	-1

Table 7: Synonym substitution is colored in red. **R-1 R**: ROUGE-1 Recall score. **Diff**: len(output) - len(reference).

words, when the desired length gets smaller (but not too small), LenAtten can help summarization models to use concise words and phrases while maintaining summary quality.

A.3 Output Examples

Synonym substitution When examining generated summaries, we find adding LenAtten can make summarizers replace long/short words with synonyms to meet the length requirement. Examples are showcased in Table 7.