Hie-BART: Document Summarization with Hierarchical BART

Kazuki Akiyama
Ehime University
k_akiyama@ai.cs.ehime-u.ac.jp

Akihiro Tamura
Doshisha University
aktamura@mail.doshisha.ac.jp

Takashi Ninomiya
Ehime University
ninomiya@cs.ehime-u.ac.jp

Abstract

This paper proposes a new abstractive summarization model for documents, hierarchical BART (Hie-BART), which captures the hierarchical structures of documents (i.e., their sentence-word structures) in the BART model. Although the existing BART model has achieved state-of-the-art performance on document summarization tasks, it does not account for interactions between sentence-level and word-level information. In machine translation tasks, the performance of neural machine translation models can be improved with the incorporation of multi-granularity self-attention (MG-SA), which captures relationships between words and phrases. Inspired by previous work, the proposed Hie-BART model incorporates MG-SA into the encoder of the BART model for capturing sentence-word structures. Evaluations performed on the CNN/Daily Mail dataset show that the proposed Hie-BART model outperforms strong baselines and improves the performance of a non-hierarchical BART model (+0.23 ROUGE-L).

1 Introduction

In recent years, improvements to abstractive document summarization models have been developed through the incorporation of pre-training. The BERTSUM model (Liu and Lapata, 2019) has been proposed as a pre-training model for document summarization tasks. For sequence-to-sequence tasks, the T5 model (Raffel et al., 2020) and the BART model (Lewis et al., 2020) have been proposed as part of generalized pre-training models. Among the existing pre-training models, the BART model achieves state-of-the-art performance on document summarization tasks. However, the BART model does not capture the hierarchical structures of documents when generating a summary.

Neural machine translation has been improved by the capture of multiple granularities of information in input texts such as "phrases and words" and "words and characters". In particular, Transformer-based machine translation model has been improved by incorporating multi-granularity self-attention (MG-SA) (Hao et al., 2019), which considers the relationships between words and phrases by decomposing an input text into its elements using multiple granularity (i.e., words and phrases) and assigning each granular element (i.e., a word or a phrase) to a head in multi-head Self-Attention Networks (SANs). This method enables interactions not only between words but also between phrases and words, through self-attentions.

Inspired by previous work, this paper proposes a new abstractive document summarization model, hierarchical BART (Hie-BART), which captures a document's hierarchical structures (i.e., sentence-word structures) through the SANs of the BART model. Here, a document is divided into elements with word-level and sentence-level granularity, where each element is assigned to a head of the SANs layers of the BART encoder. Then, information with multi-granularity is captured by combining the output of the SANs layers, where the ratio of combining word-level and sentence-level information is controlled by a hyperparameter.

We evaluated the proposed model in an abstractive summarization task with the CNN/Daily Mail dataset. Our evaluation shows that our Hie-BART model improves the F-score of ROUGE-L by 0.23 points relative to the non-hierarchical BART model, and the proposed model is better than the strong baselines, BERTSUM and T5 models.

2 Background

2.1 BART

The BART model (Lewis et al., 2020) is a generalized pre-training model based on the Transformer

Proceedings of NAACL-HLT 2021: Student Research Workshop, pages 159–165
June 6–11, 2021.

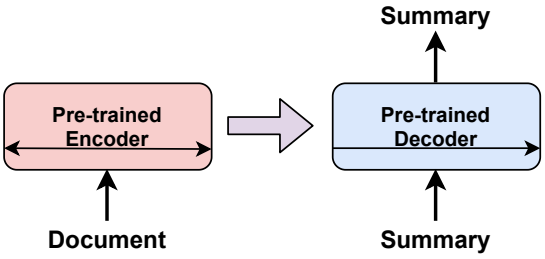

Figure 1: The overview architecture of BART. The encoder is a bidirectional model and the decoder is an autoregressive model.

model (Vaswani et al., 2017). Five pre-training techniques are introduced: token masking, sentence permutation, document rotation, token deletion, and text infilling.

Each of these is a denoising autoencoder technique that adds noise to the original text and restores the original text. **Token masking**, as used in BERT (Devlin et al., 2019), randomly masks tokens. **Sentence permutation** randomly shuffles the sentences in a document. **Document rotation** randomly selects a token from a sentence and then rotates the sentence so that it begins with that token. **Token deletion** randomly deletes a token from the original sentence. **Text infilling** replaces word sequences with a single mask token or inserts a mask token into a randomly selected position. A combination of sentence permutation and text infilling achieves the best accuracy of all techniques.

An overview of the BART model is given in Figure 1. The encoder is a bidirectional model and the decoder is an autoregressive model. This pre-trained BART model is fine-tuned to various tasks, such as the summarization task, for which, a document is provided to the encoder, and the decoder generates a document summary.

2.2 Multi-Granularity Self-Attention (MG-SA)

MG-SA (Hao et al., 2019) is used to capture multi-granularity information from an input text by dividing the input into elements with several types of granularity and preparing heads of multi-head SANs for each type of granularity. Provided with the word-level matrix H, which is an input to the SANs, this method first generates a phrase-level matrix H_g representing phrase-level information, as follows:

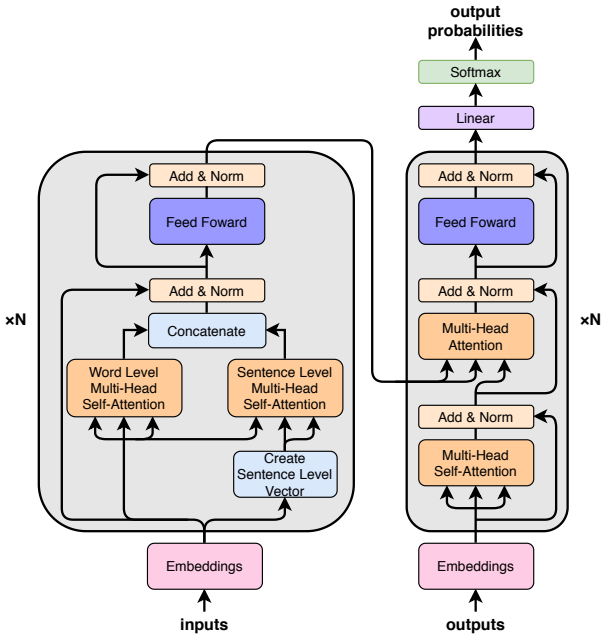

Figure 2: Overview architecture of Hie-BART. This is based on the Transformer model. The SANs in the encoder are divided into word and sentence levels and computed.

$$H_g = F_h(H),$$

where $F_h(\bullet)$ is a function that generates a phrase-level matrix for the h-th head. Specifically, a phrase-level matrix is generated by running a max pooling operation on word-level vectors in a word-level matrix. After a phrase-level matrix is generated, SANs perform the following computations:

$$Q^h, K^h, V^h = HW_Q^h, H_g W_K^h, H_g W_V^h, \quad (1)$$

$$O^h = \mathbf{ATT}(Q^h, K^h)V^h, \quad (2)$$

where $Q^h \in \mathbb{R}^{n \times d_h}$, $K^h \in \mathbb{R}^{p \times d_h}$, $V^h \in \mathbb{R}^{p \times d_h}$ are respectively the query, key, and value representations, $W_Q^h, W_K^h, W_V^h \in \mathbb{R}^{d \times d_h}$ are parameter matrices, and d, d_h, n, and p are the dimensions of the hidden layer, one head, a word vector, and a phrase vector, respectively. In addition, **ATT(X,Y)** is a function that calculates the attention weights of X and Y. From these computations, the output O^h of each head in the SANs is generated. Then, the output of MG-SA is generated by concatenating the outputs from all heads: $\mathbf{MG\text{-}SA(H)} = [\mathbf{O^1}, ..., \mathbf{O^N}]$. The outputs of each head O^h contain information between words or between words and phrases. Thus, in addition to relationships between words, the relationships between words and phrases can be captured with MG-SA.

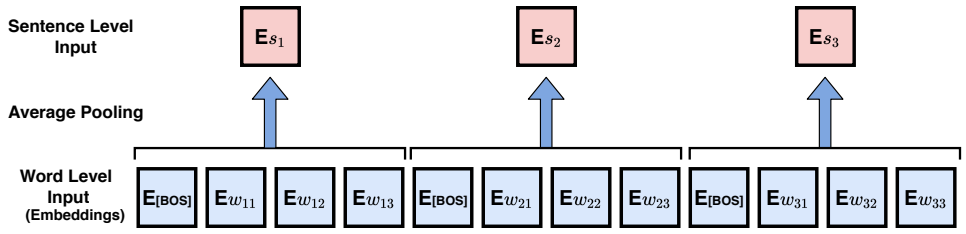

Figure 3: Behavior of the create sentence level vector layer. $E_{w_{ij}}$ and $E_{[BOS]}$ are embedded vectors for the word w_{ij} (j-th word in i-th sentence) and $[BOS]$ token, respectively. E_{s_i} is the sentence-level embedded vector for the i-th sentence s_i.

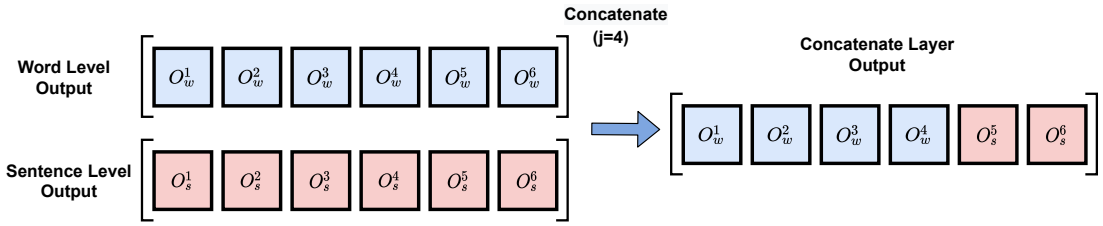

Figure 4: An example of the behavior of the concatenate layer where the number of heads of the multi-head is 6 and the join point $j = 4$. The blue $[O_w^1, ..., O_w^6]$ designates the outputs of the word-level SANs and the red $[O_s^1, ..., O_s^6]$ shows the outputs of the sentence-level SANs.

3 Hie-BART

3.1 Architecture

The Hie-BART (Hierarchical-BART) model has a sentence-to-word (sentence-level) SANs in addition to the word-to-word (word-level) SANs of the original BART model. An overview of Hie-BART is shown in Figure 2. Hie-BART has sentence-level SANs, a create sentence level vector layer and a concatenate layer, in addition to BART. In the create sentence level vector layer, a sentence-level matrix is created from a word-level matrix. The concatenate layer concatenates the outputs of word-level and sentence-level SANs. The outputs of the concatenate layer are forwarded to the subsequent feed-forward layer. To provide boundary information between the sentences, each sentence is prefixed with a $[BOS]$ token.

3.2 Create Sentence Level Vector Layer

The behavior of the create sentence level vector layer is shown in Figure 3. $E_{w_{ij}}$ and $E_{[BOS]}$ are embedded vectors for word w_{ij} (j-th word in the i-th sentence) and $[BOS]$ token, respectively. E_{s_i} is the sentence-level embedded vector for the i-th sentence s_i.

The create sentence level vector layer uses average pooling to generate a sentence-level vector from word-level vectors. Given the word sequence $W = (w_1, ..., w_N)$, it is divided into sentences $S = (s_1, ..., s_M)$, where N is the total number of words, M is the total number of sentences, and each s_i is the i-th sentence consisting of a word subsequence $w_{i1}, \ldots w_{iN_i}$, where N_i is the total number of words in the sentence. For each element of S, we apply average pooling as follows: $g_m = \mathbf{AVG}(s_m)$, where the $\mathbf{AVG}(\bullet)$ is average pooling. From this formula, $G = (g_1, ..., g_M)$ is generated. Each element of W, S, and G is an embedded vector. G is forwarded to the sentence-level SANs as its input.

3.3 Concatenate Layer

The outputs of each of the word-level and sentence-level SANs are combined in the concatenate layer. The outputs of the word-level and sentence-level SANs layer are as follows:

$$\mathbf{SANs(W)} = [O_w^1, \ldots, O_w^H] = O_w^{ALL}, \quad (3)$$

$$\mathbf{SANs(G)} = [O_s^1, \ldots, O_s^H] = O_s^{ALL}, \quad (4)$$

where H is the number of heads, $[O_w^1, \ldots, O_w^H] = O_w^{ALL}$ is the output of the word-level SANs, consisting of the word-level head's outputs, and $[O_s^1, ..., O_s^H] = O_s^{ALL}$ is the output of the sentence-level SANs, consisting of the sentence-level head's outputs. The outputs of these word/sentence-level SANs are combined as follows:

$$\mathbf{CONCAT}(\mathbf{O_w^{ALL}}, \mathbf{O_s^{ALL}}, \mathbf{j}) = [\mathbf{O_w^1}, ..., \mathbf{O_w^j}, \mathbf{O_s^{j+1}}, ..., \mathbf{O_s^H}], \quad (5)$$

Model	ROUGE-1	ROUGE-2	ROUGE-L
LEAD-3 (Nallapati et al., 2017)	40.42	17.62	36.67
PTGEN (See et al., 2017)	36.44	15.66	33.42
PTGEN+COV (See et al., 2017)	39.53	17.28	36.38
BERTSUMEXTABS (Liu and Lapata, 2019)	42.13	19.60	39.18
T5 (Raffel et al., 2020)	43.52	21.55	40.69
BART (Lewis et al., 2020)	44.16	21.28	40.90
BART (ours)	44.06	21.22	40.82
Hie-BART (ours)	**44.35***,**	**21.37**	**41.05****

Table 1: Results on the CNN/Daily Mail test set.

Word : Sentence	ROUGE		
	1	2	L
16 : 0	44.72	21.73	41.43
15 : 1	44.95	**21.92**	41.68
14 : 2	**45.01**	**21.92**	**41.75**
13 : 3	44.91	21.87	41.64
12 : 4	44.74	21.66	41.49
11 : 5	44.88	21.81	41.62
10 : 6	44.78	21.75	41.51
9 : 7	44.70	21.71	41.46
8 : 8	44.79	21.77	41.58

Table 2: Results on the CNN/Daily Mail validation set. The leftmost column shows the ratio of the number of multi-heads to combine. The highest score was achieved for the ratio "Word:Sentence = 14:2".

where **CONCAT(X, Y, j)** is a function that concatenates X and Y at the join point j of the multi-heads. In the combined multi-head, the heads from 1 to j are word-level outputs, and the heads from $j+1$ to H are sentence-level outputs.

Figure 4 shows an example of the behavior of the concatenate layer in Hie-BART, where the number of heads of the multi-head is 6 and the join point $j = 4$. The output of the word-level SANs $[O_w^1, ..., O_w^6]$ and the output of the sentence-level SANs $[O_s^1, ..., O_s^6]$ are joined at the join point $j = 4$, resulting in the output $[O_w^1, O_w^2, O_w^3, O_w^4, O_s^5, O_s^6]$.

The output of the concatenate layer is forwarded to the feed-forward layer in the encoder.

4 Experiments

4.1 Dataset

We used the CNN/Daily Mail dataset[1] (Hermann et al., 2015), a summary corpus of English news articles, consisting of 287,226 training pairs, 13,368 validation pairs, and 11,490 test pairs. On average, the source documents and summary sentences have 781 and 56 tokens, respectively. For data preprocessing, we followed the instruction provided in the CNN/Daily Mail dataset[1] and fairseq[2].

4.2 Parameters

We used the pre-trained BART model "bart.large", provided in fairseq[2] for Hie-BART. The hyper-parameters for BART and Hie-BART were determined for the validation set; the gradient accumulation parameter (update-freq) was 10, the total number of training steps was 20,000, and the number of multi-heads was set to 16. The ratio of the number of combined heads of output in word-level and sentence-level SANs was set to "Word:Sentence = 14:2" for Hie-BART. We followed fairseq's settings[2] for the other hyperparameters. In our environments, the model had 406,291,456 parameters for Hie-BART and 406,290,432 parameters for BART.

4.3 Results

The results of the evaluation of the Hie-BART model for the CNN/Daily Mail test data are shown in Table 1. The F-score for ROUGE-1, ROUGE-2, and ROUGE-L (Lin, 2004) were used as evaluation metrics. To calculate the ROUGE score, we

[1]CNN/Daily Mail dataset:
https://github.com/abisee/cnn-dailymail

[2] Usage of BART by faireseq:
https://github.com/pytorch/fairseq/tree/master/examples/bart

[Source Document] (CNN)About a dozen Native American actors have walked off the set of an Adam Sandler movie comedy, saying the satirical Western's script is insulting to Native Americans and women, according to a report. ... According to ICTMN, a Native American adviser hired to help ensure the movie's cultural authenticity also walked off the set in protest. Hill, the Choctaw actor, seemed to hold out hope that differences between the producers and Native American cast members could be resolved . ... .
[Baseline Model's Summary] The walkout occurred on the set of "The Ridiculous Six" near Las Vegas, New Mexico, according to a report . The script called for native women's names such as "Beaver's Breath" and "No Bra" and an actress portraying an Apache woman to squat and urinate while smoking a peace pipe, ICTMN reports .
[Proposed Model's Summary] About a dozen Native American actors walked off the set of an Adam Sandler movie comedy, a report says . They say the script is insulting to Native Americans and women, according to the Indian Country Today Media Network . "The Ridiculous Six" also stars Will Forte, Taylor Lautner, Steve Buscemi and John Turturro .
[Gold Summary] About a dozen Native American actors walk off set of Adam Sandler comedy, says report . Actors say satirical Western's script is insulting to Native Americans and women .

Table 3: An example of improved summaries. Note that the source document is partially omitted.

[Source Document] (CNN)It's the kind of thing you see in movies, like Robert Redford's role in "All Is Lost" or Ang Lee's "Life of Pi." But, in real life, it's hard to swallow the idea of a single person being stranded at sea for days, weeks, if not months and somehow living to talk about it. Miracles do happen, though, and not just in Hollywood. We're not talking about people who float aimlessly or run aground after running out of gas or being let down by faltering winds only to be picked up a few hours later by the U.S. Coast Guard. ... After being spotted by crew aboard a Japanese Maritime Self-Defense Force destroyer and picked up in a smaller rescue boat, he took a drink offered to him and burst into tears, Kyodo reported. Shinkawa told his rescuers, "I thought today was the last day of my life."
[Base Model's Summary] It's hard to swallow the idea of a single person being stranded at sea for days, weeks, if not months and somehow living to talk about it . Miracles do happen, though, and not just in Hollywood . Jose Salvador Alvarenga says his journey began in Paredon Viejo, a port on Mexico's Pacific coast in late 2012 .
[Proposed Model's Summary] It's hard to swallow the idea of a single person being stranded at sea for days, weeks, if not months and somehow living to talk about it . Miracles do happen, though, and not just in Hollywood . Jose Salvador Alvarenga says his journey began in Paredon Viejo, a port on Mexico's Pacific coast, in late 2012 . Louis Jordan says he used laundry to trap and scoop up fish, rigged a makeshift mast and sail .
[Gold Summary] A South Carolina man says he spent 66 days alone at sea before being rescued . Other sole survivor stories include a Japanese man washed away by a tsunami . An El Salvador man says he drifted from Mexico to Marshall Islands over a year .

Table 4: An example of summaries that are not improved.

used files2rouge[3]. Hie-BART was compared with LEAD-3 (Nallapati et al., 2017), PTGEN, PTGEN+COV (See et al., 2017), BERTSUMEXTABS (Liu and Lapata, 2019), T5 (Raffel et al., 2020), BART with our environment, and BART with Lewis et al. (2020). The **LEAD-3** method uses the first three sentences of the source document as a summary. **PTGEN** is a sequence-to-sequence model that incorporates a pointer generator network. **PTGEN+COV** introduces the coverage mechanism into PTGEN. **BERTSUMEXTABS** is a pre-training model that adapts BERT for summarization tasks. **T5** is a generalized pre-training model for sequence-to-sequence tasks based on the Transformer model. The statistical significance test was performed by the Wilcoxon-Mann-Whitney test. In Table 1, * and ** indicate that the comparisons with BART (ours) are statistically significant at 5% significance level and 10% significance level, respectively.

Hie-BART improved the F-score of ROUGE-1/2/L by 0.223 points on average relative to BART with our environment, and by 0.143 points on average from BART reported in (Lewis et al., 2020). Table 1 also shows that our Hie-BART model significantly improved ROUGE-1 and ROUGE-L scores of the baseline BART model.

4.4 Analysis

Table 2 shows a comparison of ROUGE scores for the ratio of the number of multi-heads at the word and sentence levels with the validation set of the CNN/Daily Mail dataset. The leftmost column shows the ratio of the number of multi-heads to combine. As can be seen in Table 2, the maximum ROUGE-1/2/L score was achieved for "Word:Sentence = 14:2". In ROUGE-1/2/L, smaller ratios of multi-heads at the sentence level that are compared to the word level, the higher the score tends to be. However, when the number of multi-heads at the sentence level is 0 (the original BART), the accuracy is lower than that of Hie-BART.

Table 3 shows an improved example of summaries: summaries generated by the baseline model (BART) and the proposed model (Hie-BART), and the gold summary. As can be seen in Table 3, the summary of the proposed model includes the important parts of the source document.

Table 4 shows an example of summaries that are not improved. In this example, the baseline model's summary and the proposed model's summary include almost the same contents, but they are far from and longer than the gold summary.

5 Conclusion

In this study, we proposed Hie-BART to can take into account the relationship between words and sentences in BART by dividing the self-attention layer of encoder into word and sentence levels. In the experiments, we confirmed that Hie-BART improved the F-score of ROUGE-L by 0.23 points relative to the non-hierarchical BART model, and the proposed model was better than the strong baselines, BERTSUM and T5 models for the CNN/Daily Mail dataset.

As future work, we intend to investigate methods to incorporate information between sentences in addition to word-to-word and word-to-sentence information.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jie Hao, Xing Wang, Shuming Shi, Jinfeng Zhang, and Zhaopeng Tu. 2019. Multi-granularity self-attention for neural machine translation. 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 887–897, Hong Kong, China. Association for Computational Linguistics.

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

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre–training for natural language generation, translation,

[3]files2rouge usage : https://github.com/pltrdy/files2rouge

and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880. 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.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. SummaRuNNer: A recurrent neural network based sequence model for extractive. In *Proceedings of In Association for the Advancement of Artificial Intelligence*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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.

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