Coreference-Aware Dialogue Summarization

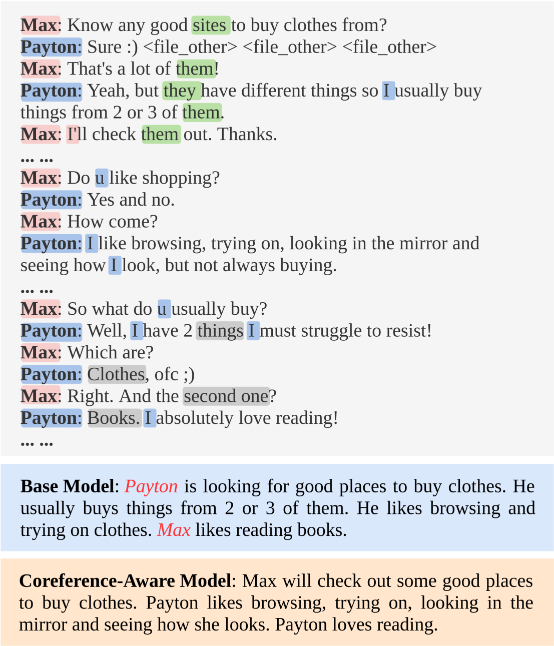
Zhengyuan Liu, Ke Shi, Nancy F. Chen
Institute for Infocomm Research, A*STAR, Singapore
{liu_zhengyuan,shi_ke,nfychen}@i2r.a-star.edu.sg

Abstract

Summarizing conversations via neural approaches has been gaining research traction lately, yet it is still challenging to obtain practical solutions. Examples of such challenges include unstructured information exchange in dialogues, informal interactions between speakers, and dynamic role changes of speakers as the dialogue evolves. Many of such challenges result in complex coreference links. Therefore, in this work, we investigate different approaches to explicitly incorporate coreference information in neural abstractive dialogue summarization models to tackle the aforementioned challenges. Experimental results show that the proposed approaches achieve state-of-the-art performance, implying it is useful to utilize coreference information in dialogue summarization. Evaluation results on factual correctness suggest such coreference-aware models are better at tracing the information flow among interlocutors and associating accurate status/actions with the corresponding interlocutors and person mentions.

1 Introduction

Text summarization condenses the source content into a shorter version while retaining essential and informative content. Most prior work focuses on summarizing well-organized single-speaker content such as news articles (Hermann et al., 2015) and encyclopedia documents (Liu* et al., 2018). Recently, models applied on text summarization benefit favorably from sophisticated neural architectures and pre-trained contextualized language backbones: on the popular benchmark corpus CNN/Daily Mail (Hermann et al., 2015), Liu and Lapata (2019) explored fine-tuning BERT (Devlin et al., 2019) to achieve state-of-the-art performance for extractive news summarization, and BART (Lewis et al., 2020) has also improved generation quality on abstractive summarization.



Max: Know any good sites to buy clothes from?
Payton: Sure :) <file_other> <file_other> <file_other>
Max: That's a lot of them!
Payton: Yeah, but they have different things so I usually buy things from 2 or 3 of them.
Max: I'll check them out. Thanks.
... ...
Max: Do u like shopping?
Payton: Yes and no.
Max: How come?
Payton: I like browsing, trying on, looking in the mirror and seeing how I look, but not always buying.
... ...
Max: So what do u usually buy?
Payton: Well, I have 2 things I must struggle to resist!
Max: Which are?
Payton: Clothes, ofc ;)
Max: Right. And the second one?
Payton: Books. I absolutely love reading!
... ...

Base Model: *Payton* is looking for good places to buy clothes. He usually buys things from 2 or 3 of them. He likes browsing and trying on clothes. *Max* likes reading books.

Coreference-Aware Model: Max will check out some good places to buy clothes. Payton likes browsing, trying on, looking in the mirror and seeing how she looks. Payton loves reading.

Figure 1: An example of dialogue summarization: The original conversation (in grey) is abbreviated; the summary generated by a baseline model is in blue; the summary generated by a coreference-aware model is in orange. While these two summaries obtain similar ROUGE scores, the summary from the baseline model is not factually correct; errors are highlighted in italic and magenta.

While there has been substantial progress on document summarization, dialogue summarization has received less attention. Unlike documents, conversations are interactions among multiple speakers, they are less structured and are interspersed with more informal linguistic usage (Sacks et al., 1978). Based on the characteristics of human-to-human conversations (Jurafsky and Martin, 2008), challenges of summarizing dialogues stem from: (1) Multiple speakers: the interactive information exchange among interlocutors implies that essential information is referred to back and forth across speakers and dialogue turns; (2) Speaker role shifting: multi-turn dialogues often involve frequent role shifting from one type of interlocutor to another type (e.g., questioner becomes responder and vice versa); (3) Ubiquitous referring expressions: aside from speakers referring to themselves and each other, speakers also mention third-party persons, concepts, and objects. Moreover, referring could also take on forms such as anaphora or cataphora where pronouns are used, making coreference chains more elusive to track. Figure 1 shows one dialogue example: two speakers exchange information among interactive turns, where the pronoun *"them"* is used multiple times, referring to the word *"sites"*. Without sufficient understanding of the coreference information, the base summarizer fails to link mentions with their antecedents, and produces an incorrect description (highlighted in magenta and italic) in the generation. From the aforementioned linguistic characteristics, dialogues possess multiple inherent sources of complex coreference, motivating us to explicitly consider coreference information for dialogue summarization to more appropriately model the context, to more dynamically track the interactive information flow throughout a conversation, and to enable the potential of multi-hop dialogue reasoning.

Previous work on dialogue summarization focuses on modeling conversation topics or dialogue acts (Goo and Chen, 2018; Liu et al., 2019; Li et al., 2019; Chen and Yang, 2020). Few, if any, leverage on features from coreference information explicitly. On the other hand, large-scale pre-trained language models are shown only to implicitly model lower-level linguistic knowledge such as part-of-speech and syntactic structure (Tenney et al., 2019; Jawahar et al., 2019). Without directly training on tasks that provide specific and explicit linguistic annotation such as coreference resolution or semantics-related reasoning, model performance remains subpar for language generation tasks (Dasigi et al., 2019). Therefore, in this paper, we propose to improve abstractive dialogue summarization by explicitly incorporating coreference information. Since entities are linked to each other in coreference chains, we postulate adding a graph neural layer could readily characterize the underlying structure, thus enhancing contextualized representation. We further explore two parameter-efficient approaches: one with an additional coreference-guided attention layer, and the other resourcefully enhancing BART's limited coreference resolution capabilities by conducting probing analysis to augment our coreference injection design.

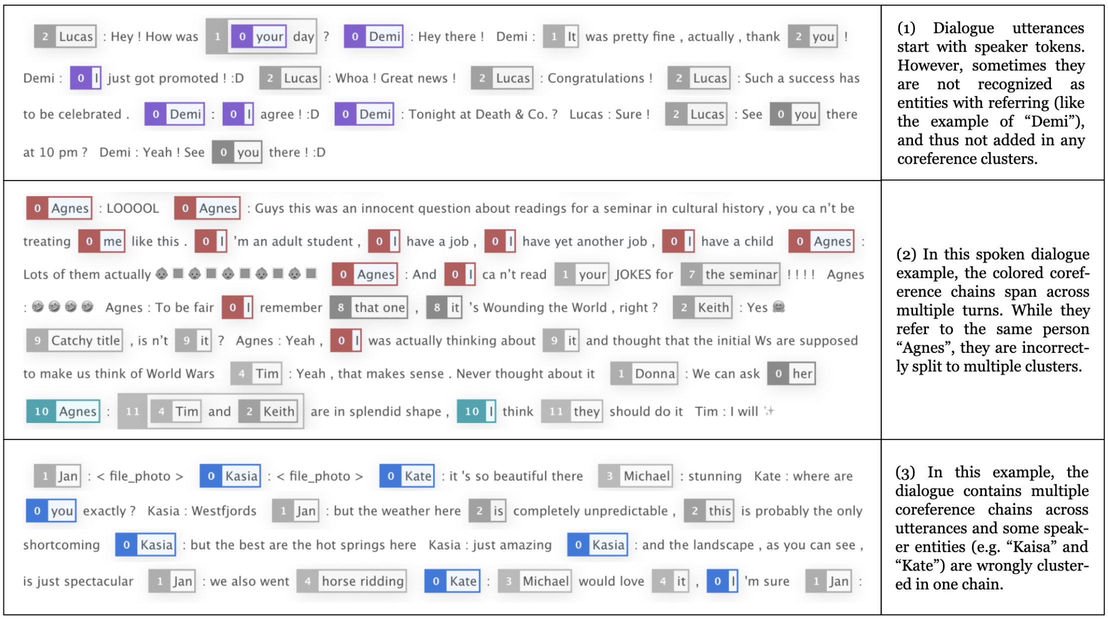
Experiments on SAMSum (Gliwa et al., 2019) show that the proposed methods achieve state-of-the-art performance. Furthermore, human evaluation and error analysis suggest our models generate more factually consistent summaries. As shown in Figure 1, a model guided with coreference information accurately associates events with their corresponding subjects, and generates more trustworthy summaries compared with the baseline.

2 Related Work

In abstractive text summarization, recent studies mainly focus on neural approaches. Rush et al. (2015) proposed an attention-based neural summarizer with sequence-to-sequence generation. Pointer-generator networks (See et al., 2017) were designed to directly copy words from the source content, which resolved out-of-vocabulary issues. Liu and Lapata (2019) leveraged the pre-trained language model BERT (Devlin et al., 2019) on both extractive and abstractive summarization. Lewis et al. (2020) proposed BART, taking advantage of the bi-directional encoder in BERT and the auto-regressive decoder of GPT (Radford et al., 2018) to obtain impressive results on language generation.

While many prior studies focus on summarizing well-organized text such as news articles (Hermann et al., 2015), dialogue summarization has been gaining traction. Shang et al. (2018) proposed an unsupervised multi-sentence compression method for meeting summarization. Goo and Chen (2018) introduced a sentence-gated mechanism to grasp the relations between dialogue acts. Liu et al. (2019) proposed to utilize topic segmentation and turn-level information (Liu and Chen, 2019) for conversational tasks. Zhao et al. (2019) proposed a neural model with a hierarchical encoder and a reinforced decoder to generate meeting summaries. Chen and Yang (2020) used diverse conversational structures like topic segments and conversational stages to design a multi-view summarizer, and achieved the current state-of-the-art performance on the SAMSum corpus (Gliwa et al., 2019).

Improving factual correctness has received keen attention in neural abstractive summarization lately. Cao et al. (2018) leveraged on dependency parsing and open information extraction to enhance the reliability of generated summaries. Zhu et al. (2021) proposed a factual corrector model based on



2 Lucas : Hey ! How was 1 0 your day ? 0 Demi : Hey there ! Demi : 1 It was pretty fine , actually , thank 2 you ! Demi : 0 I just got promoted ! :D 2 Lucas : Whoa ! Great news ! 2 Lucas : Congratulations ! 2 Lucas : Such a success has to be celebrated . 0 Demi : 0 I agree ! :D 0 Demi : Tonight at Death & Co. ? Lucas : Sure ! 2 Lucas : See 0 you there at 10 pm ? Demi : Yeah ! See 0 you there ! :D	(1) Dialogue utterances start with speaker tokens. However, sometimes they are not recognized as entities with referring (like the example of "Demi"), and thus not added in any coreference clusters.
0 Agnes : LOOOOL 0 Agnes : Guys this was an innocent question about readings for a seminar in cultural history , you ca n't be treating 0 me like this . 0 I 'm an adult student , 0 I have a job , 0 I have yet another job , 0 I have a child 0 Agnes : Lots of them actually 👶 ■ 👶 ■ 👶 ■ 👶 ■ 👶 ■ 0 Agnes : And 0 I ca n't read 1 your JOKES for 7 the seminar ! ! ! ! Agnes : 🤣 🤣 🤣 🤣 Agnes : To be fair 0 I remember 8 that one , 8 it 's Wounding the World , right ? 2 Keith : Yes 😭 9 Catchy title , is n't 9 it ? Agnes : Yeah , 0 I was actually thinking about 9 it and thought that the initial Ws are supposed to make us think of World Wars 4 Tim : Yeah , that makes sense . Never thought about it 1 Donna : We can ask 0 her 10 Agnes : 11 4 Tim and 2 Keith are in splendid shape , 10 I think 11 they should do it Tim : I will ✨	(2) In this spoken dialogue example, the colored coreference chains span across multiple turns. While they refer to the same person "Agnes", they are incorrectly split to multiple clusters.
1 Jan : < file_photo > 0 Kasia : < file_photo > 0 Kate : it 's so beautiful there 3 Michael : stunning Kate : where are 0 you exactly ? Kasia : Westfjords 1 Jan : but the weather here 2 is completely unpredictable , 2 this is probably the only shortcoming 0 Kasia : but the best are the hot springs here Kasia : just amazing 0 Kasia : and the landscape , as you can see , is just spectacular 1 Jan : we also went 4 horse ridding 0 Kate : 3 Michael would love 4 it , 0 I 'm sure 1 Jan :	(3) In this example, the dialogue contains multiple coreference chains across utterances and some speaker entities (e.g. "Kaisa" and "Kate") are wrongly clustered in one chain.

Figure 2: Examples of three common issues in adopting a document coreference resolution model for dialogues without additional domain adaptation training. Spans in blocks are items in coreference clusters with their cluster ID number. We highlight some spans for better readability.

knowledge graphs, significantly improving factual correctness in text summarization.

3 Dialogue Coreference Resolution

Since the common summarization datasets do not contain coreference annotations, automatic coreference resolution is needed to process the samples. Neural approaches (Joshi et al., 2020) have shown impressive performance on document coreference resolution. However, they are still sub-optimal for conversational scenarios (Chen et al., 2017), and there are no large-scale annotated dialogue corpora for transfer learning. When applying a document coreference resolution model (Lee et al., 2018; Joshi et al., 2020) on dialogue samples without domain adaptation,[1] as shown in Figure 2, we observed some common issues: (1) Each dialogue utterance starts with a speaker, but sometimes this speaker is not recognized as a coreference-related entity, and thus not added in any coreference clusters; (2) In dialogues, coreference chains are often spanned across multiple turns, but sometimes they are split to multiple clusters; (3) When a dialogue contains multiple coreference chain across multi-turns, speaker entities could be wrongly clustered.

Based on the observation, to improve the overall quality of dialogue coreference resolution, we conducted data post-processing on the automatic output: (1) First, we applied a model ensemble strategy to obtain more accurate cluster predictions; (2) Then, we re-assigned coreference cluster labels to the words with speaker roles that were not included in any chains; (3) Moreover, we compared the clusters and merged those that presented the same coreference chain. Human evaluation on the processed data showed that this post-processing reduced incorrect coreference assignments by approximately 19%.[2]

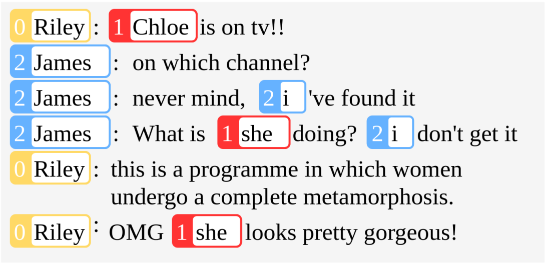
4 Coreference-Aware Summarization

In this section, we adopt a neural model for abstractive dialogue summarization, and investigate various methods to enhance it with the coreference information obtained in Section 3.

The base neural architecture is a sequence-to-sequence model Transformer (Vaswani et al., 2017). Given a conversation containing n tokens $T = \{t_1, t_2, ..., t_n\}$, a self-attention-based encoder is used to produce the contextualized hidden representations $H = \{h_1, h_2, ..., h_n\}$, then an auto-regressive decoder generates the target sequence $O = \{w_1, w_2, ..., w_k\}$ sequentially. Here, we use BART (Lewis et al., 2020) as the pre-trained lan-

[1] The off-the-shelf version of coreference resolution model we used is *allennlp-public-models/coref-spanbert-large-2021.03.10*, which is trained on OntoNotes 5.0 dataset.

[2] In our pilot experiment, we observed that models with original coreference resolution outputs showed 10% relative lower performance than that with the optimized data, validating the effectiveness of our post-processing.



0 Riley : 1 Chloe is on tv!!
2 James : on which channel?
2 James : never mind, 2 i 've found it
2 James : What is 1 she doing? 2 i don't get it
0 Riley : this is a programme in which women undergo a complete metamorphosis.
0 Riley : OMG 1 she looks pretty gorgeous!

Figure 3: One dialogue example with labeled coreference clusters: there are three coreference clusters in this conversation, where each cluster contains all mentions of one personal identity.

guage backbone, and conduct fine-tuning.

For each dialogue, there is a set of coreference clusters $\{C_1, C_2, ..., C_u\}$, and each cluster C_i contains entities $\{E_1^i, E_2^i ..., E_m^i\}$. As the multi-turn dialogue sample shown in Figure 3, there are three coreference clusters (colored in yellow, red, and blue, respectively), and each cluster consists a number of words/spans in the same coreference chain. During the conversational interaction, the referring of pronouns is important for semantic context understanding (Sacks et al., 1978), thus we postulate that incorporating coreference information explicitly can be useful for abstractive dialogue summarization. In this work, we focus on enhancing the encoder with auxiliary coreference features.

4.1 GNN-Based Coreference Fusion

As entities in coreference chains link to each other, a graphical representation could readily characterize the underlying structure and facilitate computational modeling of the inter-connected relations. In previous works, Graph Convolutional Networks (GCN) (Kipf and Welling, 2017) show strong capability of modeling graphical features in various tasks (Yasunaga et al., 2017; Xu et al., 2020), thus we use it for the coreference feature fusion.

4.1.1 Coreference Graph Construction

To build the chain of a coreference cluster, we add links between each entity and their mentions. Unlike previous work (Xu et al., 2020) where entities in one cluster are all pointed to the first occurrence, here we connect the adjacent pairs to retain more local information. More specifically, given a cluster C_i of entities $\{E_1^i, E_2^i ..., E_m^i\}$, we add a link of each E to its precedent.

Then each coreference chain is transformed to a graph, and fed to a graph neural network (GNN). Given a text input of n tokens (here we use a subword tokenization), a coreference graph G is initialized with n nodes and an empty adjacent matrix $G[:][:] = 0$. Iterating each coreference cluster C, the first token t_i of each mention (a word or a text span) is connected with the first token t_j of its antecedent in the same cluster with a bi-directional edge, *i.e.*, $G[i][j] = 1$ and $G[j][i] = 1$.

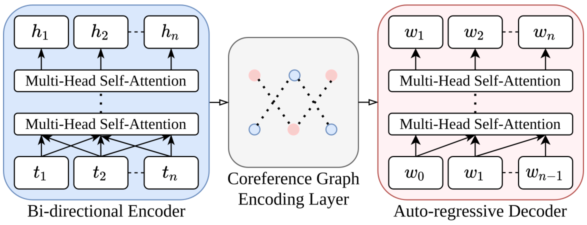


Figure 4: Architecture overview of the GNN-based coreference fusion: the encoder is employed to encode the input sequence; the coreference graph encoding layer is used to model the coreference connections between all mentions; the auto-regressive decoder generates the summaries.

4.1.2 GNN Encoder

Given a graph G with the nodes (words/spans with coreference information in the conversation) and the edges (links between mentions), we employ stacked graph modeling layers to update the hidden representations H of all nodes. Here, we take a single coreference graph encoding (CGE) layer as an example: the input of the first CGE layer is the output H from the Transformer encoder. We denote the input of k-th CGE layer as $H^k = \{h_1^k, ..., h_n^k\}$, and the representations of (k+1)-th layer H^{k+1} are updated as follows:

$$u_i^k = W_1^k \text{ReLU}(W_0^k h_i^k + b_0^k) + b_1^k \quad (1)$$

$$v_i^k = \text{LayerNorm}(h_i^k + \text{Dropout}(u_i^k)) \quad (2)$$

$$w_i^k = \text{ReLU}(\sum_{j \in N_i} \frac{1}{|N_i|} W_2^k v_j^k + b_2^k) \quad (3)$$

$$h_i^{k+1} = \text{LayerNorm}(\text{Dropout}(w_i^k) + v_i^k) \quad (4)$$

where W_i and b_i denote the trainable parameter matrix and bias, $LayerNorm(*)$ is the layer normalization component, and N_i denotes the neighborhood nodes of the i-th node. After feature propagation in all stacked CGE layers, we obtain the final representations by adding the coreference-aware hidden states $H^G = \{h_1^G, ..., h_n^G\}$ with the contextualized hidden states H (here a weight λ is used, and initialized as 0.7), then the auto-regressive decoder is applied to generate summaries.

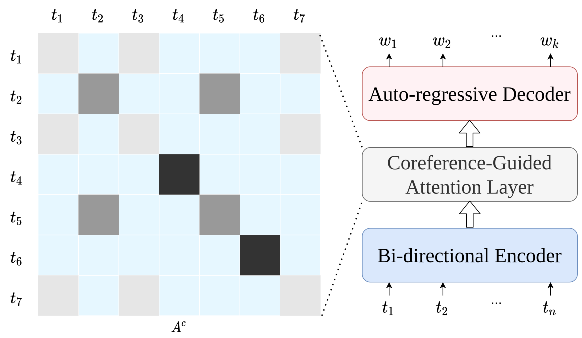


Figure 5: Architecture overview of coreference-guided attention model and an example of coreference attention weight matrix A^c, where $\{t_1,t_3,t_7\}$ are in one coreference cluster and $\{t_2,t_5\}$ are in another cluster, while t_4 and t_6 are tokens without any coreference link.

4.2 Coreference-Guided Attention

Aside from the GNN-based method which introduces a certain number of additional parameters, we further explore a parameter-free method. With the self-attention mechanism (Vaswani et al., 2017), contextualized representation can be obtained with attentive weighted sum. For entities in a coreference cluster, they all share the referring information at the semantic level. Therefore, we propose to fuse the coreference information via one additional attention layer in the contextualized representation.

Given a sample with coreference clusters, a coreference-guided attention layer is constructed to update the encoded representations H. The overview of adding the coreference-guided attention layer is shown in Figure 5. Since items in the same coreference cluster are attended to each other, values in the attention weight matrix A^c are normalized with the number of all referring mentions in one cluster, then the representation h_i of token i is updated according to the following:

$$a_i = \sum_{j \in C^*} \frac{1}{|C^*|} h_j, \quad if\ t_i \in C^* \tag{5}$$

$$h_i^A = \lambda h_i + (1 - \lambda) a_i \tag{6}$$

where a_i is the attentive representation of t_i, if t_i belongs to one coreference cluster C^*, the representation of t_i is updated, otherwise, it remains unchanged. λ is an adjustable parameter and initialized as 0.7. In our experimental settings, we observed that when λ is trainable, it is trained to be 0.69 when our coreference-guided attention model achieved the best performance on the validation set. Following the coreference-guided attention layer, we obtain the final representations with coreference information $H^A = \{h_1^A, ..., h_n^A\}$, then they are fed to the decoder for output generation.

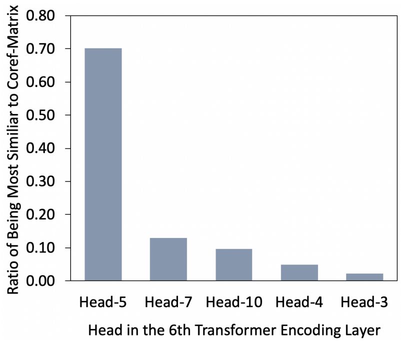


Figure 6: Similarity distribution of head probing with pre-defined coreference matrix. The X-axis shows the heads in the 6-th layer of the Transformer encoder. Values on the Y-axis denote the ratio that a head has the highest similarity with the coreference attention matrix.

4.3 Coreference-Informed Transformer

While pre-trained models bring significant improvement, they still present insufficient prior knowledge for tasks requiring high-level semantic understanding such as coreference resolution. In this section, we explore another parameter-free method by directly enhancing the language backbone. Since the encoder of our neural architecture uses the self-attention mechanism, we proposed feature injection by attention weight manipulation. In our case, the encoder of BART (Lewis et al., 2020) comprises 6 multi-head self-attention layers, and each layer has 12 heads. To incorporate coreference information, we selected heads and modified them with weights that present coreference mentions (see Figure 7).

4.3.1 Attention Head Probing and Selection

To retain prior knowledge provided by the language backbone as much as possible, we first conduct a probing task to strategically select attention heads. Since different layers and heads convey linguistic features of different granularity (Hewitt and Manning, 2019), our target is to find the head that represents the most coreference information. We probe the attention heads by measuring the cosine similarity between their attention weight matrix A^o and a pre-defined coreference attention matrix A^c as described in Section 4.2:

$$head_{probe} = \arg\max_i(\cos(A_i^o, A^c)) \tag{7}$$

where A_i^o is the attention weight matrix of the original i-th head, and $i \in (1, ..., N_h)$, N_h is the number

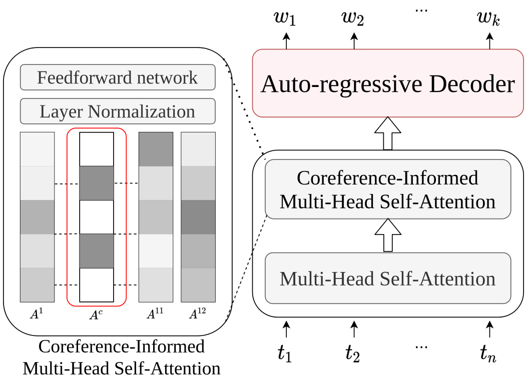


Figure 7: Architecture overview of the coreference-informed Transformer with attention head manipulation. The second attention head is selected and replaced by a coreference attention weight matrix A^c.

of heads in each layer. With all samples in the validation set, we conducted probing on all heads in the 5-th layer and 6-th layer of the *'BART-Base'* encoder. We observed that: (1) in the 5-th layer, the 7-th head obtained the highest similarity score on 95.2% evaluation samples; (2) in the 6-th layer, the 5-th head obtained the highest similarity score on 68.9% evaluation samples. The statistics of heads in 6-th encoding layer are shown in Figure 6.

4.3.2 Coreference-Informed Multi-Head Self-Attention

In order to explicitly utilize the coreference information, we replaced the two predominant attention heads with coreference-informed attention weights. The multi-head self-attention layers (Vaswani et al., 2017) are formulated as:

$$\text{Attention}(Q, K, V) = \text{Softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{8}$$

$$head_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{9}$$

$$\text{MHA}(Q, K, V) = \text{Concat}(head_1, ..., head_{N_h}) \tag{10}$$

$$\text{FFN}(x_i^l) = \text{ReLU}(x_i^l W_1^F + b_1^F)W_2^F + b_2^F \tag{11}$$

where Q, K and V are the sets of queries, keys and values respectively. W_i and b_i are the trainable parameter matrix and bias. d_k is the dimension of keys, x_i^l is the representation of i-th token after the l-th multi-head self-attention layer. FFN is the point-wise feed forward layer. Based on the probing analysis in Section 4.3.1, we selected the 7-th head of 5-th encoding layer, and the 5-th head of 6-th encoding layer for coreference injection, and observed that models with probing selection outperformed that of random head selection.

	# Conv	*# Sp*	*# Turns*	*# Ref Len*
Train	14732	2.40	11.17	23.44
Validation	818	2.39	10.83	23.42
Test	819	2.36	11.25	23.12

Table 1: Data details of the SAMSum corpus. *# Conv*, *# Sp*, *# Turns* and *# Ref Len* refer to the average number of conversations, speakers, dialogue turns and the average number of words in the gold reference summaries.

5 Experiments

5.1 Dataset

We evaluated the proposed methods on SAMSum (Gliwa et al., 2019), a dialogue summarization dataset consisting of 16,369 conversations with human-written summaries. Dataset statistics are listed in Table 1.

5.2 Model Settings

The vanilla sequence-to-sequence Transformer (Vaswani et al., 2017) was applied as the base architecture. We used the pre-trained *'BART-Base'* (Lewis et al., 2020) as language backbone. Then, we enhanced the base model with following three methods: **Coref-GNN**: Incorporating coreference information by the GNN-based fusion (see Section 4.1); **Coref-Attention**: Encoding coreference information by an additional attention layer (see Section 4.2); **Coref-Transformer**: Modeling coreference information by the attentive head probing and replacement (see Section 4.3). Several baselines were selected for comparison: (1) *Pointer-Generator Network* (See et al., 2017); (2) *DynamicConv-News* (Wu et al., 2019); (3) *Fast-Abs-RL-Enhanced* (Chen and Bansal, 2018); (4) *Multi-View BART* (Chen and Yang, 2020), which provides the state-of-the-art result.

5.3 Training Configuration

The proposed models were implemented in PyTorch (Paszke et al., 2019), and Hugging Face Transformers (Wolf et al., 2020). The Deep Graph Library (DGL) (Wang et al., 2019) was used for implementing the *Coref-GNN*. The trainable parameters were optimized by Adam (Kingma and Ba, 2014). The learning rate of the GCN component was 1e-3, and that of BART was set at 2e-5. We trained each model for 20 epochs and selected the best checkpoints on the validation set with ROUGE-2 score. All experiments were run on a single Tesla V100 GPU with 16GB memory.

Model	ROUGE-1			ROUGE-2			ROUGE-L		
	F	P	R	F	P	R	F	P	R
*Pointer-Generator**	40.1	-	-	15.3	-	-	36.6	-	-
*Fast-Abs-RL-Enhanced**	42.0	-	-	18.1	-	-	39.2	-	-
*DynamicConv-News**	45.4	-	-	20.6	-	-	41.5	-	-
*BART-Large**	48.2	49.3	51.7	24.5	25.1	26.4	46.6	47.5	49.5
*Multi-View BART-Large**	49.3	51.1	52.2	**25.6**	26.5	**27.4**	**47.7**	49.3	**49.9**
BART-Base	48.7	50.8	51.5	23.9	25.8	24.9	45.3	48.4	47.3
Coref-GNN	50.3	**56.1**	50.3	24.5	27.3	24.6	46.0	**50.9**	46.8
Coref-Attention	**50.9**	54.6	**52.8**	25.5	27.4	26.8	46.6	50.0	48.4
Coref-Transformer	50.3	55.5	50.9	25.1	**27.7**	25.6	46.2	**50.9**	46.9

Table 2: ROUGE scores of baselines and proposed models. * denotes the results from Chen and Yang (2020). F, P, and R denote F1 Score, Precision and Recall, respectively.

Model	R-1	R-2	R-L
Models trained with BART-Large			
MV-BART-Large	53.42	27.98	49.97
LM-Annotator ($\mathcal{D}_{\text{All}}$)	53.70	28.79	50.81
Our Model (Large)	53.91	28.58	50.39

Table 3: ROUGE F1 scores of baselines and our proposed framework. The reported results use the same ROUGE calculation following (Feng et al., 2021) for the benchmarked comparison.

Model	Average # Words
Reference	23.12 ± 12.20
BART-Base	22.72 ± 10.78
Coref-GNN	19.62 ± 8.75
Coref-Attention	21.68 ± 10.27
Coref-Transformer	20.54 ± 9.39

Table 4: Average word number with standard deviations of generated summaries.

Model	Average Scores
BART-Base	0.60
Coref-GNN	0.84
Coref-Attention	**1.16**
Coref-Transformer	0.96

Table 5: Human evaluation results: each summary is scored on the scale of [-2, 0, 2] as (Chen and Yang, 2020). Reported scores are averaged on 100 samples.

6 Results

6.1 Automatic Evaluation

We quantitatively evaluated the proposed methods with the standard metric ROUGE (Lin and Och, 2004), and reported ROUGE-1, ROUGE-2 and ROUGE-L.[3] As shown in Table 2, our base model *BART-Base* outperformed *Fast-Abs-RL-Enhanced* and *DynamicConv-News* significantly, showing the effectiveness of fine-tuning pre-trained language backbones for abstractive dialogue summarization. Adopting *BART-Large* could bring about relative 5% improvement, while it doubled the parameter size and training time of *BART-Base*. As shown in Table 2, compared with the base model *BART-Base*, the performance is improved significantly by our proposed methods. In particular, *Coref-Attention* performed best with 4.95%, 6.69% and 2.87% relative F-measure score improvement, and *Coref-GNN* achieved the highest scores on precision with 10.43% on ROUGE-1, 5.81% on ROUGE-2 and 5.17% on ROUGE-L. *Coref-Transformer* also showed consistent improvement.

Moreover, compared with the *BART-Base* model (Lewis et al., 2020), the proposed coref-models performed better on ROUGE-1 scores, especially on the precision metrics. More specifically, precision scores are improved 9.78%, 6.85%, and 8.61% relatively by *Coref-GNN*, *Coref-Attention* and *Coref-Transformer*, respectively. For ROUGE-2 and ROUGE-L, our models also obtain comparable performance. Recently, Feng et al. (2021) conducted a benchmarked comparison of state-of-the-art dialogue summarizers. As shown in Table 3, our method (trained with *BART-Large*) is comparable to *MV-BART-Large* (Chen and Yang, 2020) and *LM-Annotator* ($\mathcal{D}_{All}$) (Feng et al., 2021).

As shown in Table 2, we also observed that the most significant improvement is on the precision scores while the recall scores remains comparable with strong baselines. Moreover, as shown in Table

[3] We used integrated functions in HuggingFace Transformers (Wolf et al., 2020) to calculate ROUGE scores. Note that different libraries may result in different ROUGE scores.

Model	Missing Information	Redundant Information	Wrong Reference	Incorrect Reasoning
Base Model	34	26	22	20
Coref-GNN	32 [5.8% ↓]	8 [69% ↓]	14 [36% ↓]	16 [20% ↓]
Coref-Attention	**28** [17% ↓]	**4** [84% ↓]	**12** [45% ↓]	**9** [55% ↓]
Coref-Transformer	32 [5.8% ↓]	12 [53% ↓]	14 [36% ↓]	12 [40% ↓]

Table 6: Percentage of typical errors in summaries generated by the baseline and our proposed models. Values in brackets denote the relative decrease compared with the base model.

Conversation (abbreviated)	*BART-Base*	*Coref-Attention*
(i) ... Ivan : so youre coming to the wedding Eric: your brother's Ivan: yea Eric: i dont know mannn Ivan: YOU DONT KNOW?? Eric: i just have a lot to do at home, plus i dont know if my parents would let me Ivan: ill take care of your parents Eric: youre telling me you have the guts to talk to them XD Ivan: thats my problem Eric: okay man, if you say so Ivan: yea just be there Eric: alright	Eric is not sure if he's going to the wedding, because he has a lot to do at home and doesn't know if his parents would let him. Ivan will come to Eric's wedding.	Eric is coming to Ivan's brother's wedding. Eric has a lot to do at home and he can't take care of his parents. Ivan will be there.
(ii) Derek McCarthy: Filip - are you around? Would you have an Android cable I could borrow for an hour? ... Tommy : I am in Poland but can ring my wife and she will give you one ... Tommy: 67 glenoaks close Derek McCarthy: That would be great if you could!! ... Tommy: Sent her msg. She will give it to you. Approx time when she will be at home is 8:15 pm Derek McCarthy: Thanks again!! ...	Tommy will call his wife to borrow a phone charger from Derek McCarthy. Tommy will be at home at 8:15 pm.	Filip will lend Derek McCarthy his Android cable. He will call his wife at 67 glenoaks close.
(iii) Ann: Congratulations!! Ann: You did great, both of you! Sue: Thanks, Ann Julie: I'm glad it's over! Julie: That's co cute of you, my girl! Ann: Let's have a little celebration tonight! Sue: I'm in Julie: me too!!! aww	Ann congratulates Sue and Julie on their success. Ann and Julie will celebrate tonight.	Ann and Julie are congratulating Sue on their success.

Table 7: Three examples of generated summaries: For conversation *i* and conversation *ii*, *Coref-Attention* model generated correct summaries by incorporating coreference information. *Coref-Attention* model generated an imperfect summary for conversation *iii* due to inaccurate coreference resolution provided.

4, the average length of generated summaries of the base model is 22.72, and that of the *coref*-models is slightly shorter. We speculated that the proposed models tend to generate more concise summaries while preserving the important information, which is also supported by the analysis in Section 7.1.

6.2 Human Evaluation

As the example shown in Figure 1, ROUGE scores are insensitive to semantic errors such as incorrect reference, thus we conducted human evaluation to complement objective metrics. Following Gliwa et al. (2019) and Chen and Yang (2020), each summary is scored on the scale of [-2, 0, 2], where -2 means the summary is unacceptable with the wrong reference, extracted irrelevant information or does not make logical sense, 0 means the summary is acceptable but lacks of important information converge, and 2 refers to a good summary which is concise and informative. We randomly selected 100 test samples, and scored the summaries generated by the base model, *Coref-GNN*, *Coref-Attention* and *Coref-Transformer*. Four linguistic experts conducted the human evaluation, and their average scores are reported in Table 5. Compared with the base model, our *coref*-models obtain higher scores in human ratings, which is consistent with the quantitative ROUGE results.

7 Analysis

7.1 Quantitative Analysis

To further evaluate the generation quality and effectiveness of coreference fusion for dialogue summarization, we annotated four types of common errors in the automatic summaries:

Missing Information: The content is incomplete in the generated summary compared with the human-written reference.

Redundant Information: There is redundant content in the generated summary compared with the human-written reference.

Wrong References: The actions are associated with the wrong interlocutors or mentions (*e.g.,* In the example of Figure 1, the summary generated by base model confused *"Payton"* and *"Max"* in the actions of *"look for good places to buy clothes"* and *"love reading books"*).

Incorrect Reasoning: The model incorrectly reasons the conclusion from context of multiple dialogue turns. Moreover, wrong reference and incorrect reasoning will lead to factual inconsistency from source content.

We randomly sampled 100 conversations in the test set and manually annotated the summaries generated by the base and our proposed models with the four error types. As shown Table 6, 34% of summaries generated by the base model cannot summarize all the information included in the gold references, and models with coreference fusion improve the information coverage marginally. Coreference-aware models essentially reduced the redundant information: 84% relative reduction by *Coref-Attention*, 69% relative reduction by *Coref-GNN*, and 53% relative reduction by *Coref-Transformer*. *Coref-Attention* model also performed best on reducing 45% of wrong reference errors relatively, *Coref-GNN* and *Coref-Transformer* both relatively reduced 36% of that. Encoding coreference information by an additional attention layer substantially improves the reasoning capability by reducing 55% relatively in incorrect reasoning, *Coref-Transformer* and *Coref-GNN* also relatively reduced this error by 40% and 20% compared with the base model. This shows our models can generate more concise summaries with less redundant content, and incorporating coreference information is helpful to reduce wrong references, and conduct better multi-turn reasoning.

7.2 Sample Analysis

Here we conducted a sample analysis as in (Lewis et al., 2020). Table 7 shows 3 examples along with their corresponding summaries from the *BART-Base* and *Coref-Attention* model. Conversation *i* and *ii* contain multiple interlocutors and referrals. The base model made some referring mistakes: (1) in conversation *i*, *"your brother's wedding"* should refer to *"Ivan's brother's wedding"*; (2) in conversation *ii*, since *"Fillip"* and *"Tommy"* are exactly the same person, pronouns *"you"* and *"I"* in *"Would you have an Android cable I could borrow..."* should refer to *"Tommy"* and *"Derek McCarthy"*, respectively. In contrast, the *Coref-Attention* model was able to make correct statements. However, if the coreference resolution quality is poor, the coreference-aware models will be affected. For example, in the conversation *iii*, when the pronouns *"you"* and *"my girl"* in *"Julie: That's co cute of you, my girl"* are wrongly included in the coreference cluster of *"Julie"*, the model will also make referring mistakes in the summary .

8 Conclusion

In this paper, we investigated the effectiveness of utilizing coreference information for summarizing multi-party conversations. We proposed three approaches to explicitly incorporate coreference information into neural abstractive dialogue summarization: (1) GNN-based coreference fusion; (2) coreference-guided attention; and (3) coreference-informed Transformer. These methods can be adopted on various neural architectures. Quantitative results and human analysis suggest that coreference information helps track referring chains in conversations. Our proposed models compare favorably with baselines without coreference guidance and generate summaries with higher factual consistency. Our work provides empirical evidence that coreference is useful in dialogue summarization and opens up new possibilities of exploiting coreference for other dialogue related tasks.

Acknowledgments

This research was supported by funding from the Institute for Infocomm Research (I2R) under A*STAR ARES, Singapore. We thank Ai Ti Aw and Minh Nguyen for insightful discussions. We also thank the anonymous reviewers for their precious feedback to help improve and extend this piece of work.

References

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Henry Y Chen, Ethan Zhou, and Jinho D Choi. 2017. Robust coreference resolution and entity linking on dialogues: Character identification on tv show transcripts. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 216–225.

Jiaao Chen and Diyi Yang. 2020. Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4106–4118, Online. Association for Computational Linguistics.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–686, Melbourne, Australia. Association for Computational Linguistics.

Pradeep Dasigi, Nelson F. Liu, Ana Marasović, Noah A. Smith, and Matt Gardner. 2019. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. 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 5925–5932, Hong Kong, China. Association for Computational Linguistics.

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

Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021. A survey on dialogue summarization: Recent advances and new frontiers. *arXiv preprint arXiv:2107.03175*.

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

Chih-Wen Goo and Yun-Nung Chen. 2018. Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 735–742. IEEE.

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.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. 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 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Daniel Jurafsky and James H Martin. 2008. Speech and language processing: An introduction to speech recognition, computational linguistics and natural language processing. *Upper Saddle River, NJ: Prentice Hall*.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *Proceedings of International Conference on Learning Representations (ICLR)*.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 687–692.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Manling Li, Lingyu Zhang, Heng Ji, and Richard J. Radke. 2019. Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2190–2196, Florence, Italy. Association for Computational Linguistics.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612, Barcelona, Spain.

Peter J. Liu*, Mohammad Saleh*, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. In *International Conference on Learning Representations*.

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.

Zhengyuan Liu and Nancy Chen. 2019. Reading turn by turn: Hierarchical attention architecture for spoken dialogue comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5460–5466.

Zhengyuan Liu, Angela Ng, Sheldon Lee, Ai Ti Aw, and Nancy F Chen. 2019. Topic-aware pointer-generator networks for summarizing spoken conversations. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 814–821. IEEE.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in NeurIPS 2019*, pages 8024–8035.

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

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.

HARVEY Sacks, EMANUEL A. SCHEGLOFF, and GAIL JEFFERSON. 1978. A simplest systematics for the organization of turn taking for conversation. In JIM SCHENKEIN, editor, *Studies in the Organization of Conversational Interaction*, pages 7 – 55. Academic Press.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1073–1083.

Guokan Shang, Wensi Ding, Zekun Zhang, Antoine Tixier, Polykarpos Meladianos, Michalis Vazirgiannis, and Jean-Pierre Lorré. 2018. Unsupervised abstractive meeting summarization with multi-sentence compression and budgeted submodular maximization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 664–674, Melbourne, Australia. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Minjie Wang, Da Zheng, Zihao Ye, Quan Gan, Mufei Li, Xiang Song, Jinjing Zhou, Chao Ma, Lingfan Yu, Yu Gai, Tianjun Xiao, Tong He, George Karypis, Jinyang Li, and Zheng Zhang. 2019. Deep graph library: A graph-centric, highly-performant package for graph neural networks. *arXiv preprint arXiv:1909.01315*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Felix Wu, Angela Fan, Alexei Baevski, Yann Dauphin, and Michael Auli. 2019. Pay less attention with lightweight and dynamic convolutions. In *International Conference on Learning Representations*.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Discourse-aware neural extractive text summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5021–5031, Online. Association for Computational Linguistics.

Michihiro Yasunaga, Rui Zhang, Kshitijh Meelu, Ayush Pareek, Krishnan Srinivasan, and Dragomir Radev. 2017. Graph-based neural multi-document summarization. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 452–462, Vancouver, Canada. Association for Computational Linguistics.

Zhou Zhao, Haojie Pan, Changjie Fan, Yan Liu, Linlin Li, Min Yang, and Deng Cai. 2019. Abstractive meeting summarization via hierarchical adaptive segmental network learning. In *The World Wide Web Conference*, pages 3455–3461.

Chenguang Zhu, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. 2021. Enhancing factual consistency of abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 718–733, Online. Association for Computational Linguistics.