Knowledge-Based Systems 245 (2022) 108550

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Leveraging speaker-aware structure and factual knowledge for faithful dialogue summarization

Lulu Zhao [a], Weiran Xu [a,*], Chunyun Zhang [b], Jun Guo [a]

[a] Beijing University of Posts and Telecommunications, Beijing, China
[b] Shandong University of Finance and Economics, Jinan, China

ARTICLE INFO

Article history:
Received 11 October 2021
Received in revised form 4 March 2022
Accepted 5 March 2022
Available online 18 March 2022

Keywords:
Abstractive dialogue summarization
Speaker-aware structure
Dual-copy
Factual consistency

ABSTRACT

Currently, sequence/graph-to-sequence models for abstractive dialogue summarization are being studied extensively. However, previous methods strive to integrate complex events spanning multiple utterances, and the generated summaries are often filled with incorrect facts. In this study, we first utilize the speaker-aware structure to model the information interaction process in the dialogue, which shows an excellent ability to settle the cross-sentence dependency. Then, we incorporate the factual representations via a dual-copy decoder to obtain summaries conditioned on both the tokens from source sequences and the factual knowledge from our designed fact graph, which enhances the factual consistency for dialogue summarization. We also propose some fact-level factual consistency metrics. Adequate experimental results demonstrate that our model outperforms the state-of-the-art baselines by a significant margin on the SAMSum and DialSumm datasets. A comprehensive analysis also proves the effectiveness of our model. Furthermore, human judges confirm that the outputs of our model contain more informative and faithful information.

1. Introduction

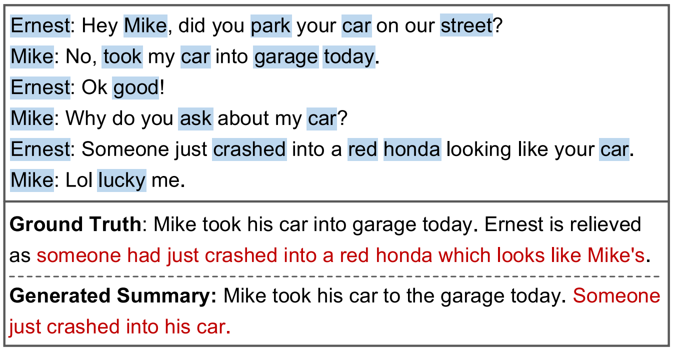
Abstractive summarization is one of the most challenging and important tasks in natural language processing (NLP), which aims to condense a piece of text to a shorter version, containing the main information from the original data. The majority of previous studies have focused on summarizing single-speaker structured text such as news reports and scientific publications [1–4]. Recently, with the explosive growth of dialogic texts, more attention has been paid to abstractive dialogue summarization, which is useful for speakers to recap the salient information in the conversation and for absentees to grasp the key points [5]. However, because dialogue is a multi-speaker view exchange process and its spoken language style is casual [6], the descriptions of one event are usually scattered among multiple utterances, which makes it difficult to integrate the salient elements, and the generated summaries often suffer from factual inconsistency. In Fig. 1, we provide an example of an unfaithful generation.

Many previous studies have proposed pushing the frontiers of abstractive dialogue summarization. Some of them directly apply the existing document summarization models [7], some use expert annotation information, such as dialogue acts [8], topic segmentations [9,10], key point sequences [11], and conversation stages [12], whereas others model the relations of diverse textual units in dialogues via topic-word interaction graphs [13], discourse relations, and action graphs [14]. However, the aforementioned methods do not utilize the speaker-aware structure, a type of special structural information in the conversation, to model the dialogue context. In addition, most of them only strive to improve the ROUGE scores [15] but ignore the exploration of factual consistency. We believe that simulating the information interaction process among multiple speakers can facilitate modeling dependencies between utterances [16,17], to integrate the salient fragments of the dialogue. Encoding factual knowledge into the generation process will further improve the informativeness and faithfulness of the summary.

To mitigate these issues, we propose a multi-interaction-guided dual-copy knowledge (MIDCK) network, in which a heterogeneous dialogue graph (HDG) is designed to simulate the information interaction process and promote the integration of cross-sentence textual fragments. Concretely, it is necessary to aggregate the self-speaker information via speaker dependency because of the logical inertia that people tend to prioritize integrating their views and then attempt to understand the views of others [17]. Considering that people may continue to insist on or change their views by obtaining information from others based on holding their views, the inter-speaker information is also modeled via sequential context dependency and co-occurring keyword dependency [18], where the core event is navigated to its other occurrences raised by different speakers. The graph

[*] Corresponding author.
E-mail addresses: zhaoll@bupt.edu.cn (L. Zhao), xuweiran@bupt.edu.cn (W. Xu), zhangchunyun1009@126.com (C. Zhang), guojun@bupt.edu.cn (J. Guo).

https://doi.org/10.1016/j.knosys.2022.108550
0950-7051/© 2022 Elsevier B.V. All rights reserved.



Ernest: Hey Mike, did you park your car on our street?
Mike: No, took my car into garage today.
Ernest: Ok good!
Mike: Why do you ask about my car?
Ernest: Someone just crashed into a red honda looking like your car.
Mike: Lol lucky me.

Ground Truth: Mike took his car into garage today. Ernest is relieved as someone had just crashed into a red honda which looks like Mike's.

Generated Summary: Mike took his car to the garage today. Someone just crashed into his car.

Fig. 1. Example from the SAMSum dataset. Factually inconsistent fragments are marked in red, and the keywords are highlighted in blue.

encoder utilizes graph neural networks to perform multi-hop reasoning over an HDG and aggregate the utterance-level features. A sequence encoder is also used to extract sequential features at the token level. These two encoders cooperate to express the conversational contents via two different granularities, showing an excellent ability to settle the cross-sentence dependency.

In addition, a fact graph (FG) is constructed to encode existing factual knowledge into the summarization system. We applied the Stanford CoreNLP [19], a dependency parser tool, to extract factual knowledge in the form of relational tuples (subject, dependency, and object), which describe facts and are considered as the skeletons of dialogues. Note that the node in this graph denotes the subject or object, and the edge is the dependency relation between the subject and object. In addition to the FG, we used graph neural networks to integrate the expressions of concise concepts or events. We then designed a dual-copy mechanism to copy content from the tokens of the dialogue text and the factual knowledge of the knowledge graph in parallel, which would clearly provide the right guidance for summarization.

We conducted extensive experiments to evaluate the performance of the proposed models. Experimental results on two real-world datasets from different domains demonstrate that our methods outperform strong baselines in abstractive dialogue summarization, particularly in factual consistency. Human judges further confirmed that the MIDCK generates more informative summaries with fewer unfaithful errors than other models without an HDG and FG. Finally, we discuss the existing challenges in abstractive dialogue summarization.

The main contributions of this study are summarized as follows:

- To the best of our knowledge, we are the first to model an HDG to simulate human communication via a speaker-aware structure and effectively capture cross-sentence dependency.
- We constructed an FG with factual knowledge and designed a dual-copy decoder to conduct neural attention computations both on the source text and FG, which alleviates the factual inconsistency in the summary generation to some extent.
- We performed extensive experiments and qualitative analysis on two large-scale datasets, the SAMSum and DialSumm, to prove the effectiveness of our methods. Some novel fact-level factual consistency metrics have been proposed to evaluate our proposed methods from different perspectives. We also discuss the current challenges of existing dialogue summarizers for abstractive dialogue summarization.

2. Related work

Our methods draw inspiration from the research fields of abstractive dialogue summarization, factual consistency in summarization, and dialogue systems. We introduce these related studies in this section.

2.1. Abstractive dialogue summarization

Owing to the lack of publicly available resources, studies on dialogue summarization have rarely been conducted in the early stages. Some previous studies benchmarked the abstractive dialogue summarization task using the AMI meeting corpus, which contains a wide range of annotations, including dialogue acts and topic descriptions [20–22]. Goo et al. [8] proposed the use of high-level topic descriptions (e.g., cost evaluation of project process) as the golden summaries and leveraged dialogue act signals in a neural summarization model. They assumed that dialogue acts indicated interactive signals and used this information to improve performance. Customer service interaction is also a common form of dialogue that includes questions from the user and solutions from the agent. Liu et al. [11] collected a dialogue-summary dataset from the logs of the DiDi customer service center. They proposed a novel leader–writer network that relies on auxiliary key point sequences to ensure the logic and integrity of dialogue summaries and designed a hierarchical decoder. The rules for labeling the key point sequences are provided by domain experts, which requires considerable human effort. Zou et al. [23] constructed a real-world Chinese customer-service dataset from the call center of an e-commerce company. They proposed a topic-augmented two-stage summarizer with a multi-role topic modeling mechanism for customer-service dialogues, which can generate highly abstractive summaries that highlight role-specific information. In addition, Fabbri et al. [24] benchmarked state-of-the-art models on diverse online conversation forms including news comments, discussion forums, community question answering forums, and email threads. However, these datasets are either small in scale or not publicly available and there are only a few scattered studies, preventing subsequent researchers from following and reproducing the previous studies, and the research on these datasets is not coherent and inheritable. In addition, the form of meeting scripts is different from the dialogues in real-life scenarios, which affects the exploration of the unique characteristics and structures of the dialogues. Therefore, the development of technologies for summarizing dialogues has been limited to a certain extent in the past.

To fill this gap, some large-scale and public datasets on open-domain chitchats have recently been released, including the SAMSum dataset [7] and DialSumm dataset [25]. For the SAMSum dataset, a number of methods have been proposed that involve

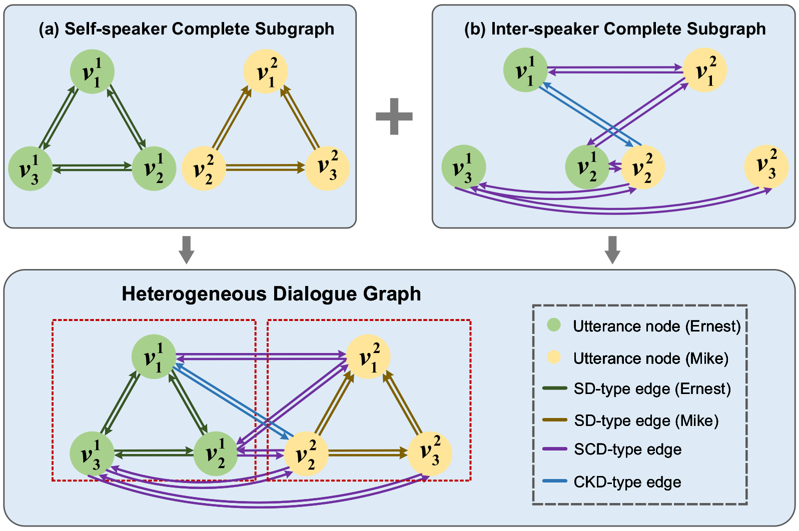


Fig. 2. Detailed illustration of heterogeneous dialogue graph. (a) Construction of a self-speaker complete subgraph. (b) Construction of an inter-speaker sparse subgraph.

the utilization of topic word information [13,26], conversational structures [12,14], personal named entity planning [27], and semantic slots [28]. In addition, because the DialSumm dataset has recently been proposed, there is no related research available. However, current models pay less attention to the interaction of speakers to capture long-distance cross-sentence dependency, which leads to incorrect combinations among salient elements and many unfaithful summarization errors.

2.2. Factual consistency in summarization

In terms of factual errors, some studies focused on designing evaluation metrics for factual consistency, as many human evaluations have shown that ROUGE scores correlate poorly with faithfulness [29]. They range from using fact triples [30,31], textual entailment predictions [32], adversarially pre-trained classifiers [33], and question answering (QA) systems [34,35]. Another line of related studies focused on enforcing factuality in summarization models. Cao et al. [36] and Zhu et al. [37] proposed RNN-based and transformer-based decoders that address the source texts and extracted knowledge triples, respectively. Li et al. [38] proposed an entailment-reward-augmented maximum-likelihood training objective. Dong et al. [39] and Cao et al. [40] designed post-editing correctors to boost the factual consistency in generated summaries. Nevertheless, these methods do not leverage the unique characteristics of dialogues. We are the first to address the issue of factual inconsistency in multi-speaker dialogue scenarios via the speaker-aware structure and factual knowledge, which can significantly improve the performance in terms of multiple factual correctness metrics without a significant drop in ROUGE scores.

2.3. Dialogue system

Our study aims to analyze abstractive summarization in conversational scenarios; therefore, we introduce some research related to various dialogue systems. The dialogue system is one of the popular topics in NLP and is promising in real-life applications. It is generally divided into two categories. One is the task-oriented dialogue system [41], which solves specific problems in a certain domain, such as movie ticket booking, consisting of four functional modules: natural language understanding, dialogue state tracking, policy learning, and natural language generation. The other category is the open-domain dialogue system, involving many popular topics, such as context awareness [42–44], response coherence [45–47], response diversity [48,49], speaker consistency and personality-based response [50–52], empathetic response [53–55], conversation topic [56,57], knowledge-grounded system [58,59], interactive training [60,61], and visual dialogue [62,63]. In addition, the recurrent units for conversational sentiment analysis have recently gained increasing attention and have been applied to many scenarios [16,17,64–68]. Although many studies have been conducted on these topics, there are still many directions worth researching, such as multimodal dialogue systems [69] and the fusion of task-oriented and open-domain dialogues in conversational agents [70–72]. In general, abstractive dialogue summarization and dialogue systems are both dialogue-related tasks, and the design of our model can also benefit from advances in these related fields.

3. Heterogeneous dialogue graph

In this section, we introduce the construction of an HDG, which utilizes a speaker-aware structure to simulate the information interaction and integrate the cross-sentence salient fragments. It consists of two parts: (1) a self-speaker complete (SSC) subgraph, which aggregates utterances from the same speaker, and (2) an inter-speaker sparse (ISS) subgraph, which simulates the information interaction between different speakers. In this process, one person is selected as the speaker, and all other people are regarded as listeners; that is, each person plays different roles in different scenarios. To facilitate model processing, we propose the concept of a heterogeneous graph. Fig. 2 shows an HDG constructed from the instance shown in Fig. 1.

Given an HDG $G_d=(V_d, E_d)$, V_d stands for the node set, which is defined as $V_d=V_{s_1}\cup\cdots\cup V_{s_n}$, and n represents the number of speakers in the dialogue. Here, $v_i^k\in V_{s_k}\subset V_d$ represents the ith utterance of the kth speaker. $E_d=E_{d,hom}\cup E_{d,het}$ is an edge set, where $E_{d,hom}$ is the homogeneous edge set of nodes belonging to the same speaker, and $E_{d,het}$ is the heterogeneous edge set of nodes belonging to different speakers. In the following subsections, we describe the construction process of the two subgraphs in detail.

3.1. Self-speaker complete subgraph construction

Speakers in a conversation may be willing to stick to their views unless their counterparts invoke a change. According to the

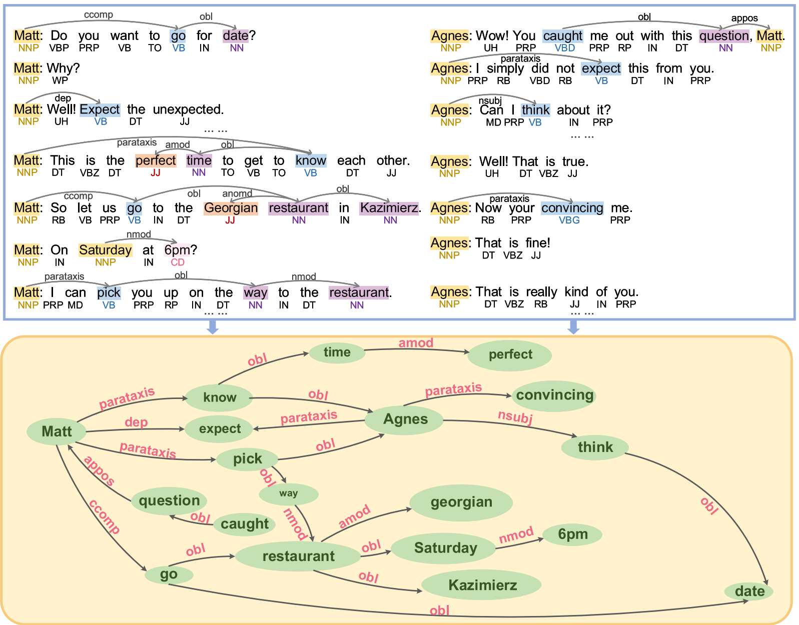

Matt: Do you want to go for date?
NNP VBP PRP VB TO VB IN NN
ccomp
obl
Matt: Why?
NNP WP
Matt: Well! Expect the unexpected.
NNP UH VB DT JJ
dep
Matt: This is the perfect time to get to know each other.
NNP DT VBZ DT JJ NN TO VB TO VB DT JJ
parataxis
amod
obl
Matt: So let us go to the Georgian restaurant in Kazimierz.
NNP RB VB PRP VB IN DT JJ NN IN NN
ccomp
obl
anomd
obl
Matt: On Saturday at 6pm?
NNP IN NNP IN CD
nmod
Matt: I can pick you up on the way to the restaurant.
NNP PRP MD VB PRP RP IN DT NN IN DT NN
parataxis
obl
nmod
Agnes: Wow! You caught me out with this question, Matt.
NNP UH PRP VBD PRP RP IN DT NN NNP
obl
appos
Agnes: I simply did not expect this from you.
NNP PRP RB VBD RB VB DT IN PRP
parataxis
Agnes: Can I think about it?
NNP MD PRP VB IN PRP
nsubj
Agnes: Well! That is true.
NNP UH DT VBZ JJ
Agnes: Now your convincing me.
NNP RB PRP VBG PRP
parataxis
Agnes: That is fine!
NNP DT VBZ JJ
Agnes: That is really kind of you.
NNP DT VBZ RB JJ IN PRP
… …
Matt
know
time
perfect
expect
pick
question
caught
go
way
restaurant
georgian
Saturday
6pm
Kazimierz
Agnes
convincing
think
date
obl
amod
parataxis
dep
parataxis
appos
ccomp
obl
obl
parataxis
obl
parataxis
nsubj
obl
obl
nmod
amod
obl
nmod
obl
obl
obl

Fig. 3. Example of the constructed fact graph from a dialogue instance. The highlighted words represent keywords.

speaker-aware structure, we aggregate all utterances conducted by the same speaker. For each speaker, we construct a fully connected graph, and $e^{hom}_{ij,k}\in E_{d,hom}$ $(i\in\{1,\ldots,|s_k|\}, j\in\{1,\ldots,|s_k|\}$ indicates the **speaker dependency (SD)** edge between the ith and jth utterances belonging to the kth speaker. For example, as shown in Fig. 2(a), the edges with SD-type labels of green nodes (Ernest) are $\{(v^1_1, v^1_2), (v^1_1, v^1_3); (v^1_2, v^1_3)\}$ and the SD-type edges of yellow nodes (Mike) are $\{(v^2_1, v^2_2); (v^2_1, v^2_3); (v^2_2, v^2_3)\}$.

3.2. Inter-speaker sparse subgraph construction

Inter-speaker interaction refers to the influence that the counterparts have on a speaker. This influence is closely related to a certain event and often changes with dynamic information flows. This subgraph is constructed to simulate the information interactions among different speakers. We consider the selected utterance nodes associated with one speaker as a type of node group and those associated with all other speakers as another type of node group, which can handle the scenario with a dynamic number of speakers. In addition to the SD-type edge, we define two types of edges, and $e^{het}_{ij,kp}\in E_{d,het}$ $(i\in\{1,\ldots,|s_k|\}, j\in\{1,\ldots,|s_p|\})$ indicates the edge between the ith utterance of the kth speaker and the jth utterance of the pth speaker.

Sequential context dependency (SCD): This relation describes the sequential utterance nodes of different speakers that occur within a fixed-size sliding window, containing the n_p past source nodes and the n_f future source nodes. For example, as shown in Fig. 2(b), this type of edge is $\{(v^1_1, v^2_1);(v^2_1, v^1_2);(v^1_2, v^2_2);(v^2_2, v^1_3);(v^1_3, v^2_3)\}$.

Co-occurring keyword dependency (CKD): The keywords are obtained by performing POS tagging on tokens according to Manning et al. [19], and they are nouns, numerals, adjectives, adverbs, and notional verbs. This relation means that the utterance nodes of different speakers containing the same keyword are connected. For example, as shown in Fig. 2 (b), this type of edge is $\{(v^1_1, v^2_2)\}$.

4. Fact graph

In this section, we describe an FG. To enable the summarization model to be fact-aware, we extract and represent factual knowledge structurally. In essence, dialogue is a process of repeated discussion around some events. The elements of "person'', "action'', "time'', and "location'', which are indispensable for the event description, ensure the factual consistency for dialogue summarization. Here, we first follow some rules to transform the first-person point-of-view utterances into third-person point-of-view forms: (i) substituting first/second-person pronouns with the names of current speakers or surrounding speakers and (ii) replacing third-person pronouns based on coreference clusters in dialogues detected by Manning et al. [19]. We then provide POS tagging of each token in the utterances. The nouns, numerals, adjectives, adverbs, and notional verbs, which can express practical significance, are selected as the keywords, based on which a dependency parser [19] is leveraged to extract the appropriate tuples (subject, dependency, and object) as fact descriptions.[1] Specifically, the dependency parser transforms the utterance into a dependency tree with multiple labeled tuples, and only the tuples related to keywords are chosen. Finally, we merge the tuples containing the same words, collapse co-referential mentions of the same entity into one word, and order words based on the original sentence to form the complete fact descriptions, as shown in Fig. 3. Given a graph G_f=(V_f, E_f), nodes V_f=$\{v_1,\ldots, v_{|V_f|}\}$ represent the subjects and objects of triples, and directed edges E_f point from subjects to objects associated with dependencies.

5. Methodology

In this section, we introduce the MIDCK network for an abstractive dialogue summarization task. As illustrated in Fig. 4, the model consists of a sequence encoder, two graph encoders, and a dual-copy decoder.

5.1. Sequence encoder

The dialogue D is segmented into m utterances and D=$\{d_1, \ldots, d_m\}$, where d_i=$\{w_{i1}, \ldots, w_{il_i}\}$ denotes the ith utterance with

[1] We also attempted to use OpenIE to extract triples. However, approximately 65% and 62% of the OpenIE outputs are empty in the SAMSum and DialSumm datasets, respectively. Considering that the (subject; predicate) or (predicate; object) tuples are present nearly in each sentence, we merge these tuples to form fact descriptions.

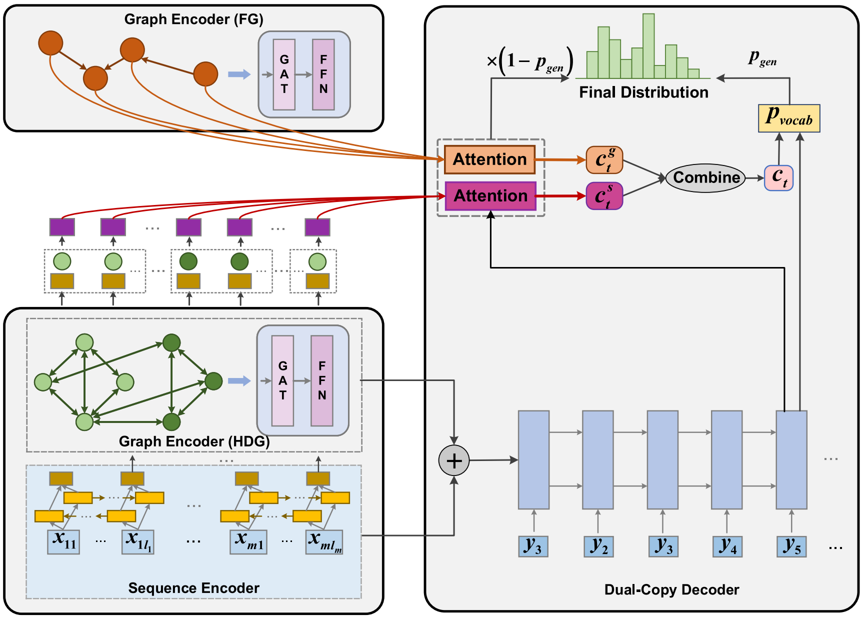


Fig. 4. General architecture of multi-interaction-guided dual-copy knowledge network for abstractive dialogue summarization.

l_i tokens. We encode d_i using a pre-trained encoder, BART [73], and consider the last layer outputs as token embeddings:

$$\{x_{i1}, \ldots, x_{il_i}\} = \text{BART}(\{w_{i1}, \ldots, w_{il_i}\}) \tag{1}$$

We then feed the BART-based embeddings one-by-one into a single-layer bidirectional LSTM, producing a sequence of encoder states $\{h^s_{i1}, \ldots, h^s_{il_i}\}$.

5.2. Graph encoders

In the following subsections, we describe in detail the main components of the two graph encoders, which encode the constructed HDG and FG into hidden vectors.

5.2.1. Initializers

Heterogeneous Dialogue Graph We utilize the output embedding of the last hidden state for the ith utterance, i.e., $h^s_{il_i}$, to initialize the corresponding node in G_d. The edge label matrix E_d is fed into the BART, and we consider the output embedding of the last layer as the relation representation r^d_{ij} between utterances i and j.

Fact Graph We also employ the BART to encode each token within node v_i in G_f and average their output embeddings as the initial representation of the node. The relation representation r^f_{ij} between nodes i and j is initialized via the output embedding of the last layer of the BART.

5.2.2. Graph attention network

To aggregate the features of nodes, we improve upon the graph attention networks [74] by adding residual connections between layers and apply them to the SSC subgraph, ISS subgraph, and FG. The graph attention (GAT) layer is designed as follows:

$$\begin{aligned}
z_{ij} &= \text{LeakyReLU}(\mathbf{W}_a[\mathbf{W}_q v_i; \mathbf{W}_k v_j; r_{ij}]) \\
\alpha_{ij} &= \frac{exp(z_{ij})}{\sum_{l\in\mathcal{N}_i} exp(z_{il})} \\
g_i &= \sigma(\sum_{j\in\mathcal{N}_i} \alpha_{ij}\mathbf{W}_v v_i) + v_i
\end{aligned} \tag{2}$$

where $\mathbf{W}_*$ are weight matrices, σ is the activation function, and $\mathcal{N}$ is the neighborhood of v_i in the graph.

Following each GAT layer, we apply a feed-forward neural (FFN) layer, which contains two linear combinations with a ReLU activation, similar to Transformer [16], to obtain the output representation h^g_i.

It is important to mention that we need to fuse the representations of the two subgraphs, that is, SSC and ISS, to obtain the node representation for HDG, $h^g_{d,i}$. We propose three methods to fuse the information: (1) sequential (from SSC to ISS), (2) sequential (from ISS to SSC), and (3) parallel, which operates on the two subgraphs separately and then sums up the representations (see Section 7.3).

5.3. Dual-copy decoder

Our decoder is a hybrid between a single-layer unidirectional LSTM and a pointer network, which can copy content by pointing and generate tokens from a fixed vocabulary. To obtain summaries with high faithfulness, we devised a dual-copy mechanism to focus on both the input tokens in the source sequence and factual knowledge in the FG.

5.3.1. Copy from fact graph

At each decoding step t, we compute a graph context vector c^g_t using the attention mechanism [75]:

$$\begin{aligned}
c^g_t &= \sum_i a^g_{i,t} h^g_{f,i} \\
a^g_{i,t} &= \text{softmax}({\mathbf{W}^g_a}^T \tanh(\mathbf{W}^g_v h^g_{f,i} + \mathbf{W}^g_s s_t + \mathbf{b}^g_a))
\end{aligned} \tag{3}$$

where $\mathbf{W}^g_*$ and $\mathbf{b}^g_a$ are learnable parameters. s_t is the decoder state at the tth step, and $s_0{=}[h^s_{ml_m}; h^g_d]$ is the decoder initial state, where h^g_d is the average value of all node representations of HDG, that is, $h^g_d{=}\frac{1}{|V_d|}\sum_{i=1}^{|V_d|} h^g_{d,i}$.

5.3.2. Copy from source sequence

We concatenate the token representation of the sequence encoder h^s_{ij} and its corresponding utterance representation $h^g_{d,i}$ to obtain the updated encoder states $\tilde{h}^s_{ij}$. Based on this, the sequence context vector c^s_t is computed similarly.

Table 1
Data statistics of two datasets. A_T and A_S represent the average numbers of turns and speakers, respectively. A_K and A_F represent the average numbers of keywords and fact descriptions extracted from the dialogues, respectively. C_T and C_F denote the proportions of tokens and facts that can be found in the summaries, respectively.

Data split	SAMSum dataset						DialSumm dataset					
	Avg_T	Avg_S	Avg_K	Avg_F	Cop_T	Cop_F	Avg_T	Avg_S	Avg_K	Avg_F	Cop_T	Cop_F
Training	9.7	2.2	32.8	29.8	31%	43%	9.5	2.1	45.1	33.6	26%	37%
Validation	10.2	2.2	32.4	29.0	32%	44%	9.4	2.1	44.8	33.0	26%	38%
Test	10.0	2.2	33.2	30.0	31%	44%	9.7	2.0	44.9	33.2	27%	37%

We use a multi-layer perceptron (MLP) to build a gate network and combine context vectors with the weighted sum:

$$\begin{aligned} g_t &= \mathrm{MLP}(c_t^s, c_t^g) \\ c_t &= g_t \odot c_t^s + (1 - g_t) \odot c_t^g \end{aligned} \tag{4}$$

where $\odot$ denotes the element-wise dot.

Finally, the next token is generated based on the context vector c_t, the decoder state s_t, and the previous word y_{t-1}.

$$\begin{aligned} P_{vocab} &= \mathrm{softmax}(\mathbf{U}'(\mathbf{U}[s_t, c_t] + \mathbf{b}) + \mathbf{b}'), \\ p_{gen} &= \sigma(\mathbf{W}_s^T s_t + \mathbf{W}_c^T c_t + \mathbf{W}_y^T y_{t-1} + \mathbf{b}_g), \\ P_{copy} &= \sum_{i:y_i = y_t} a_{i,t}^s + \sum_{i:y_i = y_t} a_{i,t}^g \\ P(y_t) &= p_{gen} P_{vocab}(y_t) + (1 - p_{gen}) P_{copy}(y_t) \end{aligned} \tag{5}$$

where $\mathbf{U}$, $\mathbf{U}'$, $\mathbf{b}$, $\mathbf{b}'$, $\boldsymbol{W}_*^T$, and $\boldsymbol{b}_g$ are the parameters.

6. Experimental setup

In this section, we introduce the details of the experimental setup, including the dataset, implementation details, comparison methods, and automated evaluation metrics.

6.1. Dataset

We performed our experiments on the SAMSum dataset [7] and DialSumm dataset [25]. Both are large-scale open-domain chitchat datasets for dialogue summarization, which cover a wide range of daily life topics, including schooling, work, medication, shopping, leisure, and travel. Note that the SAMSum dataset contains natural messenger-like conversations created and written by linguists fluent in English. The standard dataset was split into 14,732, 818, and 819 examples for training, validation, and testing, respectively. The DialSumm dataset is collected from three public dialogue corpora, namely Dailydialog [76], DREAM [77], and MuTual [78], as well as an English speaking practice website. The standard dataset was split into 12,460, 500, and 500 examples for training, validation, and testing, respectively. The statistics of the datasets are listed in Table 1.

6.2. Implementation details

We used the large version of BART[2] to extract 1024-dimensional token features. The dialogue sequence and ground truth were truncated to 200 and 50 tokens, respectively. In all experiments, the dimension of the hidden units of LSTM was 256 for the sequence encoder. For graph encoders, the hidden size was set to 128, the size of the sliding window for sequential context dependency was selected from {1, 2, 3},[3] and the layer number was from {1, 2, 3, 4, 5}. At the test time, the minimum length of the generated summaries was set to 35, and the beam size was 5. In our experiments, we trained 50 epochs using the Adam optimizer with mini-batches, which could be easily parallelized on a single machine with multiple cores. The learning rate was varied within the range of {0.0005, 0.0007, 0.001, 0.005}, and the batch size was within the range of {8, 16, 32}. These optimal parameters were obtained based on the best performance on the validation set and were then used to evaluate the test set. The detailed settings are listed in Table 2.

Table 2
Hyperparameter settings.

Parameter	Parameter name	Value
l_d	maximum length of dialogue sequence	200
$l_{max,s}$	maximum length of ground truth	50
$l_{min,s}$	minimum length of generated summary	35
d_{bart}	pre-trained BART embedding size	1024
$d_{s,hidden}$	hidden size of sequence encoder (BiLSTM)	256
$d_{g,hidden}$	hidden size of graph encoders	128
n_f (n_p)	sliding window size	1
n_{iter}	layer number of graph encoders	3
n_{beam}	beam size	5
n_{step}	epoch	50
n_{batch}	batch size	8
λ	learning rate	0.001

6.3. Comparison methods

We compared our method with several baselines. Some sequence-based baselines are as follows:

Pointer Generator (PG): This model was proposed by See et al. [2]; it aids the accurate reproduction of information by pointing and retains the ability to produce new words through the generator.

Fast Abs RL Enhanced (FARE): This model was proposed by Chen et al. [79]; it constructs a hybrid extractive-abstractive architecture with policy-based reinforcement learning to bridge together the two networks and the names of all other speakers at the end of the utterances.

DynamicConv (DC): This model was proposed by Wu et al. [80]; it predicts a different convolution kernel at every time-step, and the dynamic weights are a function of the current time-step only rather than the entire context.

Multi-View Seq2Seq (M-BART): This model was proposed by Chen et al. [12]; it represents conversational structures from different views and utilizes a multi-view decoder to incorporate different views to generate dialogue summaries.

We also added some graph-based models for comparison:

Topic Interaction Graph2Seq (TIG): This model was proposed by Li et al. [18]; it generates summaries with a graph-to-sequence model and models the input news as a topic interaction graph.

Topic-word Guided Dialogue Graph Attention (TGDGA): This model was proposed by Zhao et al. [13]; it models the dialogue as an interaction graph according to the topic word information and uses the topic word features to assist the decoding process.

[2] In the experiments, we set the parameters of the BART model and trained our additional BiLSTM layer. Thus, the pre-trained features could be used, and the training cost could be significantly reduced.

[3] For the ith utterance, if the size of the sliding window for sequential context dependency was 1, the sliding window contained the $(i-1)$-th and $(i+1)$-th utterances, that is, n_p=1 and n_f=1.

Table 3
Main results in terms of ROUGE scores, FEQA, FCR-s, and FCR-r on the test set of SAMSum and DialSumm datasets. The results with * are directly obtained from Gliwa et al. [7] or Zhao et al. [13], and other results were reproduced by us. ($p < 0.05$ under t-test).

Model	SAMSum Dataset						DialSumm Dataset					
	R-1	R-2	R-L	FEQA	FCR-s	FCR-r	R-1	R-2	R-L	FEQA	FCR-s	FCR-r
PG [2]	40.09*	15.28*	36.63*	48.61	30.12	35.49	33.77	9.24	32.18	37.56	22.63	24.31
FARE [79]	41.95*	18.06*	39.23*	49.29	30.75	36.80	35.52	12.40	34.72	38.31	23.47	25.25
DC [80]	45.41*	20.65*	41.45*	55.47	33.82	40.07	38.65	15.37	36.89	45.08	28.11	31.39
M-BART [12]	49.35*	25.61*	47.73*	65.82	41.56	48.79	42.51	18.93	43.45	54.72	34.45	37.83
TIG [18]	42.14	18.42	39.81	53.14	33.49	39.55	36.27	12.06	35.50	42.16	26.08	30.42
TGDGA [13]	43.11*	19.15*	40.49*	59.75	38.06	43.91	36.38	12.71	35.84	48.72	31.23	34.89
S-BART [14]	48.70	24.88	47.24	66.83	43.77	50.14	41.23	18.06	42.61	56.30	36.72	40.65
MIDCK	**51.19**	**27.03**	**49.20**	**70.27**	**52.31**	**61.52**	**44.05**	**20.93**	**44.81**	**59.38**	**45.33**	**52.38**

Structure-aware seq2seq (S-BART): This model was proposed by Chen et al. [14]; it models the rich structures in conversations by incorporating discourse relations between utterances and action triples in utterances through structured graphs and designing a multi-granularity decoder to generate summaries by combining all levels of information.

6.4. Automated evaluation metrics

The summarization system is evaluated automatically in terms of three metrics:

ROUGE Metric Because ROUGE scores are the most widely used metrics for summarization, we employed the standard F1 scores for ROUGE-1, ROUGE-2, and ROUGE-L metrics [15] to measure the word overlaps between the generated summaries and references. These three metrics evaluate the accuracy of unigrams, bigrams, and the longest common subsequence, respectively.

FEQA Metric Considering that the ROUGE scores cannot measure factual errors in the summary, we attempted to use some factual metrics for measuring factual consistency, including the triple-based, textual entailment-based, and QA-based methods. The triple-based metrics count the relation triple overlaps between the generated summary and the source document. Textual entailment-based metrics, also known as natural language inference (NLI), aim to detect whether the source document could entail the generated summary. The QA-based metrics are all based on the intuition that if we ask questions about a summary and its source document, we will receive similar answers if the summary is factually consistent with the source document.

In this study, we chose a question answering (QA)-based metric, FEQA [35], for evaluating factual consistency in an unsupervised manner because the QA models have reading comprehension ability, and the FEQA metric is more interpretable than other metrics and performs well in factual consistency evaluation.[4] Given question–answer pairs from the generated summary, an off-the-shelf QA model was executed to extract answers from the source dialogue. The average F1 score of the answer from the generated summary and the answer from the source dialogue was considered as the faithfulness score for the generated summary.

FCR Metric To verify the quality of the facts contained in the generated summary, we propose two fact-level factual metrics, which are defined as follows:

$$\begin{aligned}\text{FCR-s} &= \frac{\mathcal{N}(s \cap h)}{\mathcal{N}(s)} \\ \text{FCR-r} &= \frac{\mathcal{N}(r \cap h)}{\mathcal{N}(r)}\end{aligned} \qquad (6)$$

where $\mathcal{N}(s \cap h)$ denotes the number of facts found in the generated summary that can find a match in the source dialogue, and $\mathcal{N}(r \cap h)$ denotes the number of facts found in the generated summary that can find a match in the reference. Note that FCR-s represents the rate of copying facts from the source dialogue, whereas FCR-r represents the facts that should be copied and are relevant.

7. Results and discussions

In this section, we introduce the experimental results on the SAMSum and DialSumm datasets for the abstractive dialogue summarization task. In addition, the performance of our models is demonstrated through detailed analysis and discussions.

7.1. Main results

Results on SAMSum The results of the baselines and our model on the SAMSum dataset are shown in Table 3. We evaluated our models using F1 scores for standard ROUGE metrics. We also performed a model analysis in terms of the FEQA and FCR metrics. For sequence-based baselines, FARE first uses extraction methods to select important sentences and then fuses these sentences to obtain novel summaries, which slightly improved the results compared with the PG model. Adding pre-trained embeddings or extra document training data to dynamic-weight convolution models also led to improved performance. The best-performing model, M-BART, significantly improved all ROUGE-based evaluation metrics via different structured views. In addition, the graph-based baselines, TIG and TGDGA, outperformed the sequence-based baselines without pre-trained embeddings. In particular, the TGDGA achieved high scores in the FEQA and FCR metrics, which suggests the effectiveness of topic-word information. It is worth noting that although the ROUGE scores of S-BART were slightly lower than those of M-BART, the performance in terms of FEQA and FCR scores was better, which benefited from the structured graphs in S-BART. In general, the experiments demonstrated that our models improved the performance by different margins. In particular, the MIDCK surpassed the best-performing graph-based model, S-BART, by 2.49, 2.15, 1.96, 3.44, 8.54, and 11.38 points for ROUGE-1, ROUGE-2, ROUGE-L, FEQA, FCR-s, and FCR-r, respectively. This demonstrated that the speaker-aware structure aids in modeling the conversation context to better capture the cross-sentence dependency, and the factual knowledge significantly improves the faithfulness of the summaries.

Results on DialSumm As shown in Table 3, the performance in terms of all evaluation metrics is far lower than that on the SAMSum dataset. As stated, DialSumm is more abstract, open-domain, and spoken analogous. Another possible reason for the lower performance on DialSumm is the longer input size. However, our model performed well, achieving 44.05, 20.93, 44.81, 59.38, 45.33, and 52.38 for ROUGE-1, ROUGE-2, ROUGE-L, FEQA,

[4] For entailment-based metrics, the out-of-the-box entailment models generalize poorly in downstream tasks and do not provide the required performance for factual consistency evaluation in text summarization. One reason is that the domain shifts from the NLI dataset to the summarization dataset. Second, NLI models often rely on heuristics, such as lexical overlap, to explain the high entailment probability.

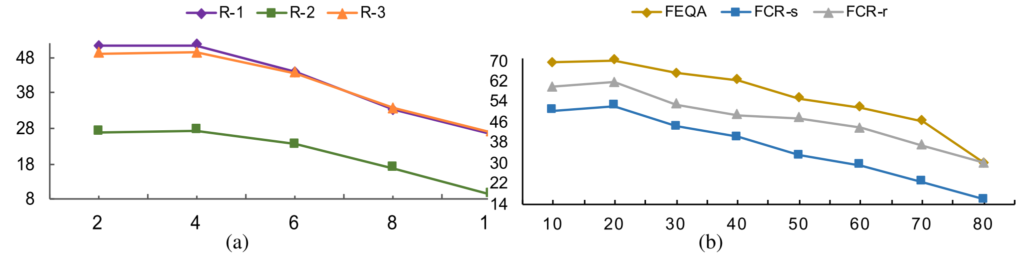


Fig. 5. (a) Impact of the number of speakers on ROUGE scores. (b) Impact of the number of fact tuples on FEQA, FCR-s, and FCR-r scores.

Table 4
Ablation studies for HDG and FG on test set of SAMSum dataset.

Model	R-1	R-2	R-L	FEQA	FCR_s	FCR_r
MIDCK	**51.19**	**27.03**	**49.20**	**70.27**	**52.31**	**61.52**
w/o FG	49.82	26.27	48.09	66.92	43.78	52.45
w/o HDG	49.58	25.92	47.86	67.31	48.15	56.88
w/o HDG&FG	48.43	24.65	46.88	65.27	41.01	47.63

FCR-s, and FCR-r, respectively. All these results demonstrated the improvement in the modeling of cross-sentence dependencies and proved the effectiveness of our proposed speaker-aware structure and factual knowledge. In addition, the significantly higher FEQA and FCR scores demonstrated that our model can not only correctly integrate the facts in the original dialogue, but also generate more facts relevant to the golden summary. This further proved that the cross-sentence dependencies were well aggregated.

7.2. Ablation study

We conducted ablation analysis on the HDG and FG, as shown in Table 4. First, we deleted one graph at a time and found that removing either of them leads to a significant drop in performance. In particular, the HDG significantly affected the ROUGE scores and resulted in a decrease of 1.61%, 1.11%, and 1.34%, respectively. The FG is more important for FEQA and FCR. In particular, its removal resulted in a drop of 8.53% for FCR-s and a drop of 9.07% for FCR-r. Removing both the HDG and FG resulted in very low ROUGE-1, ROUGE-2, ROUGE-L, FEQA, FCR-s, and FCR-r scores of 48.43%, 24.65%, 46.88%, 65.27%, 41.01%, and 47.63%, respectively, because the multi-speaker interaction could not be well modeled and the salient fragments in multiple utterances of dialogues failed to be integrated.

7.3. Quantitative analysis

Impact of Speakers and Fact Tuples Based on our MIDCK, the impact of two critical elements, i.e., the number of speakers and the number of fact tuples, is depicted in Fig. 5(a) and Fig. 5(b), respectively. As the number of speakers increases, ROUGE scores decrease, which has the same pattern as FEQA, FCR-s, and FCR-r scores for the number of fact tuples. This demonstrates that the greater the number of speakers and fact descriptions in the dialogue, the more complex is the information interaction process, which poses more challenges for the abstractness and faithfulness of the dialogue summarization model.

Impact of Fuse Method We conducted an experiment with three methods to fuse the SSC and ISS subgraphs of HDG for the SAMSum dataset. From Table 5, we observe that fusing the two subgraphs in the sequential strategy, that is, from the SSC subgraph to the ISS subgraph, achieves the best results. This verifies the logical inertia of people for information exchange in dialogues, that is, accepting or denying the views of others on the basis of holding their views.

Table 5
Results of three methods for fusing the self-speaker complete (SSC) subgraph and inter-speaker sparse (ISS) subgraph.

Fuse method	R-1	R-2	R-L	FEQA	FCR-s	FCR-r
SSC→ISS	**51.19**	**27.03**	**49.20**	**70.27**	**52.31**	**61.52**
ISS→SSC	50.06	26.17	48.25	68.83	51.16	60.37
Parallel	50.04	26.16	48.22	68.79	50.98	60.12

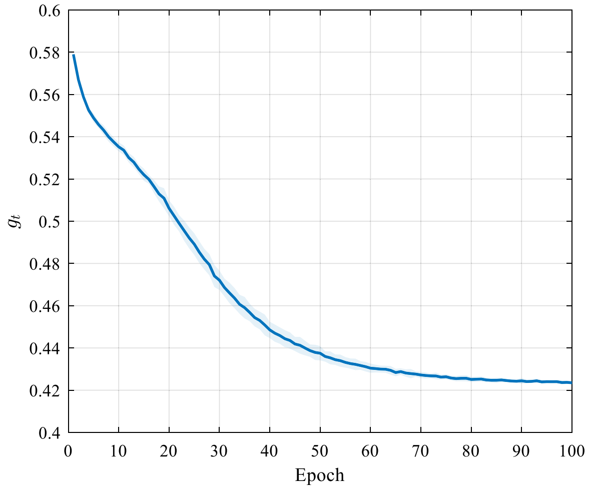


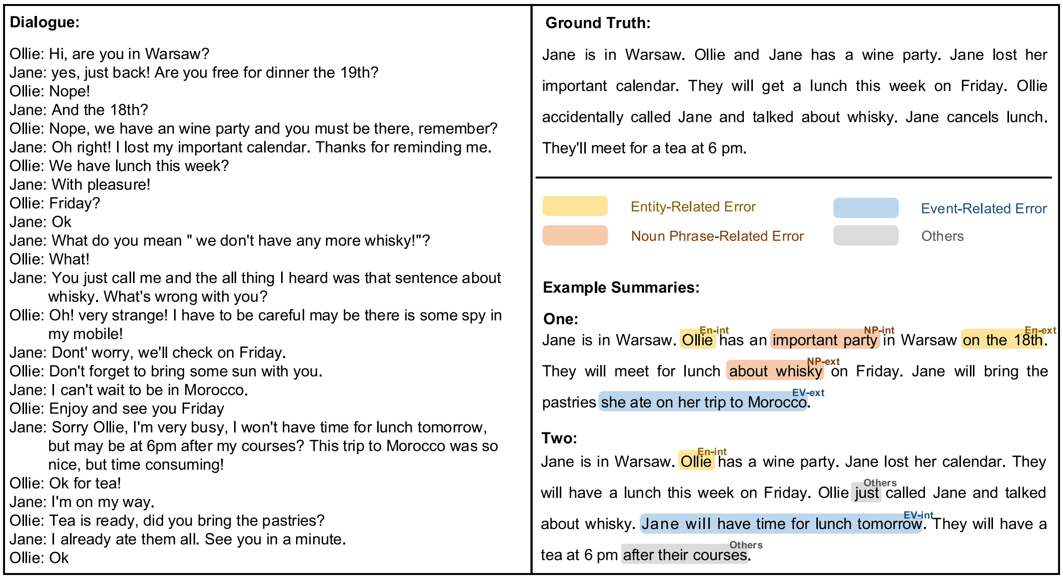
Fig. 6. Gate changes during training on the validation set of the SAMSum dataset. Shaded area spans±std.

7.4. Gate analysis

We investigated what the gate network (Eq. (4)) learned and the ratio in which $\mathbf{c}^s$ and $\mathbf{c}^g$ were combined. Fig. 6 shows the changes in the gate value g_t for the validation set during training. Initially, the average g_t exceeded 0.5, which suggested that the generated content was biased toward choosing source dialogues. As the training progressed, our model gradually learnt that fact descriptions were more reliable, which led to a consecutive drop in g_t. Finally, the average g_t gradually stabilized at 0.422. Notably, the ratio of g_t is $(1 - 0.422)/0.422 \approx 1.37$, which is very close to the ratio of copying proportions $0.44/0.32 \approx 1.38$, as shown in Table 1, indicating that our model predicts copy proportions relatively accurately and normalizes them as the gate value.

7.5. Error analysis of factual inconsistency

We identified four categories of errors defined by Goyal et al. [81] through manual inspection, including entity-related, event-related, noun phrase-related, and other errors. Each of these categories is further divided into intrinsic (errors that arise because of misinterpreting information from the source dialogue) and extrinsic (errors that hallucinate new information or facts



Dialogue:

Ollie: Hi, are you in Warsaw?
Jane: yes, just back! Are you free for dinner the 19th?
Ollie: Nope!
Jane: And the 18th?
Ollie: Nope, we have an wine party and you must be there, remember?
Jane: Oh right! I lost my important calendar. Thanks for reminding me.
Ollie: We have lunch this week?
Jane: With pleasure!
Ollie: Friday?
Jane: Ok
Jane: What do you mean " we don't have any more whisky!"?
Ollie: What!
Jane: You just call me and the all thing I heard was that sentence about whisky. What's wrong with you?
Ollie: Oh! very strange! I have to be careful may be there is some spy in my mobile!
Jane: Dont' worry, we'll check on Friday.
Ollie: Don't forget to bring some sun with you.
Jane: I can't wait to be in Morocco.
Ollie: Enjoy and see you Friday
Jane: Sorry Ollie, I'm very busy, I won't have time for lunch tomorrow, but may be at 6pm after my courses? This trip to Morocco was so nice, but time consuming!
Ollie: Ok for tea!
Jane: I'm on my way.
Ollie: Tea is ready, did you bring the pastries?
Jane: I already ate them all. See you in a minute.
Ollie: Ok

Ground Truth:

Jane is in Warsaw. Ollie and Jane has a wine party. Jane lost her important calendar. They will get a lunch this week on Friday. Ollie accidentally called Jane and talked about whisky. Jane cancels lunch. They'll meet for a tea at 6 pm.

Example Summaries:

One:

Jane is in Warsaw. Ollie has an important party in Warsaw on the 18th. They will meet for lunch about whisky on Friday. Jane will bring the pastries she ate on her trip to Morocco.

Two:

Jane is in Warsaw. Ollie has a wine party. Jane lost her calendar. They will have a lunch this week on Friday. Ollie just called Jane and talked about whisky. Jane will have time for lunch tomorrow. They will have a tea at 6 pm after their courses.

Fig. 7. Taxonomy of error types considered in our manual annotation. On the left is an example of a dialogue, and on the right are the corresponding ground truth and the two manually constructed examples of summaries. En-int, En-ext, Ev-int, Ev-ext, NP-int, and NP-ext denote the entity-related intrinsic error, entity-related extrinsic error, event-related intrinsic error, event-related extrinsic error, noun phrase-related intrinsic error, and noun phrase-related extrinsic error, respectively.

not present in the source dialogue) [29]. The guidelines of the taxonomy for manual annotation are as follows:

(1) Entity-Related: Errors specifically related to the surface realization of named entities, quantities, dates, etc. Incorrectly identifying distinct entities from the source dialogue is an intrinsic error (“*Ollie*” in Fig. 7) and the hallucination of new entities is an extrinsic error (“*on the 18th*” in Fig. 7).

(2) Event-Related: Errors with incorrect claims about events in the summary, such as predicates with arguments filled by incorrect entities. The mixed attributes from within the source dialogue are intrinsic (“*Jane will have time for lunch tomorrow*” in Fig. 7) and hallucinations of new events are extrinsic (“*she ate on her trip to Morocco*” in Fig. 7).

(3) Noun Phrase-Related: Errors related to noun phrases other than the entity-specific errors. Combining with an incorrect modifier from the dialogue is intrinsic (“*important party*” in Fig. 7) and hallucinations of new noun phrase modifiers are extrinsic (“*about whisky*” in Fig. 7).

(4) Other Errors: Errors such as ungrammatical text, repeated words, highly erroneous spans, etc., that do not fall into one of the aforementioned categories. These are not broken down as intrinsic and extrinsic (“*after their courses*” in Fig. 7).

For the two datasets, we used the aforementioned taxonomy to manually annotate 100 examples generated by S-BART and MIDCK. For SAMSum, as shown in Fig. 8(a), 56% of the summaries generated by S-BART contain inconsistencies and 45% of these are intrinsic, indicating that the current model lacks the ability to integrate the salient elements of an event in the dialogue. The number of intrinsic errors in the summaries generated by our model was significantly reduced to 19% (Fig. 8(b)) because our FG directly integrated factual knowledge. For DialSumm, 83% of the summaries generated by S-BART were unfaithful, with the majority of errors being extrinsic (Fig. 8(c)). As shown in Fig. 8(d), the MIDCK only reduces all errors by 10%, which indicates that the DialSumm dataset is more abstract and more difficult for our model to handle.

7.6. Human evaluation

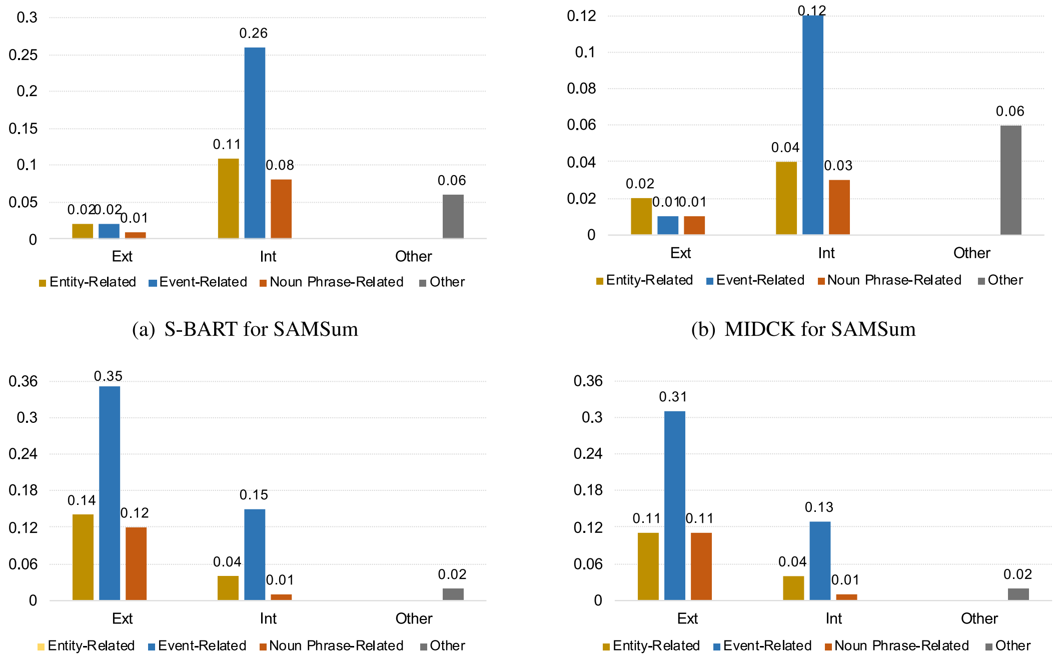
We conducted a manual evaluation to assess the models. We randomly selected 100 samples from the test set of the SAMSum dataset and hired five native speakers of English to rate the references and summaries generated by different models. Each judge rated summaries from 1 (worst) to 5 (best) on fluency, grammaticality, informativeness, and consistency. Each example is rated by three workers. The scores for each summary were averaged. The intra-class correlation was 0.543, showing moderate agreement [82]. As shown in Table 6, the TGDGA does not perform well in these four metrics, that is, the overall quality of the summary is low. S-BART achieved high scores in fluency, grammaticality, and informativeness; however, the factual consistency score was low. This suggests that fact inconsistency is a significant problem in the current dialogue summarization system. For our MIDCK, significant improvement was achieved in all metrics, particularly in fact consistency. In addition, the removal of HDG and FG had a significant impact on the consistency and the results decreased by 0.27 and 0.51, respectively, which indicated that HDG and FG can effectively integrate the cross-sentence scattered salient information in the dialogue to solve the problem of fact inconsistency.

Table 6
Human evaluation on fluency (Flu.), grammaticality (Gra.), informativeness (Inf.), and consistency (Con.) for the SAMSum dataset.

Model	Flu.	Gra.	Inf.	Con.
Ground Truth	**4.73**	**4.69**	**4.86**	**4.21**
TGDGA	3.76	3.27	3.91	3.02
S-BART	4.35	3.82	4.24	3.47
MIDCK	**4.39**	**3.95**	**4.46**	**4.03**
w/o HDG	4.26	3.80	4.33	3.76
w/o FG	4.30	3.93	4.28	3.52
w/o HDG&FG	4.16	3.69	4.11	3.08

7.7. Case study

Fig. 9 shows three examples of abstractive dialogue summarization from the SAMSum dataset. In example one, the summary generated by S-BART does not show the meaning of “*fast*”, which belongs to noun phrase-related errors. In example two, S-BART only extracts the speakers of “*Chloe*” and “*Riley*”, and



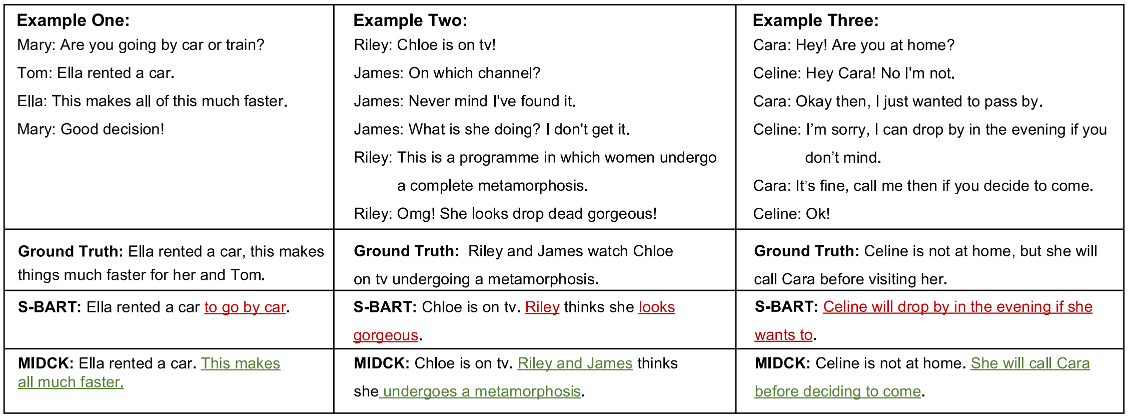
(a) S-BART for SAMSum

(b) MIDCK for SAMSum

(c) S-BART for DialSumm

(d) MIDCK for DialSumm

Fig. 8. Fractions of examples in two datasets exhibiting different error types.



Example One: Mary: Are you going by car or train? Tom: Ella rented a car. Ella: This makes all of this much faster. Mary: Good decision!	**Example Two:** Riley: Chloe is on tv! James: On which channel? James: Never mind I've found it. James: What is she doing? I don't get it. Riley: This is a programme in which women undergo a complete metamorphosis. Riley: Omg! She looks drop dead gorgeous!	**Example Three:** Cara: Hey! Are you at home? Celine: Hey Cara! No I'm not. Cara: Okay then, I just wanted to pass by. Celine: I'm sorry, I can drop by in the evening if you don't mind. Cara: It's fine, call me then if you decide to come. Celine: Ok!
Ground Truth: Ella rented a car, this makes things much faster for her and Tom.	**Ground Truth:** Riley and James watch Chloe on tv undergoing a metamorphosis.	**Ground Truth:** Celine is not at home, but she will call Cara before visiting her.
S-BART: Ella rented a car to go by car.	**S-BART:** Chloe is on tv. Riley thinks she looks gorgeous.	**S-BART:** Celine will drop by in the evening if she wants to.
MIDCK: Ella rented a car. This makes all much faster.	**MIDCK:** Chloe is on tv. Riley and James thinks she undergoes a metamorphosis.	**MIDCK:** Celine is not at home. She will call Cara before deciding to come.

Fig. 9. Case study on three examples from the SAMSum dataset.

ignores "*James*", which is an entity-related error. The fact description is "*Chloe undergoes a metamorphosis*" rather than "*she looks gorgeous*", which is an event-related error. The meaning of the summaries generated by S-BART for example three is completely different from the reference, which is also an event-related error. In addition, we can observe that most of the summaries generated by S-BART only arise from a single utterance, whereas the summaries generated by MIDCK in the three examples integrate the key information from different utterances in the original dialogue, which proves that our model has the ability to handle cross-sentence dependencies. Our MIDCK avoids these errors to some extent. By using the speaker-aware structure, our model constructs an HDG to capture long-distance cross-sentence dependencies and combines the fragments of events correctly. Furthermore, factual knowledge is enhanced via the dual-copy mechanism, which generates summaries that identify more key information and contain more faithful facts. However, in the summary generated by MIDCK in example two, the verb "*think*" weakens the original statement "*watch*", an event-related error, which shows that there are certainly some unresolved errors. We further analyze the current issues and challenges of existing summarizers in detail (see Section 8).

8. Challenges

Through the analysis of cases generated by our model, we summarize three challenges that still exist in the abstractive dialogue summarization task:

(1) Missing Information: The key elements, e.g., entities, and the description of events mentioned in golden summaries are missing in the generated summaries. For example, the missing element of person's name "*for her and Tom*" in example one in Fig. 9 and the missing event "*Riley and James watch Chloe on tv*" in example two in Fig. 9.

(2) Lack of Reasoning: The current model still lacks the context reasoning capability while facing multi-turn negotiation, which only memorizes the surface words or phrases but disregards the logical relationships between conversations, leading to the wrong selection and combination of key elements. For example, the wrong representation "*She will call Cara before deciding to come*" in example three in Fig. 9.

(3) Redundancy: The golden summaries do not mention what appears in the generated summaries, particularly some fact descriptions.

9. Conclusion

In this paper, we presented an MIDCK network for abstractive dialogue summarization. Our model utilizes the speaker-aware structure to construct an HDG, which can simulate the information interaction process of humans and handle cross-sentence dependencies. We also constructed an FG from the original dialogue to intuitively show the structured facts in the dialogue. In tandem with the graph information, our dual-copy decoder uses the tokens from the source sequence and factual knowledge from the FG to enhance the faithfulness of the generated summaries. To better evaluate the factual consistency, we proposed some fact-level factual consistency metrics. On two benchmark datasets, the SAMSum and DialSumm datasets, our proposal outperformed the strong baselines and the existing state-of-the-art model by a significant margin. Through the ablation study, quantitative analysis, gate analysis, error analysis of factual inconsistency, and case study, the effectiveness of each module of our model was further proved. Finally, we discussed the challenges that still exist in the current dialogue summarizers, which will serve as future research directions.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was partially supported by the Shandong Provincial Natural Science Foundation Key Project, China (ZR2020KF015), National Key R&D Program of China No. 2019YFF0303300 and Subject II No. 2019YFF0303302, DOCOMO Beijing Communications Laboratories Co., Ltd., China, MoE-CMCC "Artificial Intelligence" Project No. MCM20190701, BUPT, China Excellent Ph.D. Students Foundation CX2021204.

References

[1] A.M. Rush, S. Chopra, J. Weston, A neural attention model for abstractive sentence summarization, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 379–389, http://dx.doi.org/10.18653/v1/D15-1044, URL https://www.aclweb.org/anthology/D15-1044.

[2] A. See, P.J. Liu, C.D. Manning, Get to the point: Summarization with pointer-generator networks, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 1073–1083, http://dx.doi.org/10.18653/v1/P17-1099, URL https://www.aclweb.org/anthology/P17-1099.

[3] S. Gehrmann, Y. Deng, A. Rush, Bottom-up abstractive summarization, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 4098–4109, http://dx.doi.org/10.18653/v1/D18-1443, URL https://www.aclweb.org/anthology/D18-1443.

[4] E. Sharma, L. Huang, Z. Hu, L. Wang, An entity-driven framework for abstractive summarization, 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, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3280–3291, http://dx.doi.org/10.18653/v1/D19-1323, URL https://www.aclweb.org/anthology/D19-1323.

[5] S. Gao, X. Chen, Z. Ren, D. Zhao, R. Yan, From standard summarization to new tasks and beyond: Summarization with manifold information, in: C. Bessiere (Ed.), Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20, International Joint Conferences on Artificial Intelligence Organization, 2020, pp. 4854–4860, Survey track.

[6] H. Sacks, E.A. Schegloff, G. Jefferson, A simplest systematics for the organization of turn-taking for conversation, Language 50 (4) (1974) 696–735, URL http://www.jstor.org/stable/412243.

[7] B. Gliwa, I. Mochol, M. Biesek, A. Wawer, SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization, in: Proceedings of the 2nd Workshop on New Frontiers in Summarization, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 70–79, http://dx.doi.org/10.18653/v1/D19-5409, URL https://www.aclweb.org/anthology/D19-5409.

[8] C. Goo, Y. Chen, Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts, in: 2018 IEEE Spoken Language Technology Workshop, SLT, 2018, pp. 735–742, http://dx.doi.org/10.1109/SLT.2018.8639531.

[9] M. Li, L. Zhang, H. Ji, R.J. Radke, Keep meeting summaries on topic: Abstractive multi-modal meeting summarization, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 2190–2196, http://dx.doi.org/10.18653/v1/P19-1210, URL https://www.aclweb.org/anthology/P19-1210.

[10] Z. Liu, A. Ng, S. Lee, A.T. Aw, N.F. Chen, Topic-aware pointer-generator networks for summarizing spoken conversations, in: 2019 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU, 2019, pp. 814–821, http://dx.doi.org/10.1109/ASRU46091.2019.9003764.

[11] C. Liu, P. Wang, J. Xu, Z. Li, J. Ye, Automatic dialogue summary generation for customer service, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, in: KDD '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 1957–1965, http://dx.doi.org/10.1145/3292500.3330683.

[12] J. Chen, D. Yang, 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, Association for Computational Linguistics, 2020, pp. 4106–4118, http://dx.doi.org/10.18653/v1/2020.emnlp-main.336, Online URL https://www.aclweb.org/anthology/2020.emnlp-main.336.

[13] L. Zhao, W. Xu, J. Guo, Improving abstractive dialogue summarization with graph structures and topic words, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 437–449, http://dx.doi.org/10.18653/v1/2020.coling-main.39, URL https://www.aclweb.org/anthology/2020.coling-main.39.

[14] J. Chen, D. Yang, Structure-aware abstractive conversation summarization via discourse and action graphs, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2021, Online.

[15] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81, URL https://www.aclweb.org/anthology/W04-1013.

[16] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, E. Cambria, DialogueRNN: An attentive RNN for emotion detection in conversations, Proc. AAAI Conf. Artif. Intell. 33 (01) (2019) 6818–6825, http://dx.doi.org/10.1609/aaai.v33i01.33016818, URL https://ojs.aaai.org/index.php/AAAI/article/view/4657.

[17] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, A. Gelbukh, DialogueGCN: A graph convolutional neural network for emotion recognition in conversation, 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, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 154–164, http://dx.doi.org/10.18653/v1/D19-1015, URL https://www.aclweb.org/anthology/D19-1015.

[18] W. Li, J. Xu, Y. He, S. Yan, Y. Wu, X. Sun, Coherent comments generation for Chinese articles with a graph-to-sequence model, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 4843–4852, http://dx.doi.org/10.18653/v1/P19-1479, URL https://www.aclweb.org/anthology/P19-1479.

[19] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, D. McClosky, The stanford coreNLP natural language processing toolkit, in: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Baltimore, Maryland, 2014, pp. 55–60, http://dx.doi.org/10.3115/v1/P14-5010, URL https://www.aclweb.org/anthology/P14-5010.

[20] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, P. Wellner, The AMI meeting corpus: A pre-announcement, in: Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction, in: MLMI'05, Springer-Verlag, Berlin, Heidelberg, 2005, pp. 28–39, http://dx.doi.org/10.1007/11677482_3.

[21] Y. Mehdad, G. Carenini, R.T. Ng, Abstractive summarization of spoken and written conversations based on phrasal queries, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Baltimore, Maryland, 2014, pp. 1220–1230, http://dx.doi.org/10.3115/v1/P14-1115, URL https://www.aclweb.org/anthology/P14-1115.

[22] S. Banerjee, P. Mitra, K. Sugiyama, Abstractive meeting summarization using dependency graph fusion, in: Proceedings of the 24th International Conference on World Wide Web, in: WWW '15 Companion, Association for Computing Machinery, New York, NY, USA, 2015, pp. 5–6, http://dx.doi.org/10.1145/2740908.2742751.

[23] Y. Zou, L. Zhao, Y. Kang, J. Lin, M. Peng, Z. Jiang, C. Sun, Q. Zhang, X. Huang, X. Liu, Topic-oriented spoken dialogue summarization for customer service with saliency-aware topic modeling, in: AAAI, 2021.

[24] A. Fabbri, F. Rahman, I. Rizvi, B. Wang, H. Li, Y. Mehdad, D. Radev, ConvoSumm: Conversation summarization benchmark and improved abstractive summarization with argument mining, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, 2021, pp. 6866–6880, http://dx.doi.org/10.18653/v1/2021.acl-long.535, Online URL https://aclanthology.org/2021.acl-long.535.

[25] Y. Chen, Y. Liu, L. Chen, Y. Zhang, DialogSum: A real-life scenario dialogue summarization dataset, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, 2021, pp. 5062–5074, http://dx.doi.org/10.18653/v1/2021.findings-acl.449, Online URL https://aclanthology.org/2021.findings-acl.449.

[26] J. Liu, Y. Zou, H. Zhang, H. Chen, Z. Ding, C. Yuan, X. Wang, Topic-aware contrastive learning for abstractive dialogue summarization, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 1229–1243, URL https://aclanthology.org/2021.findings-emnlp.106.

[27] Z. Liu, N. Chen, Controllable neural dialogue summarization with personal named entity planning, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 92–106, URL https://aclanthology.org/2021.emnlp-main.8.

[28] L. Zhao, W. Zeng, W. Xu, J. Guo, Give the truth: Incorporate semantic slot into abstractive dialogue summarization, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 2435–2446, URL https://aclanthology.org/2021.findings-emnlp.209.

[29] J. Maynez, S. Narayan, B. Bohnet, R. McDonald, On faithfulness and factuality in abstractive summarization, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 1906–1919, http://dx.doi.org/10.18653/v1/2020.acl-main.173, Online URL https://www.aclweb.org/anthology/2020.acl-main.173.

[30] B. Goodrich, V. Rao, P.J. Liu, M. Saleh, Assessing the factual accuracy of generated text, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, in: KDD19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 166–175, http://dx.doi.org/10.1145/3292500.3330955.

[31] Y. Zhang, D. Merck, E. Tsai, C.D. Manning, C. Langlotz, Optimizing the factual correctness of a summary: A study of summarizing radiology reports, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 5108–5120, http://dx.doi.org/10.18653/v1/2020.acl-main.458, Online URL https://www.aclweb.org/anthology/2020.acl-main.458.

[32] T. Falke, L.F.R. Ribeiro, P.A. Utama, I. Dagan, I. Gurevych, Ranking generated summaries by correctness: An interesting but challenging application for natural language inference, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 2214–2220, http://dx.doi.org/10.18653/v1/P19-1213, URL https://www.aclweb.org/anthology/P19-1213.

[33] W. Kryscinski, B. McCann, C. Xiong, R. Socher, Evaluating the factual consistency of abstractive text summarization, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, Association for Computational Linguistics, 2020, pp. 9332–9346, http://dx.doi.org/10.18653/v1/2020.emnlp-main.750, Online URL https://www.aclweb.org/anthology/2020.emnlp-main.750.

[34] A. Wang, K. Cho, M. Lewis, Asking and answering questions to evaluate the factual consistency of summaries, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 5008–5020, http://dx.doi.org/10.18653/v1/2020.acl-main.450, Online URL https://www.aclweb.org/anthology/2020.acl-main.450.

[35] E. Durmus, H. He, M. Diab, FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 5055–5070, http://dx.doi.org/10.18653/v1/2020.acl-main.454, Online URL https://www.aclweb.org/anthology/2020.acl-main.454.

[36] Z. Cao, F. Wei, W. Li, S. Li, Faithful to the original: Fact aware neural abstractive summarization, in: Proceedings of the 32th AAAI Conference on Artificial Intelligence, 2017.

[37] C. Zhu, W. Hinthorn, R. Xu, Q. Zeng, M. Zeng, X. Huang, M. Jiang, Enhancing factual consistency of abstractive summarization, 2021, arXiv:2003.08612.

[38] H. Li, J. Zhu, J. Zhang, C. Zong, Ensure the correctness of the summary: Incorporate entailment knowledge into abstractive sentence summarization, in: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 1430–1441, URL https://www.aclweb.org/anthology/C18-1121.

[39] Y. Dong, S. Wang, Z. Gan, Y. Cheng, J.C.K. Cheung, J. Liu, Multi-fact correction in abstractive text summarization, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, Association for Computational Linguistics, 2020, pp. 9320–9331, http://dx.doi.org/10.18653/v1/2020.emnlp-main.749, Online URL https://www.aclweb.org/anthology/2020.emnlp-main.749.

[40] M. Cao, Y. Dong, J. Wu, J.C.K. Cheung, Factual error correction for abstractive summarization models, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, Association for Computational Linguistics, 2020, pp. 6251–6258, http://dx.doi.org/10.18653/v1/2020.emnlp-main.506, Online URL https://www.aclweb.org/anthology/2020.emnlp-main.506.

[41] T.-H. Wen, D. Vandyke, N. Mrkšić, M. Gašić, L.M. Rojas-Barahona, P.-H. Su, S. Ultes, S. Young, A network-based end-to-end trainable task-oriented dialogue system, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 438–449, URL https://aclanthology.org/E17-1042.

[42] C. Tao, W. Wu, C. Xu, W. Hu, D. Zhao, R. Yan, One time of interaction may not be enough: Go deep with an interaction-over-interaction network for response selection in dialogues, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 1–11, http://dx.doi.org/10.18653/v1/P19-1001, URL https://aclanthology.org/P19-1001.

[43] Q. Jia, Y. Liu, S. Ren, K. Zhu, H. Tang, Multi-turn response selection using dialogue dependency relations, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, Association for Computational Linguistics, 2020, pp. 1911–1920, http://dx.doi.org/10.18653/v1/2020.emnlp-main.150, Online URL https://aclanthology.org/2020.emnlp-main.150.

[44] Z. Lin, D. Cai, Y. Wang, X. Liu, H. Zheng, S. Shi, The world is not binary: Learning to rank with grayscale data for dialogue response selection, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, Association for Computational Linguistics, 2020, pp. 9220–9229, http://dx.doi.org/10.18653/v1/2020.emnlp-main.741, Online URL https://aclanthology.org/2020.emnlp-main.741.

[45] L. Shen, Y. Feng, H. Zhan, Modeling semantic relationship in multi-turn conversations with hierarchical latent variables, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 5497–5502, http://dx.doi.org/10.18653/v1/P19-1549, URL https://aclanthology.org/P19-1549.

[46] X. Gao, Y. Zhang, M. Galley, C. Brockett, B. Dolan, Dialogue response ranking training with large-scale human feedback data, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, Association for Computational Linguistics, 2020, pp. 386–395, http://dx.doi.org/10.18653/v1/2020.emnlp-main.28, Online URL https://aclanthology.org/2020.emnlp-main.28.

[47] J. Xu, H. Wang, Z.-Y. Niu, H. Wu, W. Che, T. Liu, Conversational graph grounded policy learning for open-domain conversation generation, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 1835–1845, http://dx.doi.org/10.18653/v1/2020.acl-main.166, Online URL https://aclanthology.org/2020.acl-main.166.

[48] W. Du, A.W. Black, Boosting dialog response generation, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 38–43, http://dx.doi.org/10.18653/v1/P19-1005, URL https://aclanthology.org/P19-1005.

[49] H. Su, X. Shen, S. Zhao, Z. Xiao, P. Hu, R. Zhong, C. Niu, J. Zhou, Diversifying dialogue generation with non-conversational text, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 7087–7097, http://dx.doi.org/10.18653/v1/2020.acl-main.634, Online URL https://aclanthology.org/2020.acl-main.634.

[50] H. Kim, B. Kim, G. Kim, Will I sound like me? Improving persona consistency in dialogues through pragmatic self-consciousness, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, Association for Computational Linguistics, 2020, pp. 904–916, http://dx.doi.org/10.18653/v1/2020.emnlp-main.65, Online URL https://aclanthology.org/2020.emnlp-main.65.

[51] A. Boyd, R. Puri, M. Shoeybi, M. Patwary, B. Catanzaro, Large scale multi-actor generative dialog modeling, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 66–84, http://dx.doi.org/10.18653/v1/2020.acl-main.8, Online URL https://aclanthology.org/2020.acl-main.8.

[52] B.P. Majumder, H. Jhamtani, T. Berg-Kirkpatrick, J. McAuley, Like hiking? You probably enjoy nature: Persona-grounded dialog with commonsense expansions, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, Association for Computational Linguistics, 2020, pp. 9194–9206, http://dx.doi.org/10.18653/v1/2020.emnlp-main.739, Online URL https://aclanthology.org/2020.emnlp-main.739.

[53] Y. Ma, K.L. Nguyen, F.Z. Xing, E. Cambria, A survey on empathetic dialogue systems, Inf. Fusion 64 (2020) 50–70, http://dx.doi.org/10.1016/j.inffus.2020.06.011, URL https://www.sciencedirect.com/science/article/pii/S1566253520303092.

[54] P. Zhong, C. Zhang, H. Wang, Y. Liu, C. Miao, Towards persona-based empathetic conversational models, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, Association for Computational Linguistics, 2020, pp. 6556–6566, http://dx.doi.org/10.18653/v1/2020.emnlp-main.531, Online URL https://aclanthology.org/2020.emnlp-main.531.

[55] E.M. Smith, M. Williamson, K. Shuster, J. Weston, Y.-L. Boureau, Can you put it all together: Evaluating conversational agents' ability to blend skills, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 2021–2030, http://dx.doi.org/10.18653/v1/2020.acl-main.183, Online URL https://aclanthology.org/2020.acl-main.183.

[56] J. Tang, T. Zhao, C. Xiong, X. Liang, E. Xing, Z. Hu, Target-guided open-domain conversation, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 5624–5634, http://dx.doi.org/10.18653/v1/P19-1565, URL https://aclanthology.org/P19-1565.

[57] Z. Liu, H. Wang, Z.-Y. Niu, H. Wu, W. Che, T. Liu, Towards conversational recommendation over multi-type dialogs, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 1036–1049, http://dx.doi.org/10.18653/v1/2020.acl-main.98, Online URL https://aclanthology.org/2020.acl-main.98.

[58] T. Young, E. Cambria, I. Chaturvedi, H. Zhou, S. Biswas, M. Huang, Augmenting end-to-end dialogue systems with commonsense knowledge, 2018, URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16573.

[59] S. Ji, S. Pan, E. Cambria, P. Marttinen, P.S. Yu, A survey on knowledge graphs: Representation, acquisition, and applications, IEEE Trans. Neural Netw. Learn. Syst. (2021) 1–21, http://dx.doi.org/10.1109/TNNLS.2021.3070843.

[60] D. Bouchacourt, M. Baroni, Miss tools and mr fruit: Emergent communication in agents learning about object affordances, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 3909–3918, http://dx.doi.org/10.18653/v1/P19-1380, URL https://aclanthology.org/P19-1380.

[61] B. Hancock, A. Bordes, P.-E. Mazare, J. Weston, Learning from dialogue after deployment: Feed yourself, chatbot!, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 3667–3684, http://dx.doi.org/10.18653/v1/P19-1358, URL https://aclanthology.org/P19-1358.

[62] Y. Wang, S. Joty, M. Lyu, I. King, C. Xiong, S.C. Hoi, VD-BERT: A unified vision and dialog transformer with BERT, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, Association for Computational Linguistics, 2020, pp. 3325–3338, http://dx.doi.org/10.18653/v1/2020.emnlp-main.269, Online URL https://aclanthology.org/2020.emnlp-main.269.

[63] H. Le, D. Sahoo, N. Chen, S.C. Hoi, BiST: Bi-directional spatio-temporal reasoning for video-grounded dialogues, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, Association for Computational Linguistics, 2020, pp. 1846–1859, http://dx.doi.org/10.18653/v1/2020.emnlp-main.145, Online URL https://aclanthology.org/2020.emnlp-main.145.

[64] W. Li, W. Shao, S. Ji, E. Cambria, BiERU: Bidirectional emotional recurrent unit for conversational sentiment analysis, Neurocomputing 467 (2022) 73–82, http://dx.doi.org/10.1016/j.neucom.2021.09.057, URL https://www.sciencedirect.com/science/article/pii/S0925231221014351.

[65] D. Ghosal, N. Majumder, A. Gelbukh, R. Mihalcea, S. Poria, COSMIC: COmmonsense knowledge for emotion identification in conversations, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, 2020, pp. 2470–2481, http://dx.doi.org/10.18653/v1/2020.findings-emnlp.224, Online URL https://aclanthology.org/2020.findings-emnlp.224.

[66] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, R. Zimmermann, Conversational memory network for emotion recognition in dyadic dialogue videos, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 2122–2132, http://dx.doi.org/10.18653/v1/N18-1193, URL https://aclanthology.org/N18-1193.

[67] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, R. Zimmermann, ICON: Interactive conversational memory network for multimodal emotion detection, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 2594–2604, http://dx.doi.org/10.18653/v1/D18-1280, URL https://aclanthology.org/D18-1280.

[68] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, L.-P. Morency, Context-dependent sentiment analysis in user-generated videos, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 873–883, http://dx.doi.org/10.18653/v1/P17-1081, URL https://aclanthology.org/P17-1081.

[69] T. Young, V. Pandelea, S. Poria, E. Cambria, Dialogue systems with audio context, Neurocomputing 388 (2020) 102–109, http://dx.doi.org/10.1016/j.neucom.2019.12.126, URL https://www.sciencedirect.com/science/article/pii/S0925231220300758.

[70] T. Young, F. Xing, V. Pandelea, J. Ni, E. Cambria, Fusing task-oriented and open-domain dialogues in conversational agents, 2021, arXiv:2109.04137.

[71] K. Sun, S. Moon, P. Crook, S. Roller, B. Silvert, B. Liu, Z. Wang, H. Liu, E. Cho, C. Cardie, Adding chit-chat to enhance task-oriented dialogues, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2021, pp. 1570–1583, http://dx.doi.org/10.18653/v1/2021.naacl-main.124, Online URL https://aclanthology.org/2021.naacl-main.124.

[72] T. Zhao, A. Lu, K. Lee, M. Eskenazi, Generative encoder-decoder models for task-oriented spoken dialog systems with chatting capability, in: Proceedings of the 18th Annual SIGDial Meeting on Discourse and Dialogue, Association for Computational Linguistics, Saarbrücken, Germany, 2017, pp. 27–36, http://dx.doi.org/10.18653/v1/W17-5505, URL https://aclanthology.org/W17-5505.

[73] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019, arXiv:1910.13461.

[74] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, Y. Bengio, Graph attention networks, in: International Conference on Learning Representations, 2018, URL https://openreview.net/forum?id=rJXMpikCZ.

[75] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, 2016, arXiv:1409.0473.

[76] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, S. Niu, DailyDialog: A manually labelled multi-turn dialogue dataset, in: Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Asian Federation of Natural Language Processing, Taipei, Taiwan, 2017, pp. 986–995, URL https://aclanthology.org/I17-1099.

[77] K. Sun, D. Yu, J. Chen, D. Yu, Y. Choi, C. Cardie, DREAM: A Challenge data set and models for dialogue-based reading comprehension, Trans. Assoc. Comput. Linguist. 7 (2019) 217–231, http://dx.doi.org/10.1162/tacl_a_00264, arXiv:https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00264/1923111/tacl_a_00264.pdf.

[78] L. Cui, Y. Wu, S. Liu, Y. Zhang, M. Zhou, [MuTual]: A Dataset for multi-turn dialogue reasoning, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 1406–1416, http://dx.doi.org/10.18653/v1/2020.acl-main.130, Online URL https://aclanthology.org/2020.acl-main.130.

[79] Y.-C. Chen, M. Bansal, 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),

Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 675–686, http://dx.doi.org/10.18653/v1/P18-1063, URL https://www.aclweb.org/anthology/P18-1063.

[80] F. Wu, A. Fan, A. Baevski, Y. Dauphin, M. Auli, Pay less attention with lightweight and dynamic convolutions, in: International Conference on Learning Representations, 2019, URL https://openreview.net/forum?id=SkVhlh09tX.

[81] T. Goyal, G. Durrett, Annotating and modeling fine-grained factuality in summarization, 2021, arXiv:2104.04302.

[82] T.K. Koo, M.Y. Li, A guideline of selecting and reporting intraclass correlation coefficients for reliability research, J. Chiropr. Med. 15 (2) (2016) 155–163, http://dx.doi.org/10.1016/j.jcm.2016.02.012, URL https://www.sciencedirect.com/science/article/pii/S1556370716000158.