Fact-Driven Abstractive Summarization by Utilizing Multi-Granular Multi-Relational Knowledge

Qianren Mao, Jianxin Li, *Member, IEEE*, Hao Peng, Shizhu He, Lihong Wang, Philip S. Yu, *Life Fellow, IEEE*, and Zheng Wang

***Abstract*—Abstractive summarization generates a concise summary to capture the key ideas of the source text. This task underpins important applications like information retrieval, document comprehension, and event tracking. While much progress has been achieved, state-of-the-art summarization approaches often fail to generate high-quality summaries to reproduce factual details accurately. One of the key limitations of existing solutions is that they are primarily concerned about extracting facts from the source text but overlook other crucial factual information, such as the related time, locations, reasons, consequences, purposes, participants and involved parties. Furthermore, the current summarization frameworks are inadequate in modeling the complex semantic relations among facts and the corresponding factual information, leaving much room for improvement. This paper presents FFSUM, a novel summarization framework for exploiting multi-grained factual information to improve text summarization. To this end, FFSUM constructs an individual fine-grained factual graph with multiple relations among facts and the corresponding factual information. It employs a fact-driven graph attention network to integrate multi-granular factual representations at the encoding stage. It then uses a hybrid pointer network to retrieve factual pieces from the graph for the summary generation. We evaluate the FFSUM by applying it to two real-world datasets. Experimental results show that the FFSUM consistently outperforms a state-of-the-art approach across evaluation datasets.**

***Index Terms*—Fact consistency, graph neural network, language model, pointer network, text summarization.**

I. INTRODUCTION

BY CONDENSING long documents into a shorter form while preserving primary factual information, text-summarization underpins important applications like event tracking [1] and information retrieval [2]. Abstractive summarization aims to generate concise expressions as document summaries, similar to how humans summarize texts. The SEQ2SEQ framework [3] is a widely used abstractive summarization framework. Recent works attempted to enhance SEQ2SEQ by incorporating techniques like autoencoder-based pre-trained language models [4], autoregressive-based pre-trained language models [5], [6], or hybrid pre-trained models [7]–[9], leading to performance improvements.

While promising, prior approaches often produce imprecise summaries containing errors with utterly different semantics and meanings from the original text. This is because they fail to capitalize on the structured linguistic content existing in documents or can not explicitly model the dependencies between nested complex factual pieces [10]. Most recent works address this problem by introducing a fact-driven strategy [11]–[14]. The idea is to first extract factual pieces from the source text, such as fact triples (e.g., *somebody-do(be)-something*), and then encode them into the summarization framework to improve the generated summary.

Although representing a step forward, these recent works only consider coarse-granular factual pieces but overlook the corresponding details of a given fact. More detailed information in a precise summary should be composed of a multitude of fine-granular pieces of information since events/facts typically come with their arguments. These fine-granular details are first defined as facets by Prasojo *et al.*, [15], including time, locations, reasons, consequences, purposes, participants and involved parties. The existing models ignore these essential multi-granular factual information and produce imprecise summaries that confuse the end-users. As we will show in the paper, the multi-granular factual pieces (facts and facets) often provide helpful details and cannot be omitted.

Fig. 1 gives an example to illustrate the usefulness of multi-granular factual information. In this example, facts are mentions with factual information stating '*somebody-do(be)-something*' (coded with dash lines). These detailed information phrases (coded with colors) are denoted as facets, such as time '*(*around 5:30 p.m),' locations ('*in the northeastern state of Borno*'), or numeric values ('*more than 70 members*'). We see that facets can provide complementary information to a fact (i.e., an event in this example). For instance, the phrase '*in Damaturu*' is the location of the event '*suspected members attacked a military checkpoint*,'

Manuscript received August 2, 2021; revised December 22, 2021, February 25, 2022, and March 7, 2022; accepted March 7, 2022. Date of publication March 22, 2022; date of current version May 12, 2022. This work was supported by the National Natural Science Foundation of China under Grant U20B2053. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Rohit Prabhavalkar. *(Corresponding author: Jianxin Li.)*

Qianren Mao, Jianxin Li, and Hao Peng are with the Beijing Advanced Innovation Center for Big Data and Brain Computing and the State Key Laboratory of Software Development Environment, Beihang University, Beijing 100190, China (e-mail: maoqr@act.buaa.edu.cn; lijx@act.buaa.edu.cn; penghao@act.buaa.edu.cn).

Shizhu He is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100045, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 101408, China (e-mail: shizhu.he@nlpr.ia.ac.cn).

Lihong Wang is with the CNCERT/CC, Beijing 100029, China (e-mail: wlh@isc.org.cn).

Philip S. Yu is with the Department of Computer Science, University of Illinois Chicago, Chicago, IL 60607 USA (e-mail: psyu@cs.uic.edu).

Zheng Wang is with the School of Computing, University of Leeds, LS2 9JT Leeds, U.K. (e-mail: z.wang5@leeds.ac.uk).

Digital Object Identifier 10.1109/TASLP.2022.3161157

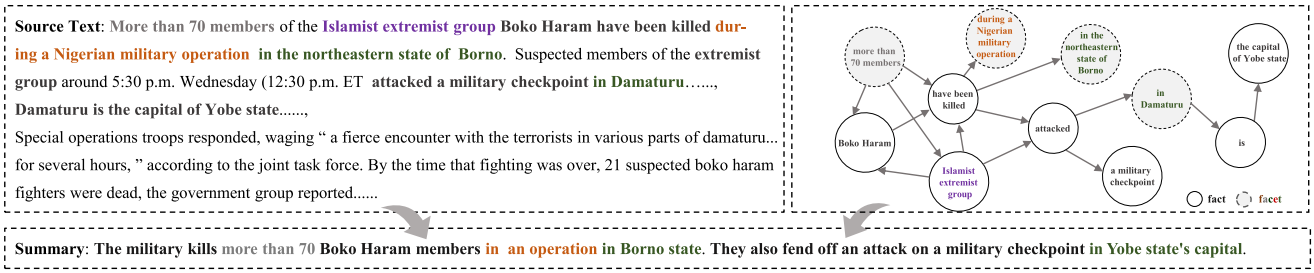


Fig. 1. Example of facts and facets in article and summary from the CNN/DailyMail dataset. In this case, event mentions are facts or events (colored with **DarkBlack**), which indicate '*somebody-do(be)-something*. Phrases (with other colors) are relevant complementary details of facts or events, e.g., **Brown**-colored detail phrase of '*during a Nigerian military operation*,' **DarkGreen**-colored location phrase of '*in the northeastern state of Borno*'.

and it is also the subject of the fact '*Damaturu is the capital of Yobe state*'. Hence, we argue that a better fact extraction method can be developed by modeling the relationship between the fact and facet, which can help the summarization system generate a more informative summary and avoid factual errors.

This paper thus presents a new fact-driven summarization system to explicitly model the facts and their facets. We do so by first employing a multi-granular information extraction tool [15] to obtain facts and facets from the source texts. We then construct an individual fine-grained factual graph with multiple relations for each source article which will be integrated into the summarization.

We present **FFSUM**, a novel framework to consolidate the fine-grained factual pieces[1] of source text into summarization. We implement FFSUM upon the BART [8], a state-of-the-art, SEQ2SEQ-based summarization framework. FFSUM leverages the BART's checkpoints to warm-start the generation framework. FFSUM enhances BART by utilizing a fact-driven graph attention network (FGAT) to integrate multi-granular fact representations at the encoding stage. FFSUM further employs a hybrid pointer (Ptr-Net) in the decoder for abstractive summarization. The hybrid pointer allows the generation framework to retrieve fact and facet knowledge from the factual graph and copy faithful tokens from the source article. By incorporating multi-granular factual pieces, FFSUM provides richer contexts to boost informativeness and factual correctness.

We evaluate FFSUM by applying it to two canonical abstractive summarization datasets, CNN/Daily Mail [16] and BBC XSUM [17]. We compare FFSUM against BART and various implementation variants. Experimental results show that FFSUM significantly outperforms alternative schemes by generating more informative and faithful summaries.

This paper makes the following contributions.

- It is the first to exploit multi-granular factual information (events/facts and their facets) for faithful text summarization.
- It develops a new fact-driven graph attention network to integrate factual information into summarization effectively.
- It shows how the graph encoding and hybrid pointer networks can be combined to collect multi-granular factual information for better text summarization.

II. Related Work

Our work builds upon the following past foundations but is different from them regarding summarization quality.

Pre-trained frameworks. Pre-trained language models have recently advanced a wide range of text summarization tasks. Since the SEQ2SEQ based Transformer [18] is naturally suitable for summary generation, almost all the language models' pre-trained checkpoints can be adapted to text generation and summarization. Owing to large amounts of unlabeled data and sufficient pre-training, language models can capture intricate world knowledge with informative language representations [19], [20]. The salient pre-trained frameworks for summarization include BERTSum [4], UniLM [7], and BART [8]. Very recently, Rothe *et al.*, [9] integrate pre-trained BERT, GPT-2, and RoBERTa checkpoints[2] to warm-start SEQ2SEQ based generation models. The warm-starting with pre-trained representations brings substantial improvements to generate informative summaries.

However, the superior performance is not a guarantee of a perfect system since existing models exhibit an inability to assure semantic-level consistency between the generated summary and source article. Factual inconsistency is a common problem that is hard to be avoided because neural abstractive approaches involve summary rewriting.

Graph-augmented summarization. Graph-based abstractive summarization works [21]–[23] explore augmenting SEQ2SEQ generative frameworks with structural graphs. Fernandes *et al.*, [24] introduce a graph model to integrate highly structured data such as entity relationships, molecules, and programs. To address factual-incorrectness, recent researchers use OpenIE to extract fact triples or construct factual knowledge graphs from the article to integrate them into encoding [12] or decoding process [13], via graph attention networks [25].

However, these methods are limited to OpenIE, which can only extract coarse-grained factual pieces as a series of fact/event mentions. It makes the summarization system unable to integrate

[1]Note that our goal is to generate summaries that do not conflict with the facts presented in the source documents but not to detect the authenticity of the facts in the source texts.

[2][Online]. Available: https://github.com/google-research/google-research/tree/master/bertseq2seq

TABLE I
USING SEMANTIC LABELS TO EXTRACT FACTS AND FACETS FROM THE CNN/DAILYMAIL DATASET. SUBJECTS AND OBJECTS ARE OBTAINED FROM THE EXTRACTED FACTS

Subject	Object	Post	Location	Temporal	Purpose	Manner	Other/Details	Source
16,439,732	16,387,937	1,475,181	1,565,244	1,312,802	1,151,674	1,069,083	720,483	275,465
Contrast	Profession	Attribution	Separation	Comparison	Numeric	Partwhole	Cause	Recipient
109,628	31,257	88,349	26,162	22,860	14,698	14,369	11,624	6,212

TABLE II
FACT AND FACET WITH SEMANTIC LABELS EXTRACTED BY STUFFIE WHICH ARE REPAIRED WITH OUR CORRECTION FOR SUMMARIES ON CNN/DAILYMAIL DATASET. SUBJECT AND OBJECT COME FROM EXTRACTED FACTS

Subject	Object	Post	Location	Temporal	Purpose	Manner	Other/Details	Source
1,926,476	1,913,447	211,593	191,531	172,494	134,489	115,914	63,696	34,678
Contrast	Profession	Separation	Comparison	Attribution	Numeric	Partwhole	Cause	Condition
12,153	3,647	2,907	2,802	2,421	2,345	1,277	1,025	18

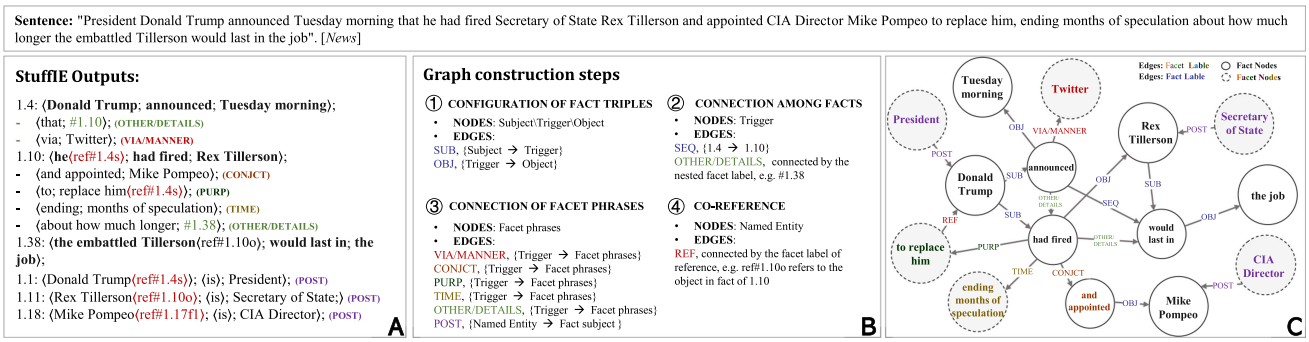


Fig. 2. An overview of the StuffIE output, our graph construction steps, and the constructed multi-granular multi-relational knowledge graph. We use the blue color to represent facts and other colors for facets.

fine-grained factual pieces and generate detailed fact/event arguments.

Pointer-generator network. Vinyals *et al.*, [26] first introduce the pointer network to select tokens from the input as an output rather than to pick tokens from a predefined vocabulary. The pointer mechanism has been used to create hybrid approaches for NMT [27], task-oriented dialogue [28], and summarization [29], [30]. It is also referred to as a copying mechanism [31], [32] in text generation, which can also choose tokens from the input sequence and put them at proper places in the output sequence. We have achieved a hybrid pointer used to copy tokens from the input sequence and retrieve tokens from graph nodes in this work which will be described in detail in next.

III. PRELIMINARIES

A. Factual Knowledge Extraction

Our work utilizes StuffIE,[3] a fine-grained information extraction tool to extract facts and facets. We construct the multi-granular factual graph by utilizing StuffIE to obtain the co-reference resolution among facts and facets. Compared to traditional information extraction tools like OpenIE [33], StuffIE has the advantages of supporting the extraction of multi-granular factual pieces containing facts and facets, as well as finer-grained information. The co-reference resolution provides nested relations between facts and facets, which is naturally suitable for graph construction without any fallible handcrafted rules. StuffIE also exploits existing SRL techniques to label facets with semantic roles, like via/manner, temporal, location and attribution, which are useful for our purpose.

1) Working Example: As a working example for factual pieces, consider Fig. 2A which shows a news sentence and the outputs extracted by the StuffIE from this sentence.

- Fact: The facts are associated with a serial number. In this example, facts of 1.4, 1.10 and 1.38 have a form of ⟨*a subject, predicate, an object*⟩.
- Facet: A facet has the form of ⟨*connector*; *content*⟩. The facet can be (1) verbless (e.g., *ending months of speculation*) or (2) verbal (e.g., *via Twitter*), and thus dependent on another fact (e.g., *that #1.10* which means nesting the next fact.). Each facet has a label that represents the semantic role corresponding to the fact. For the example given in Fig. 2, '*via Twitter*' is a facet of the fact '*Donald Trump announced Tuesday morning*' because it completes the fact's action. In other words, it answers the question, '*how did Donald Trump announce Tuesday morning*'. Other labels can be 'OTHER/DETAILS,' 'CONJUCTION, (CONJCT),' 'PURPOSE, (PURP),' 'TIME,' etc. It should be noted that there are also three specific facts, 1.1, 1.11, 1.18, in the StuffIE outputs, whose predicates are verb '*be*'. We treat these three cases as POST facets in our

[3][Online].Available: https://gitlab.inf.unibz.it/rprasojo/stuffie

graphs since these triples are always the reference for the subjects/objects of other facts.

2) Extraction Results: To improve the reliability of the extraction results, we set simple artificial semantic rules using SpaCy's NER[4] to match entity phrases and to rectify the incorrect labels. The StuffIE defines facet semantic labels using a connector. For example, if a fact and a facet are related via the connector 'because,' the facet is the 'reason' of the fact. However, some connector words (e.g., prepositions) have multiple semantics, which easily leads to errors. For example, in the sentence 'the Chung dropped their bid in August 2007,' the StuffIE identifies the facet 'in August 2007' with a 'LOCATION' label which should be a TIME label.

To correct these errors, we first align the facet label related to the StuffIE with SpaCy's NER labels. Then, we revise StuffIE's semantic rules to label the facet by matching the Entity label of SpaCy's NER results, e.g., identifying the 'TEMPORAL' of the facet phrase should also be 'TIME' of SpaCy's NER or should be corrected by the SpaCy's NER if not. Tables I and II show the final extracted results on articles and summaries on two datasets. Specifically, in the CNN/DailyMail dataset, many facts and facets have around 18 types, suggesting the source text has rich factual pieces. Except for facts composed of subjects and objects, the rest of this dataset contains facets, accounting for 20% of all factual knowledge in summaries.

B. Factual Graph Construction

The pseudocode in Fig. 2B shows the steps for constructing a factual-knowledge graph. One of such graphs is given in Fig. 2C, which has two types of edges. The first is the FACT LABEL edge for representing referential transfer among fact tokens. The FACT LABEL edge is used to connect two sequential facts, consisting of SUB and OBJ to connect internal triples of the fact, and SEQ to connect two facts. The second is the FACET LABEL edge for linking a fact with its facets. This edge only applies to facet labels and is visualized in colors in Fig. 2C.

1) The FACT LABEL edge: In this work, we use a FACT LABEL edge to connect the subject, object and predicated phrases within a fact triple. In other words, we use two edges labeled '*sbj*' or '*obj*' to connect the three nodes of a fact. For the graph example given in Fig. 2C, a '*sub*' label is going out from the predicate node of '*would last in*' to subject node of '*the embattled Tillerson*'. Similarly, the '*obj*' label is going out from the predicate node of '*would last in*' to object node of '*the job*'. For multiple facts, we connect two facts with a co-reference. For the example graph given in Fig. 2A, $ref\#1.4s$ which refers to '*he*' in the fact of 1.10 representing the subject node (*Donald Trump*) of the fact *1.4*. An alternative strategy to connect two facts is to indicate the sequential relation between two facts. To this end, we use a label '*Seq*' to connect the two facts directly, where the connection refers to a sequence of the two facts described in the original text. For our working example, the former predicate node '*had fired*' is connected with the second node '*would last in*' by the '*Seq*'.

[4][Online]. Available: https://github.com/explosion/spaCy

2) The FACET LABEL edge: We use the FACET LABEL to link a fact and its facets. In our working example, the predicate node '*had fired*' is connected to a facet node '*ending months of speculation*' by the edge of 'TIME,' a semantic role tag.

We use the steps described in Fig. 2B to obtain multi-granular and multi-relational factual-knowledge graphs for each article. As we will show later in the paper, our knowledge graphs can be integrated with a summarization system to improve the quality of summary generation.

C. Factual Correctness Evaluator

For summarization systems, a superior ROUGE performance is not a guarantee of a perfect system [34], [35]. Several studies [35]–[40] also observe an unexpected situation. The existing ROUGE-favoring summarization systems can generate highly informative summaries, but they are very likely to produce factual information of low faithfulness that is not complete enough, wrong, or even expresses a somewhat different semantic meaning. Cao *et al.*, [11] show that up to 30% of summaries generated by abstractive models contain incorrect facts. Recent works [13], [36], [38], [39], [41], [42] propose a model-based fact-correctness verification method. The verification of fact-checking is closely related to natural language inference (NLI) which can be regarded as a binary classification problem: a summary is either fact-consistent or fact-inconsistent with the article. Based on the NLI, Kryściński *et al.*, [36], Zhu *et al.*, [13] and Cao *et al.*, [42] propose fact evaluators, FactCC, FactCC$^+$ and FEC, respectively. FactCC and FactCC$^+$ are two BERT-based language inference models. FEC is a predictor with the assumption that a generated summary is inconsistent if it decides to be rectified.

Despite their positive effects, the above models are token-level fact-checking models and cannot evaluate the correctness of phrase-level facts and facets explicitly. We argue that the challenge is the lack of a training dataset related to fine-grained factual consistency evaluation. To generate training data of multi-granular factual samples, we sample *claims* from golden summaries and annotate their labels. Claims are then passed through textual transformations to generate positive and negative samples. A detailed text transformation algorithm of the data generation is presented in Algorithm 1. Compared with the FactCC [13], [36], there are two main differences in our algorithm. Our positive and negative samples come directly from the summaries (at least one sentence should be sampled). Multi-grained factual information is transformed to generate negative samples. Following Kryscinski *et al.*, [43], we adopt back translation to produce the positive samples by translating a sentence into an intermediate language, including French, German, Chinese, Spanish, and Russian and then translating them back to English. We swap fact and facet phrases to produce negative samples by displacing them in a claim to other fact and facet phrases in the articles, as shown in Fig. 3.

We create 1,441,800 document-claim pairs to train the fact-checking evaluator, out of which 50.66% are labeled as negative (`INCONSISTENT`), and the remaining 49.34% are labeled as positive (`CONSISTENT`). The constructed dataset of

Transformation	Original sentence	Transformed sentence
Black-translation	A Norwegian citizen of Somali descent is investigated in the Kenya mall attack .	In the Kenya shopping mall attack, a Norwegian citizen of Somali descent was investigated .
Subject swap	…Los Angeles Superior Court Judge **Upinder Kalra** said he had no choice but to set bail .	…Los Angeles Superior Court Judge **Carney** said he had no choice but to set bail .
Object swap	Carney told **the court Burkhart** would flee the country if he was allowed out…	Carney told **Kalra** would flee the country if he was allowed out…
Via/manner swap	France boosted their Euro 2012 qualifying campaign **with a 2-0 victory over Romania** .	France boosted their Euro 2012 qualifying campaign **with a victory at the Liberty stadium** .
Temporal swap	India 's rape problem needs a rewiring of society 's attitude **on November 18 .**	India 's rape problem needs a rewiring of society 's attitude **on September** .
Location swap	Russian navy soldier Gen. David Petraeus oversees U.S. operations **in the Middle East Asia** .	Russian navy soldier Gen. David Petraeus oversees U.S. operations **in the Russian** .
Attribution swap	American troop levels in Iraq peaked at 166,300 **according to the U.S. Defense Department** .	American troop levels in Iraq peaked at 166,300 **according to reports by CNN** .
Post swap	**Australian 8th seed** Samantha Stosur dumped out by Gisela Dulko .	**Defending champion** Samantha Stosur dumped out by Gisela Dulko .
Other swaps……		

Fig. 3. Factual claim examples of all transformations used to generate training data. Black-translation is a semantically invariant transformation. Swaps of 'SUBJECT,' 'OBJECT,' 'VIA/MANNER,' 'TEMPORAL,' 'LOCATION,' 'ATTRIBUTION,' and 'POST' are semantically variant transformation.

Algorithm 1: Texual Transformation Algorithm.

Require: $(\mathcal{A}, \mathcal{S})$ - set of source article-summary pairs
$\mathcal{T}^+$ - set of semantically invariant transformations
$\mathcal{T}^-$ - set of semantically variant transformations

function GENERATE _ DATA($(\mathcal{A}, \mathcal{S}), \mathcal{T}^+, \mathcal{T}^-$)
  $\mathcal{D} \leftarrow \emptyset$
  for $doc \in \mathcal{A}$ **do**
    $sum\_sent \leftarrow$ choose_summary$(\mathcal{A}, \mathcal{S})$
    $sent \leftarrow$ sentence_tokenizer(sum_sent)
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{(doc, sent, +)\}$
    for fn $\in \mathcal{T}^+$ **do**
      $new\_sent \leftarrow$ fn$(doc, sent)$
      $\mathcal{D} \leftarrow \mathcal{D} \cup \{(doc, new\_sent, +)\}$
    end for
  end for
  for $example \in \mathcal{D}$ **do**
    $(doc, sent, -) \leftarrow example$
    for fn $\in \mathcal{T}^-$
      $new\_sent \leftarrow$ fn$(doc, sent)$
      $\mathcal{D} \leftarrow \mathcal{D} \cup \{(doc, new\_sent, -)\}$
    end for
  end for
  return $\mathcal{D}$
end function

fact-checking on CNN/DailyMail has training samples with 1,225,530 document-claim pairs, 144,180 pairs for the validation set, and 72,090 pairs for the test. Those numbers are 613,150, 72,135, 36,078 for the fact-checking models in BBC XSUM dataset. Then, the document-claim pairs are fed as input to the BERT for classification.[5] We denote our fine-grained factual consistent corrector as **FFCC**[6] which is different from previous methods in two aspects:

- Our FFCC explicitly identifies multi-granular factual consistency in which those factual pieces consist not only of facts but also their facets.
- Our FFCC focuses on factual consistency on phrase-level multi-granular factual information beyond entity tokens.

[5]The two-way classification is realized by a single-layer classifier based on the hidden representation of [CLS] in BERT.

[6]Code and are available at: https://github.com/OpenSUM/FFCC

TABLE III
PERCENTAGE OF INCORRECTLY ORDERED SENTENCE PAIRS USING DIFFERENT CONSISTENCY PREDICTION MODELS

Model	Incorrect	
Random	50.00	
BERTNLI [37]	35.90	-14.10
ESIM [37]	67.60	-17.60
FactCC [36]	30.00	-20.00
FactCC$^+$ [13]	26.80	-23.20
QAGS [38]	27.90	-22.10
FEC [42]	26.80	-23.20
FFCC	25.77	-24.23

TABLE IV
PERFORMANCE OF FACT-CHECKING MODELS TESTED BY MEANS OF WEIGHTED (CLASS-BALANCED) ACCURACY AND F1 SCORE ON THE CONSTRUCTED TWO FACT CHECKING DATASETS

MODEL	CNN/DailyMail		XSUM	
	Weighted Accuracy	F1-score	Weighted Accuracy	F1-score
BERT+MNLI [36]	42.51	8.17	50.20	5.10
BERT+FEVER [36]	43.07	8.22	45.33	3.22
FactCC [36]	50.15	51.03	43.03	42.11
FactCC$^+$ [13]	63.80	63.77	42.11	40.76
FEC [42]	63.22	63.31	50.33	47.14
FFCC	65.55	64.37	51.51	48.07

To compare our FFCC to other fact-checking models, we conduct the sentence ranking experiment described by Falke *et al.*, [37] as other fact-checking models [13], [36] have done. This experiment is to verify how often a model assigns a higher probability of being correct to the positive rather than the negative claim. Results are presented in Table III, where our FFCC substantially outperforms other models in checking the correct sequential sentences. To further verify the ability of fact-checking, we evaluate the fine-grained factual consistency results in Table IV. Our FFCC models substantially outperform classifiers trained on the two fact-checking datasets constructed by the CNN/DailyMail and BBC XSUM. The accuracy result is generally below 0.50 on more abstractive BBC XSUM, showing the difficulty of understanding implicit factual information. The results also indicate that the current fact-checking models, such as FactCC and FEC, evaluate factual consistency at the entity level to a certain extent. However, they are not robust enough to verify multi-granular factual consistency. We use our superior FFCC for all verification of the fine-grained fact-checking in the following test of summarization framework.

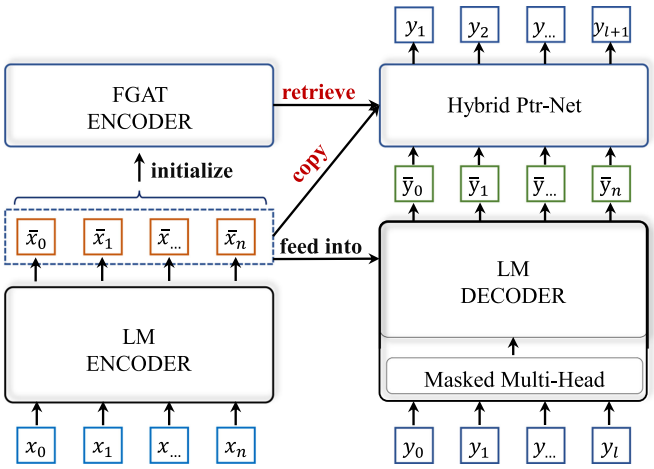


Fig. 4. The overview architecture of FFSUM for summarization. It is based on the pre-trained language model architecture of BART. The framework integrates two augmented sub-modules: a fact-driven graph encoder and a hybrid pointer network.

IV. Summarization Framework

A. Problem Formulation

In order to integrate the factual graph $\mathbf{G}_{1:m}$, our framework formalizes abstractive summarization as a supervised SEQ2SEQ problem to find a mapping of a text sequence $\mathbf{X}_{1:n}$ with invariant length n to an output sequence $\mathbf{Y}_{1:l}$ with variable length l. The source text $\mathbf{X}_{1:n}$ and graph $\mathbf{G}_{1:m}$ are consumed by a text encoder and a graph encoder to obtain token representations $\overline{\mathbf{X}}_{1:n}$ and node representations $\overline{\mathbf{G}}_{1:m}$, respectively. The summary decoder then generates tokens by computing the distribution over tokens in the vocabulary. The distribution can be factorized to a product of conditional probability distribution of the target token $\mathbf{y}_i$ with the $\overline{\mathbf{X}}_{1:n}$, $\overline{\mathbf{G}}_{1:m}$ and all previous generated tokens $\mathbf{Y}_{0:i-1}$. The probability distribution is produced by:

$$p_\theta(\mathbf{Y}_{1:l}) = \prod_{i=1}^{l} p_\theta(\mathbf{y}_i|\mathbf{Y}_{0:i-1}, \overline{\mathbf{X}}_{1:n}, \overline{\mathbf{G}}_{1:m}, \theta, \theta^*), \quad (1)$$

where θ is the parameters to be trained, and θ^* will be warm-started by the pre-trained language model for further fine-tuning.

B. Summarization Model

As shown in Fig. 4, our framework is based on the skeleton of the pre-trained language model, such as BART. BART uses a standard Transformer-based SEQ2SEQ architecture which, despite its simplicity, can be seen as generalizing BERT [44] (due to the bidirectional encoder) and GPT [6] (with the left-to-right decoder). Based on the SEQ2SEQ framework of BART, we integrate two sub-modules: a fact-driven graph encoder, FGAT, and a hybrid pointer, Hybrid Ptr-Net. The FGAT obtains graph nodes representations. The Hybrid Ptr-Net retrieves factual nodes from the graph and copies tokens from the article to generate faithful summaries.

C. Encoder

The FFSUM encoder contains two modules, BART encoder to learn text tokens' representations and FGAT to learn graph nodes' representations.

1) Pre-trained LM Encoder: To encode the input sequence $\mathbf{X}_{1:n}$ into a sequence of hidden states $\overline{\mathbf{X}}_{1:n}$, we feed the $\mathbf{X}_{1:n}$ to the BART encoder. Thus, we define the mapping:

$$f_{\theta_{\text{enc}}} : \mathbf{X}_{1:n}, \theta^* \rightarrow \overline{\mathbf{X}}_{1:n}, \quad (2)$$

where θ^* is parameters warm-started by BART's checkpoint and will be fine-tuned during the model training.

2) Fact-driven Graph Encoder: We use an undirected graph $\mathbf{G} = (\mathbf{V}; \mathbf{E})$ to represent the fine-grained factual nodes, where each node $v \in \mathbf{V}$ is associated with textual tokens. $e \in \mathbf{E}$ is the relation edge. We use BART output to initialize the representations of nodes by using the average embedding of their tokens. We also add the location embedding to each node to signify its original sequential location in the source text. All factual nodes representations are transformed by:

$$g_{\theta_{\text{enc}}} : \mathbf{G}_{1:m}, \theta' \rightarrow \overline{\mathbf{G}}_{1:m}, \quad (3)$$

where m is the total number of nodes, and θ' is warm-started by BART's output and trained by the graph encoder.

Since there exist different relations and nodes in the factual graph, we propose a series of Fact-driven Graph Attention Networks (FGAT) to learn node representations.

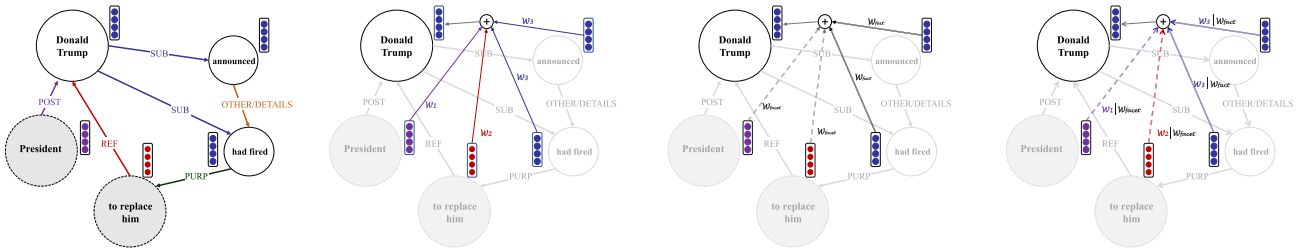
- Edge-type-aware GAT (**EGAT**). As shown in Fig. 5(b), the updating of factual nodes can be relation-specific transformations depending on the type of edges. Thus, EGAT is based on multiple relations to parameterize the weight matrices and to calculate nodes' attention over each edge. Thus, each node $\mathbf{g}_i$ is represented by a weighted average of its neighbors with different edge types:

$$\widehat{\mathbf{g}_i} = \mathbf{g}_i + ||_{k=1}^{K} \sigma \left(\sum_{e \in \mathcal{E}} \sum_{j \in N_i^e} a_{i,j}^k W_e^k \mathbf{g}_j \right), \quad (4)$$

  where N_i^e denotes the set of neighbors of node i with the edge type e and $e \in \mathcal{E}$, $\mathcal{E}$ is the set of all types of edges. W_e^k is the corresponding input linear transformation's weight matrix of edge type e. $a_{i,j}^k$ denotes the attention score between two nodes and $||_{k=1}^{K}$ denotes the concatenation of K heads.
- Node-type-aware GAT (**NGAT**). As shown in Fig. 5(c), factual nodes have two types (fact and facet) and their transformations can depend on the type of nodes. Thus, each node $\mathbf{g}_i$ is represented by a weighted average of its neighbors with different node types:

$$\widehat{\mathbf{g}_i} = \mathbf{g}_i + ||_{k=1}^{K} \sigma \left(\sum_{n \in \mathcal{N}} \sum_{j \in N_i^n} a_{i,j}^k W_n^k \mathbf{g}_j \right), \quad (5)$$

  where N_i^n denotes the set of neighbors of node i with the node type n. $n \in \mathcal{N}$ and $\mathcal{N}$ is all node types. W_n^k is weight matrix of node type n in k-th attention head.



(a) Original Graph (b) EGAT (c) NGAT (d) ENGAT

Fig. 5. Illustration of fact-driven graph attention networks (FGAT). The updating for a single node is calculated by aggregating its relational neighbors in the graph. The neighbor nodes are gathered by edge types (b), node types (c) and combining two types (d) for both ingoing and outgoing relations.

- Edge type & Node type combining-aware GAT (**ENGAT**). As shown in Fig. 5(d), the current node, 'Donald Trump,' has two facet neighbors. The two neighbors have different relations (edges) to the current node. The node representations of ENGAT are transformed by combining node types and edge types to parameterize the weight matrices:

$$\widehat{\mathbf{g}_i} = \mathbf{g}_i + ||_{k=1}^{K} \sigma \left(\sum_{n \in \mathcal{N}} \sum_{e \in \mathcal{E}} \sum_{j \in N_i^{ne}} a_{i,j}^k W_{ne}^k \mathbf{g}_j \right), \quad (6)$$

where N_i^{ne} denotes the set of neighbors of node i with different node types and edge types. W_{ne}^k is weight matrix of node type n nested edge type e.

D. Decoder

FFSUM decoder contains two modules, BART decoder to learn target representations and Hybrid Ptr-Net to copy source tokens and retrieve graph nodes for the generation.

1) BART Decoder: The BART decoder is a stack of autoregressive blocks. As shown in Fig. 4, the encoder's output $\overline{\mathbf{X}}_{1:n}$ is fed into this stack. Concurrently, the stack integrates previous generated sequence $\mathbf{Y}_{0:t-1}$ to produce the t-th target's hidden vector $\overline{\mathbf{y}}_t$. We define this kind of procedure as:

$$f_{\theta_{\text{dec}}} : \overline{\mathbf{X}}_{1:n}, \mathbf{Y}_{0:t-1}, \theta^* \rightarrow \overline{\mathbf{y}}_t, \quad (7)$$

where θ^* is parameters warm-started by BART's checkpoint and then will be fine-tuned by the downstream task.

2) Hybrid Ptr-Net: The Hybrid Ptr-Net is proposed to simultaneously copy tokens from the source text and retrieve fine-grained factual nodes (fact or facet nodes) from the graph:

$$f_{\theta_{\text{gen}}} : \overline{\mathbf{X}}_{1:n}, \overline{\mathbf{G}}_{1:m}, \overline{\mathbf{y}}_t, \mathbf{Y}_{0:t-1}, \theta'' \rightarrow \mathbf{y}_t, \quad (8)$$

where θ'' is randomly initialized. At each decoding step t, we first compute contextual vectors: $\mathbf{c}_t^{src} = \sum_i a_{i,t}^{src} \overline{\mathbf{x}}_i$ of source text, $\mathbf{c}_t^{grp} = \sum_j a_{j,t}^{grp} \widehat{\mathbf{g}}_j$ of factual nodes:

$$a_{i,t}^{src} = softmax(u^\top (W_1 \overline{\mathbf{y}}_t + W_2 \overline{\mathbf{x}}_i) + b_{src}), \quad (9)$$

$$a_{j,t}^{grp} = softmax(u^\top (W_3 \overline{\mathbf{y}}_t + W_4 \widehat{\mathbf{g}}_j) + b_{grp}), \quad (10)$$

where $\overline{\mathbf{x}}_i$ is the hidden representation of i-th token in encoder outputs and $\widehat{\mathbf{g}}_j$ is the j-th factual nodes' representations.

At last, the probability distribution over the vocabulary can be obtained by:

$$p_{vcb} = softmax\left(MLP[\overline{\mathbf{y}}_t || \mathbf{c}_t^{src} || \mathbf{c}_t^{grp}]\right). \quad (11)$$

In addition, the generation probability p_{gen} for timestep t is calculated from the two context vectors, the current decoder state $\overline{\mathbf{y}}_t$, and the embedding of previous token $\mathbf{y}_{t-1}$:

$$p_{gen} = \sigma\left(W_g \left[\overline{\mathbf{y}}_t || \mathbf{c}_t^{src} || \mathbf{c}_t^{grp} || \mathbf{y}_{t-1}\right] + b_g\right), \quad (12)$$

where σ is the sigmoid function. We further add a hybrid pointer to copy source text and to retrieve graph nodes for token prediction. The copy probability of $y_t = w$ is:

$$p_{copy} = \lambda_{src} \sum_{i:w_i=w} a_{i,t}^{src} + \lambda_{grp} \sum_{j:w_j=w} a_{j,t}^{grp}, \quad (13)$$

Thus, the model can learn to copy a important word w from different encoders by adjusting the gating weights λ_{src} and λ_{grp}. p_{copy} is the hybrid copy probability. Next, p_{gen} is used as a soft switch for generation by sampling from p_{vcb}, or sampling from p_{copy}: Next, p_{gen} is used as a soft switch for generation by sampling from p_{vcb} or p_{copy}:

$$P_{final} = p_{gen} p_{vcb} + (1 - p_{gen}) p_{copy}. \quad (14)$$

Among all the equations above, all W, u, and scalar of b are trainable parameters.

V. COMPARATIVE BASELINES

We compare our FFSUM against four classes of baselines: extractive methods, copy-based abstractive methods, language model-based methods and fact-aware methods.

A. Extractive methods

LEAD-3 uses the first three sentences of the article as its summary. TransformerEXT [4] is a neural extractive method that the encoder is the Transformer [18].

B. Copy-based abstractive methods

Ptr-Net [29] and its variant Ptr-Net+Cov [29] are the pointer generator networks without the coverage mechanism. GPG [30] is a generalized pointer that can either generate from the vocabulary or copy and edit some source words. BOTTOMUP [45] is a content selector that applies the copy mechanism to pre-select

phrases in an article during decoding. For language model-based methods, we compare those language model-oriented summarizers based on language pre-training by Transformer.

C. Language model-based methods

BERTSUMEXTABS [4] is a two-stage fine-tuned model based on BERT (first on an extractor, then on an abstractor). UniLM [7] is a unified BERT-Large pre-trained for bidirectional, unidirectional, and SEQ2SEQ language modeling objectives. BART [8] pre-trains a language model combining auto-encoder and auto-regressive Transformers. Other systematic pre-trained sequence generation methods are developed by Rothe *et al.*, [9]. They introduce another line of pre-trained models which are compatible with publicly available pre-trained GPT [6], BERT [44], and RoBERTa [46] checkpoints. These models contain GPT, RND2GPT, BERT2GPT, RND2RND, BERT2RND, RND2BERT, BERT2BERT, BERTSHARE, and RoBERTaSHARE. BERTSHARE and RoBERTaSHARE share the pre-trained parameters between the encoder and decoder, greatly reducing the memory footprint.

D. Fact-aware summarizers

We compare FASUM [13] with ASGARD [12]. FASUM extracts coarse-grained fact triples by OpenIE to build a graph and integrates it into the decoding process via neural graph computation. ASGARD[7] utilizes a graph encoder to encode those coarse-grained and entity-centered information.

VI. Experimental Setups

A. Datasets

We perform experiments on two popular datasets for single-document summarization: CNN/DailyMail [16] and BBC XSUM [17]. CNN/DailyMail contains online news with multiple sentence summaries and strongly favors extractive summarization. XSUM corpus provides a single-sentence summary for each BBC long story. XSUM needs to perform more information fusion and inference since the source is much longer than the target, and the summaries are more abstractive than CNN/DailyMail. We follow the preprocessing steps and experimental setups from prior work [12], [16], [29] for datasets. The CNN/DailyMail dataset consists of 287 k document-summary pairs, whereas the XSUM dataset includes 204 k pairs. During training, the input documents are truncated to 512 tokens for CNN/DailyMail and XSUM. The length of the summaries is limited to 128 tokens for CNN/DailyMail, 64 for XSUM. For CNN/DailyMail, the training, validation, and test samples are 287,188/13,367/11,490, respectively. For XSUM, the amounts of the three categories are 204,045/11,332/11,334.

B. Evaluation

We employ the official ROUGE F1 (version 1.5.5) as our evaluation metric. ROUGE-1 (R1) and ROUGE-2 (R2) are reported for informativeness and ROUGE-L (RL) for fluency. Additionally, the informativeness of a summary can be evaluated by the number of unique name entities in the generated text [47]. The informativeness is calculated by $\mathrm{INF.score} = \frac{\mathrm{No.}unique\_entity}{\mathrm{No.}summary}$. Taking the golden summary into account, we introduce a relative informativeness score, **RINF.score** $= \frac{\mathrm{INF.score}\_of\_Generation}{\mathrm{INF.score}\_of\_GoldenSummary}$. We utilize the SpaCy NER tagger to extract the entities from the summaries. The entities contain regular entities, names of persons and institutions, and numeric entities. Moreover, we leverage the phrase-level fact-checking model FFCC introduced in Section III-C to evaluate the factual correctness of summarization models. The fine-grained factual-consistency score is denoted as **FF.score**. To be fair, we also release the evaluation score of the token-aware fact-checking using FactCC$^+$ [13], denoted as FactCC.score. In the test of the generated summaries, the factual score is $f(A,S) = \frac{1}{k}\sum_{i-1}^{k} f(A, C_i)$. C_i is one of the sentences of summary S. $f(A, C_i)$ represents the probability that C_i is factually correct with respect to the article A. Besides, to verify the extractive property of the summarization systems, we measure the coverage (denoted as **Entity.COV**) of entities that exist in fact and facet pieces. We set the coverage function as $\mathrm{Entity.COV}(A,S) = \frac{1}{|S|}\sum_{e\in E(A,S)} |e|$, where $E(A,S)$ is the set of entities shared between an article A and its summary S. $|e|$ is the number of unique entities, and $|S|$ is that of tokens in the summary.

C. Training Details and Parameters

We use the base checkpoint of BART with 12 layers, a hidden size of 1024, and 12 attention heads. The model is fine-tuned on two datasets using Adam optimizer with a cross-entropy loss function. The ranges of the hyper-parameters are 1e-6, 5e-5, 3e-5, 2e-5 for learning rate for θ^* and 1e-2, 5e-2, 1e-3, 5e-3, 1e-4, 5e-4 for θ' and θ''. We use the toolkit NNI (Neural Network Intelligence)[8] to automatically run experiments' trial jobs to search the best hyper-parameters. We use a linear learning rate warmup with 20 k steps, normalization by the square root of the hidden size, and square root decay. BART trains with a dropout of 0.1 on all layers and attention weights and a GELU activation function. The training is done with a global batch size of 8 for CNN/DailyMail and BBC XSUM datasets. We set the beam size as 5 during generation and removed duplicated trigrams in beam search [48] on the validation set. All models are trained on four GPUs of Tesla V100-PCIE-32 GB with a distributed data-parallel trainer.

VII. Experimental Results

This section will analyze to show insights into the proposed FFSum by answering the following research questions.

- RQ1: What is the ROUGE performance of our FFSum?
- RQ2: How faithful is the generation of FFSum?
- RQ3: How does each sub-module (FGAT encoder and hybrid Ptr-Net) affect the model's overall performance?
- RQ4: What is the quality of the summaries generated by the model in terms of different metrics and case studies?

[7]To verify the performance being brought by its graph encoder, we use their model variant of 'ASGARD-DOC' as a baseline without additional reinforcement Learning and cloze reward.

[8][Online]. Available: https://github.com/microsoft/nni

TABLE V
ROUGE F1 RESULTS OF MODELS ON CNN/DAILYMAIL. WE COMPARE OUR MODEL (THE BOTTOM BLOCK) AGAINST EXTRACTIVE MODELS (THE TOP 1 BLOCK), COPY-BASED MODELS (THE TOP 2 BLOCKS), PRE-TRAINED TRANSFORMER MODELS (THE TOP 3 BLOCKS), AND FACT GRAPH AUGMENTED MODELS (THE PENULTIMATE BLOCK). STATISTICALLY, STATE-OF-THE-ART RESULTS ARE IN BOLD

Model	ROUGE-1	ROUGE-2	ROUGE-L
LEAD-3	40.42	17.62	36.67
TransformerEXT	40.90	18.02	37.17
Ptr-Net	36.44	15.66	33.42
Ptr-Net+Cov	39.53	17.28	36.38
GPG	40.95	18.01	37.46
BOTTOMUP	41.22	18.68	38.34
GPT	37.26	15.83	34.47
RND2GPT	32.08	8.81	29.03
BERT2GPT	25.20	4.96	22.99
RoBERTa2GPT	36.35	14.72	33.79
RND2RND	35.77	14.00	32.96
BERT2RND	38.74	17.76	35.95
RND2BERT	36.65	15.55	33.97
BERT2BERT	39.02	17.84	36.29
BERTSHARE	39.09	18.10	36.33
RoBERTa2RoBERTa	40.03	18.57	36.82
RoBERTaSHARE	40.10	18.95	37.39
BERTSUMEXTABS	42.13	19.60	39.18
UniLM	43.47	20.30	40.63
BART	44.16	21.28	40.90
FASUM	38.80	17.23	35.70
ASGARD	40.38	18.40	37.51
FFSUM (ENGAT)	**45.36**	**22.03**	**42.11**

TABLE VI
ROUGE F1 RESULTS OF MODELS ON BBC XSUM. WE CAREFULLY RE-IMPLEMENT THE GRAPH AUGMENTED METHOD OF ASGARD PROPOSED BY HUANG *ET AL.*, [12], SINCE THEY DID NOT VERIFY THE PERFORMANCE OF THEIR MODEL ON THE XSUM DATASET

Model	ROUGE-1	ROUGE-2	ROUGE-L
LEAD-3	16.30	1.61	11.95
Ptr-Net	29.70	9.21	23.24
Ptr-Net+Cov	28.10	8.02	21.72
BOTTOMUP	28.21	8.00	20.69
GPT	22.21	4.89	16.69
RND2GPT	28.48	8.77	22.30
BERT2GPT	27.79	8.37	21.91
RoBERTa2GPT	19.91	5.20	15.88
RND2RND	30.90	10.23	24.24
BERT2RND	38.42	15.83	30.80
BERT2BERT	37.53	15.24	30.05
BERTSHARE	38.52	16.12	31.13
RoBERTaSHARE	39.87	17.50	32.37
BERTSUMABS	38.76	16.33	31.15
BERTSUMEXTABS	38.81	16.50	31.27
BART	45.14	22.27	37.25
FASUM	28.60	8.97	22.80
ASGARD	39.82	17.11	32.51
FFSUM (ENGAT)	**45.72**	**22.73**	**37.84**

A. ROUGE Results (RQ1)

Table V lists the ROUGE results for all models on CNN/DailyMail dataset, where FFSUM outperforms all baselines. In particular, our FFSUM performs better than salient pre-trained models, such as BERTSUMEXTABS by a large margin in ROUGE, as shown in Table VI. The baselines combining auto-regressive or auto-encoding pre-trained language models, such as UniLM and BART, have consistently achieved robust performance. Their results demonstrate that the models with pre-trained representations can obtain high ROUGE performance. Despite such progress in pre-trained abstractive systems, our FFSUM achieves distinct improvements, compared with most pre-trained models, including the salient BART. Moreover, compared with FASUM and ASGARD, only integrating coarse-grained fact triples or graphs into summarization, our model achieves noticeable improvements. These results indicate that our fact-driven framework enhances the pre-trained models by integrating multi-granular factual pieces and can improve ROUGE performance. It should be noted that our FFSUM takes slightly longer inference time than BART as shown in Table VII, since it is stacked upon the BART model and has a larger number of parameters.

TABLE VII
THE MODEL SIZE, TRAINING TIME ($\mathcal{T}$.TRAIN), THE USAGE RATIO OF GPU MEMORY ($\mathcal{U}$.MEMORY) AND INFERENCE TIME ($\mathcal{T}$.INF) OF MODELS ON CNN/DAILYMAIL DATASET USING A SINGLE GPU OF TESLA V100-PCIE-32 GB. THE MODEL TRAINING IS ON FOUR GPUS

Model	Size	$\mathcal{T}$.train (/20k steps)	$\mathcal{U}$.memory	$\mathcal{T}$.inf (/sample)
BART	406.3 M	10.4 hours	45.3%	756.9ms
FFSUM (ENGAT)	410.1 M	11.2 hours	51.1%	872.1ms

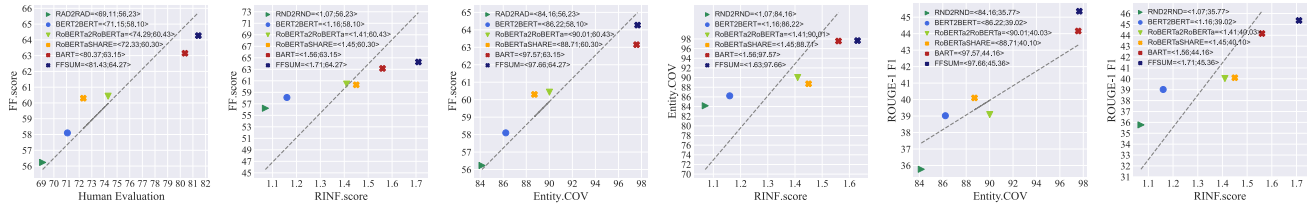
TABLE VIII
THE PERCENTAGE OF FINE-GRAINED FACTUAL-CONSISTENCY USING OUR EVALUATION MODEL FFCC OR USING FACTCC$^+$ [13]. THE TESTED SUMMARIES ARE GENERATIONS FROM CNN/DAILYMAIL

Model	FF.score	FactCC.score
RND2RND	56.23	52.19
BERT2BERT	58.10	55.42
RoBERTa2RoBERTa	60.43	58.75
RoBERTaSHARE	60.30	58.52
ASGARD	60.33	61.04
BART	63.15	62.21
FFSUM (ENGAT)	**64.27**	**63.59**

B. Factual-Consistency Performance (RQ2)

We select several salient pre-trained summary systems for factual consistency verification experiments. All generated summaries are tested by our evaluation model FFCC introduced in Section III-C. Although the RoBERTaSHARE is superior to RoBERTa2RoBERTa on ROUGE in Table V, its superiority is not maintained in factual correctness as shown in Table VIII. This comparison indicates that the ROUGE metric does not always reflect factual correctness, similar to what Zhu *et al.*, [13] have observed. Besides, compared with these five baselines, our FFSUM performs well in checking factual consistency by a large margin no matter testing by our fact-checking model (FFCC) or by FactCC on FactCC.score.

To further verify whether our approach can improve the factual consistency for a pre-trained model by stacking on it, we integrate our fact-driven sub-modules, FGAT and Hybrid Pre-Net, on other pre-trained encoder-decoder skeletons. The experimental results are shown in Table IX. Our approach generally improves about 1.80 in factual consistency performance on account of fine-grained factual information being consolidated during summarization.



(a) Human Evaluation VS. FF.score (b) RINF.score VS. FF.score (c) Entity.COV VS. FF.score (d) RINF.score VS. Entity.COV (e) Entity.COV VS. ROUGE-1 F1 (f) RINF.score VS. ROUGE-1 F1

Fig. 6. The comparison of neural summarization systems under ROUGE performance, fine-grained factual-consistency score (FF.score), the coverage score of entities (Entity.COV) and relative informativeness score (RINF.score) on CNN/DailyMail dataset. We show a regression line to distinguish differences by calculating the ordinary least squares (OLS).

TABLE IX
THE INFLUENCE DIFFERENT PRE-TRAINED GENERATION SKELETONS INTEGRATING WITH (W/) ON OUR FACT-DRIVEN SUB-MODULES

Model	ROUGE-1	FF.score
RND2RND w/	39.24 ↑ 3.47	60.01 ↑ 3.78
BERT2BERT w/	41.86 ↑ 2.84	59.33 ↑ 1.23
RoBERTa2RoBERTa w/	42.04 ↑ 2.01	61.82 ↑ 1.39
RoBERTaSHARE w/	42.13 ↑ 2.03	61.77 ↑ 1.47
BART w/	45.36 ↑ 1.20	64.27 ↑ 1.12
Average Improvement	2.31	1.80

TABLE X
INFLUENCE ON OUR SUB-MODULES BY TESTING ABLATION EXPERIMENTS. THE TESTED SUMMARIES ARE GENERATIONS FROM CNN/DAILYMAIL DATASET. 'W/O' MEANS 'WITHOUT'

Model	ROUGE-1	FF.score
FFSUM (GAT)	44.50 ↓ 0.86	63.33 ↓ 0.94
FFSUM (EGAT)	44.77 ↓ 0.59	63.77 ↓ 0.50
FFSUM (NGAT)	45.09 ↓ 0.27	63.80 ↓ 0.47
FFSUM (w/o copying source tokens)	45.11 ↓ 0.25	64.09 ↓ 0.18
FFSUM (w/o retrieving graph nodes)	44.41 ↓ 0.95	63.41 ↓ 0.86
FFSUM (w/o retrieving fact nodes)	44.53 ↓ 0.83	63.50 ↓ 0.77
FFSUM (w/o retrieving facet nodes)	45.00 ↓ 0.36	64.03 ↓ 0.24

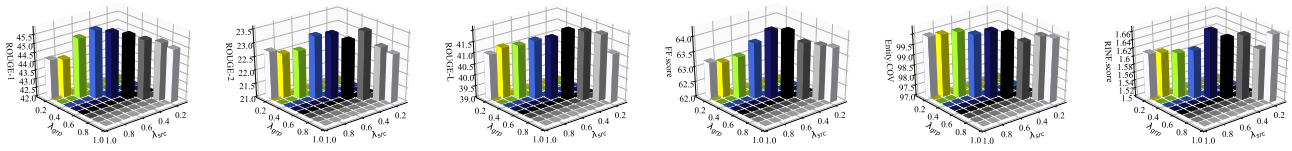
C. Ablation Study (RQ3)

To better understand the contribution of different sub-modules to the last performance, we conduct ablation studies using our proposed FFSUM model on CNN/DailyMail dataset. First, BART can be viewed as an ablation study of FFSUM eliminating the two fact-driven modules: FGAT encoder and Hybrid Ptr-Net. Without the fact-driven modules, the base model of BART suffers a noticeable performance loss, whether on ROUGE or FF.score. In addition, other ablation studies are without (w/o) copying source tokens and without retrieving graph nodes. The results are shown in Table X. Compared with the one without copying tokens from the source text, FFSUM w/o retrieving graph nodes has suffered great performance degradation, about 0.95 ROUGE drops, and 0.86 FF.score drops. Specifically, the model's performance that only copies the facet node decreases more than that that only copies the fact node. This is mainly because the amount of facts determines the main content. Moreover, we also analyze the impact of different FGAT models on overall performance, as shown in Table X. To equip with the vanilla GAT [25] in FFSUM is inferior to that equipped with edge-type-aware GAT (EGAT) and node-type-aware GAT (NGAT) or edge-type & node-type aware GAT (ENGAT). Integrated the ENGAT, our FFSUM distinguishes the semantics of different edges and nodes for fact-driven graph modeling, improving the overall effect. Based on the above results, we conclude: i) the hybrid Ptr-Net improves performance by copying words from the original texts and duplicating fact and facet phrases from the factual graphs. ii) the EGAT and NGAT boost the fact-driven encoder to generate expressive node representations by aggregating neighbor nodes through syntactical relations or node types. Combining the edge-type and node-type translations for fact-driven GAT can accumulate more evidential information for node representations.

D. Informativeness vs. Factual-Correctness (RQ4)

To verify the properties of the generated summaries, we illustrate the performance of the informativeness (RINF.score and Entity.COV) and factual-correctness (FF.score). The results can be found in Fig. 6. The BART and our FFSUM achieve 97.57 and 97.66 on the Entity.COV respectively, which indicates they have an excellent extractive property. In Fig. 6(f), the higher the ROUGE-1 is, the higher the RINF.score is obtained. Besides, all pre-trained frameworks obtain RINF.score higher than 1, indicating that these pre-trained frameworks generate more informative summaries than manual-crafted summaries. However, obtaining a high ROUGE-1 and RINF.score does not mean that the quality of summaries can be guaranteed. For example, in Fig. 6(b), RoBERTa2RoBERTa obtains a higher factual correctness score (60.43), but its RINF.score (1.41) is inferior to RoBERTaSHARE with 1.45 RINF.score. Then, we conclude that pre-trained summarization systems can guarantee informative summaries. However, these systems need to balance multiple evaluation aspects, especially information correctness, since a high ROUGE or a high informativeness performance does not mean a high factual consistency score.

We also study how the λ_{src} (related to copying source text) and λ_{grp} (related to retrieving graph nodes) affect the performance of the proposed method. The results are shown in Fig. 7. From the results, we can see that increasing λ_{grp} does not always result in better performance of ROUGE. The ROUGE performance of FFSUM is better when the two hyper-parameters reach the middle interval of [0.4,0.5,0.6]. We find that the Entity.COV is not very sensitive to these two hyper-parameters because the skeleton model of BART affects capturing rich contextualized representations. However, Fig. 7(d) shows that λ_{grp} has a



(a) ROUGE 1 F1 (b) ROUGE 2 F1 (c) ROUGE L F1 (d) FF.score (e) Entity.COV (f) RINF.score

Fig. 7. The performance comparison of ROUGE, factual-consistency (FF.score) and two informativeness metrics (RINF.score and Entity.COV) under different hyper-parameters of λ_{src} and λ_{grp} to our FFSUM on CNN/DailyMail dataset.

TABLE XI
HUMAN EVALUATION SCORES OF FACTUAL CORRECTNESS BASED ON QUESTIONS ANSWERED BY PARTICIPANTS. THE FINAL SCORE OF A SUMMARIZATION SYSTEM IS THE AVERAGE OF ALL QUESTION SCORES

Model	FF.score(Human Evaluation)	Kappa
RND2RND	69.11	65.04
BERT2BERT	71.07	65.22
RoBERTa2RoBERTa	74.29	65.20
RoBERTaSHARE	72.33	65.00
BART	80.37	70.00
FFSUM (ENGAT)	81.43	70.07

dominant effect on factual-correctness. These two hyper-parameters are sensitive to factual correctness so that both of them can be neither too large nor too small. To get more correct summaries needs to copy both of the original texts and retrieve graph nodes. It is meaningful to improve the model's capability in generating correct factual pieces, even though these models have generated informative summaries.

E. Human Evaluation (RQ4)

We conduct a human evaluation to compare our model-based evaluation method of factual correctness. Firstly, we create a set of questions based on the 20 gold summaries under the assumption that it contains the most (100%) correct factual information from the source text. Taking the generated summary in Fig. 2 as an example, we can design a series of questions based on the summary, such as '*From the generated summary, are you informed that 'the military kills more than 70 Boko Haram members'?*,' or '*From the generated summary, are you informed that 'the kill is in Borno state'?*'. Here, we invite three human annotators (excluding the authors of this paper) who have good knowledge of natural language generation to assign scores to the samples. Three annotators evaluate the factual correctness based on each article and question with a linguistic background. The participants need to select `YES`, `NO`, or `NOT SURE`. The `YES` is 1 for correct ones if the participants can judge the answers according to the summary generated by the system. The `NO` is 0 which refers to that the summary description is not correct. 0.5 refers to `NOT SURE`. A system's score is the average of all question scores. We evaluate the agreement among human annotators by Fleiss' kappa-ratio [49].

As shown in Table XI of the human evaluation of factual consistency, all of Cohen's kappa coefficients are higher than 0.60, indicating a high correlation and agreement among the three human annotators. We find that participants overwhelmingly prefer BART and our FFSUM, whose scores are statistically significant with a p-value smaller than 0.05 under the paired t-test. The score of BERT2BERT is higher than that of RoBERTaSHARE, which can verify that the evaluation model of our FFCC correlates with human preference. The results can also indicate that it is somewhat hard to verify the factual consistency by the model-based evaluation methods. There is still much room to improve to check factual consistency by semantic understanding.

F. Case Study (RQ4)

We conduct qualitative analyses on our model's predicted summaries, including some samples here. In the first case in Fig. 8, all the generations of BART and our FFSUM obtain high ROUGE (80% or higher) and RINF.score (100%). Our FFSUM respects the fine-grained factual pieces with the best ROUGE score of 94.73%. BART and our FFSUM obtain 100% RINF.score, which is the same as the golden summary. RND2RND and BERT2BERT obtain high Entity.COV since the summary is relatively short, making the information more compact. Thus, the summary is very extractive.

However, the baselines make several factual errors: (i), RND2RND generates a wrong 'TEMPORAL' facet of '*by 2050*,' which should be '*by 2070*' and BERT2BERT fabricates a fact in the first case. (ii), these two models also easily falsify the second case's numbers. The generated claim '*Christians will outnumber Christians*,' should be that '*Muslims will outnumber Christians*'. Both of these generated results obtain high Entity.COV, but have incorrect facts. The results indicate more entities can be extracted, but the extracted facts are not always with high correctness.

VIII. CONCLUSION

In this work, we propose a synthetic summarization framework FFSUM to ensure fine-grained factual consistency and factual informativeness. FFSUM is based on salient language generation model BART and is augmented by two sub-modules: a fact-driven graph attention network and a hybrid pointer network. During summary generation, it retrieves and copies fine-grained factual pieces containing facts and facets.

Besides, we reform an existing model-based fact-checking method of FactCC to an ameliorative model FFCC. FFCC is used to verify the factual correctness of phrase-level multi-granular factual-correctness for summarization systems. Experimental

Source Text: Islam will become America's second-largest religion by 2050 according to a report outlining the world's religious landscape 35 years from now. The number of Christians in the US will decline from three quarters of the population in 2010 to just two thirds in 2050, researchers claim. Statistics revealed by the Pew Research Center show the percentage of atheists across the globe is expected to fall across the same time frame while Muslims will outnumber Christians by 2070. Click on the graphic below for a breakdown on each country Research shows that although Islam will be the fastest growing religion in the world over the next four decades, Christians will remain the largest group in 2050. ..... By 2050 in Europe, Muslims will make up one in ten of the overall population. The region's Christian population will shrink from 553million to 454million, according to projections. Islam will be the fastest growing religion in the world over the next four decades but Christians will remain the largest group in 2050, the report reveals. ..... Washington DC's Pew Research Center and Austria's International Institute for Applied Systems Analysis spent six years gathering the data from 2,500 censuses and surveys.

Golden Summary, Entity.COV=78.57, RINF=100
Christians will decline from 75% of US population to just two thirds in 2050. Muslims will outnumber Christians in the world **by 2070**, report predicts. Christians are expected to fall below 50% of the population in UK by 2050. Statistics have been revealed by Washington DC's Pew Research Center.

Generation of RND2RND, ROUGE-1 F1=57.49, Entity.COV=100, RINF.score=64.29
The number of Christians in the United States will decline from three quarters of the population in 2010 to just two thirds in 2050. Muslims will outnumber Christians **by 2050**.

Generation of BERT2BERT, ROUGE-1 F1=76.19, Entity.COV=90, RINF.score=71.43
Christians will decline from 75% of United States population to just two thirds in 2050. **Christians will outnumber Christians** in the world by 2070, report predicts. Statistics of Washington DC's Pew Research Center.

Generation of RoBERTa2RoBERTa: ROUGE-1 F1=78.16, Entity.COV=72.72, RINF.score=78.57
Christians will decline from 75% of United States population to just two thirds in 2050. Christians are expected to fall below 50% of the population in UK by 2050. Statistics of Washington DC's Pew Research Center. Samantha Stosurs beaten by Gisla Dulko.

Generation of RoBERTaSHARE: ROUGE-1 F1=81.39, Entity.COV=72.72, RINF.score=78.57
Christians will decline from 75% of US population to just two thirds in 2050. Christians are expected to fall below 50% of the population in UK by 2050. Statistics of Washington DC's Pew Research Center.

Generation of BART: ROUGE-1 F1=89.13, Entity.COV=78.57, RINF.score=100
Christians will decline from 75% of US population to just two thirds in 2050. Muslims will outnumber Christians in the world by 2070. Christians fall below 50% of the population in UK by 2050. Statistics from Washington DC's Pew Research Center.

Generation of FFSUM: ROUGE-1 F1=94.73, Entity.COV=78.57, RINF.score=100
Christians will decline from 75% of US population to just two thirds in 2050. Muslims will outnumber Christians in the world by 2070. Christians fall below 50% of the population in UK by 2050. Statistics have been revealed by Washington DC's Pew Research Center.

Source Text: As well as moving Aston Villa up one place in the Premier League table on Tuesday night, Christian Benteke's hat-trick against QPR was also a triumph for the top-flight's Belgian contingent. The 24-year-old's treble accounted for the 46th, 47th and 48th goals scored by players from Belgium this season, making them the third most prolific nation in the division. Benteke's goals - which arrived in a memorable 3-3 draw at Villa Park - moved Belgium above France (46 goals) and within 19 of second-placed Spain (67 goals). ..... Diego Costa scores one of the 67 goals scored by players with Spanish nationality against Southampton Manchester City midfielder David Silva has netted 11 top-flight goals so far this season A total of 255 goals have been scored by 81 different English players during the current campaign, with Tottenham's latest hero Harry Kane (19) and QPR hitman Charlie Austin (17) leading the way..... Harry Kane has contributed to England's tally of 255 goals this season with 19 strikes for Tottenham QPR striker Charlie Austin netted the most recent goal scored by an Englishman against Villa on Tuesday Olivier Giroud, pictured celebrating a goal against Liverpool, is the top-scoring Frenchman so far this season Argentina complete the top five on 37 goals, the vast majority of which have arrived from Manchester City's main man Sergio Aguero (17 goals). Senegal (35 goals), Scotland (28) and Ivory Coast (22) come next on the list, with Holland (19), Brazil (18), the Republic of Ireland (18) and Chile (17) - helped by the goalscoring form of Alexis Sanchez - further down. Players from 46 countries have found the back of the net this season, with a total of 800 goals scored so far.....

Golden Summary, Entity.COV=75.00, RINF=100
Aston Villa striker Christian Benteke **scored 47** against QPR on Tuesday. Belgium have now netted 67 Premier League goals this term. A total of 255 goals have been scored by 81 English players during the current campaign.

Generation of RND2RND, ROUGE-1 F1=55.91, Entity.COV=88.89, RINF.score=56.25
Christian Benteke of Aston Villa **scored 46 goals** against QPR on Tuesday. Belgium has scored 67 Premier League goals so far this season. Players from 46 countries have 800 goals.

Generation of BERT2BERT, ROUGE-1 F1=58.91, Entity.COV=83.33, RINF.score=75.00
Christian Benteke, of Aston Villa, **scored 46 goals** against QPR on Tuesday. Belgium has already scored 67 Premier League goals this season. During the current season, 81 English players have scored a total of 255 goals.

Generation of RoBERTa2RoBERTa: ROUGE-1 F1=61.05, Entity.COV=87.5, RINF.score=50.00
Aston Villa striker Christian Benteke scored 47 against QPR on Tuesday. English players have scored 255 goals during the current campaign. Players from 46 countries have scored a total of 800 goals.

Generation of RoBERTaSHARE: ROUGE-1 F1=64.08, Entity.COV=88.89, RINF.score=56.25
Aston Villa striker Christian Benteke scored 47 goals against QPR on Tuesday. A total of 255 goals have been scored by 81 different English players during the current campaign. Players from 46 countries have scored a total of 800 goals.

Generation of BART: ROUGE-1 F1=76.47, Entity.COV=90.90, RINF.score=68.75
Aston Villa striker Christian Benteke scored 47 against QPR on Tuesday. Players from Belgium have now netted 48 Premier League goals this term. English players have scored 255. Players from 46 countries have scored a total of 800 goals.

Generation of FFSUM: ROUGE-1 F1=76.47, Entity.COV=90.90, RINF.score=68.75
Aston Villa striker Christian Benteke scored 47 against QPR on Tuesday. Players from Belgium have now netted 48 Premier League goals this term. English players have scored 255. Players from 46 countries have scored a total of 800 goals.

Fig. 8. Comparison of the output of baselines and our proposed model. Ground truth summary is sampled from the CNN/DailyMail summarization corpus. Factual errors are marked with an underline. **BlueViolet** is the fact, **Olive** is the 'TEMPORAL' facet. **Brown** is the 'NUMERIC' facet.

results have shown that our FFSUM guarantees summaries informativeness and ensures fine-grained factual consistency. Future work includes migrating our fine-grained fact-driven summarization method to other NLG tasks, such as Q&A and dialogue generation.

ACKNOWLEDGMENT

The authors would like to thank the editor and reviewers very much for their positive and constructive comments and suggestions on manuscript.

REFERENCES

[1] H. Peng *et al.*, "Streaming social event detection and evolution discovery in heterogeneous information networks," *ACM Trans. Knowl. Discov. Data*, vol. 15, no. 5, pp. 89:1–89:33, 2021.

[2] A. Esteva *et al.*, "Co-search: COVID-19 information retrieval with semantic search, question answering, and abstractive summarization," 2020, *arXiv:2006.09595*.

[3] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 3104–3112.

[4] Y. Liu and M. Lapata, "Text summarization with pretrained encoders," in *Proc. Conf. Empir. Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 3728–3738.

[5] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," *Tech. Rep., OpenAI*, 2018.

[6] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *Tech. Rep., OpenAI*, vol. 1, no. 8, 2019, Art. no. 9.

[7] L. Dong *et al.*, "Unified language model pre-training for natural language understanding and generation," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 13042–13054.

[8] M. Lewis *et al.*, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 7871–7880.

[9] S. Rothe, S. Narayan, and A. Severyn, "Leveraging pre-trained checkpoints for sequence generation tasks," *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 264–280, 2020.

[10] V. Balachandran, A. Pagnoni, J. Y. Lee, D. Rajagopal, J. G. Carbonell, and Y. Tsvetkov, "StructSum: Summarization via structured representations," in *Proc. 16th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2021, pp. 2575–2585.

[11] Z. Cao, F. Wei, W. Li, and S. Li, "Faithful to the original: Fact aware neural abstractive summarization," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 4784–4791.

[12] L. Huang, L. Wu, and L. Wang, "Knowledge graph-augmented abstractive summarization with semantic-driven cloze reward," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 5094–5107.

[13] C. Zhu *et al.*, "Enhancing factual consistency of abstractive summarization," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2021, pp. 718–733.

[14] B. Gunel, C. Zhu, M. Zeng, and X. Huang, "Mind the facts: Knowledge-boosted coherent abstractive text summarization," 2020, *arXiv:2006.15435*.

[15] R. E. Prasojo, M. Kacimi, and W. Nutt, "StufIE: Semantic tagging of unlabeled facets using fine-grained information extraction," in *Proc. 27th ACM Int. Conf. Inf. Knowl. Manage.*, 2018, pp. 467–476.

[16] K. M. Hermann *et al.*, "Teaching machines to read and comprehend," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 1693–1701.

[17] S. Narayan, S. B. Cohen, and M. Lapata, "Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2018, pp. 1797–1807.

[18] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[19] J. Davison, J. Feldman, and A. M. Rush, "Commonsense knowledge mining from pretrained models," in *Proc. Conf. Empir. Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 1173–1178.

[20] A. Ettinger, "What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models," *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 34–48, 2020.

[21] J. Tan, X. Wan, and J. Xiao, "From neural sentence summarization to headline generation: A coarse-to-fine approach," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2017, pp. 4109–4115.

[22] C. An, M. Zhong, Y. Chen, D. Wang, X. Qiu, and X. Huang, "Enhancing scientific papers summarization with citation graph," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 12498–12506.

[23] F. Liu, J. Flanigan, S. Thomson, N. M. Sadeh, and N. A. Smith, "Toward abstractive summarization using semantic representations," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2015, pp. 1077–1086.

[24] P. Fernandes, M. Allamanis, and M. Brockschmidt, "Structured neural summarization," in *Proc. Int. Conf. Learn. Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=H1ersoRqtm

[25] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Proc. Int. Conf. Learn. Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=rJXMpikCZ

[26] O. Vinyals, M. Fortunato, and N. Jaitly, "Pointer networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 2692–2700.

[27] R. Knowles and P. Koehn, "Context and copying in neural machine translation," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2018, pp. 3034–3041.

[28] C.-S. Wu, R. Socher, and C. Xiong, "Global-to-local memory pointer networks for task-oriented dialogue," in *Proc. Int. Conf. Learn. Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=ryxnHhRqFm

[29] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 1073–1083.

[30] X. Shen, Y. Zhao, H. Su, and D. Klakow, "Improving latent alignment in text summarization by generalizing the pointer generator," in *Proc. Conf. Empir. Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 3760–3771.

[31] J. Gu, Z. Lu, H. Li, and V. O. K. Li, "Incorporating copying mechanism in sequence-to-sequence learning," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2016, pp. 1631–1640.

[32] S. He, C. Liu, K. Liu, and J. Zhao, "Generating natural answers by incorporating copying and retrieving mechanisms in sequence-to-sequence learning," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 199–208.

[33] G. Angeli, M. J. J. Premkumar, and C. D. Manning, "Leveraging linguistic structure for open domain information extraction," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2015, pp. 344–354.

[34] H. Hardy, S. Narayan, and A. Vlachos, "HighRES: Highlight-based reference-less evaluation of summarization," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 3381–3392.

[35] T. Goyal and G. Durrett, "Annotating and modeling fine-grained factuality in summarization," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2021, pp. 1449–1462.

[36] W. Kryściński, B. McCann, C. Xiong, and R. Socher, "Evaluating the factual consistency of abstractive text summarization," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2020, pp. 9332–9346.

[37] T. Falke, L. F. R. Ribeiro, P. A. Utama, I. Dagan, and I. Gurevych, "Ranking generated summaries by correctness: An interesting but challenging application for natural language inference," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 2214–2220.

[38] A. Wang, K. Cho, and M. Lewis, "Asking and answering questions to evaluate the factual consistency of summaries," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 5008–5020.

[39] E. Durmus, H. He, and M. Diab, "FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 5055–5070.

[40] J. Maynez, S. Narayan, B. Bohnet, and R. T. McDonald, "On faithfulness and factuality in abstractive summarization," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 1906–1919.

[41] Y. Dong, S. Wang, Z. Gan, Y. Cheng, J. C. K. Cheung, and J. Liu, "Multi-fact correction in abstractive text summarization," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2020, pp. 9320–9331.

[42] M. Cao, Y. Dong, J. Wu, and J. C. K. Cheung, "Factual error correction for abstractive summarization models," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2020, pp. 6251–6258.
[43] W. Kryscinski, R. Paulus, C. Xiong, and R. Socher, "Improving abstraction in text summarization," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2018, pp. 1808–1817.
[44] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, pp. 4171–4186.
[45] S. Gehrmann, Y. Deng, and A. M. Rush, "Bottom-up abstractive summarization," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2018, pp. 4098–4109.
[46] Y. Liu *et al.*, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.
[47] K. Shu, Y. Li, K. Ding, and H. Liu, "Fact-enhanced synthetic news generation," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 13825–13833.
[48] A. Fan, D. Grangier, and M. Auli, "Controllable abstractive summarization," in *Proc. 2nd Workshop Neural Mach. Transl. Gener.*, 2018, pp. 45–54.
[49] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychol. Bull.*, vol. 76, no. 5, 1971, Art. no. 378.

Qianren Mao is currently working toward the Ph.D. degree with the School of Computer Science and Engineering, Beihang University, Beijing, China. His research interests include natural language generation and text summarization.

Jianxin Li (Member, IEEE) is currently a Professor with the School of Computer Science and Engineering, and the Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing, China. His research interests include the big data, machine learning and trustworthy computing.

Hao Peng is currently an Assistant Professor with the School of Cyber Science and Technology, and Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing, China. His research interests include representation learning, machine learning, and graph mining.

Shizhu He is currently an Associate Researcher with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include question & answering, dialogue system, knowledge reasoning and text mining.

Lihong Wang is currently a Professor with the National Computer Network Emergency Response Technical Team/Coordination Center of China. Her research interests include data mining and analytics, information retrieval, and graph mining.

Philip S. Yu (Life Fellow, IEEE) is currently a Distinguished Professor and the Wexler Chair in Information Technology with the Department of Computer Science, University of Illinois Chicago, Chicago, IL, USA. He is a Fellow of the ACM and IEEE. He has authored or coauthored more than 1,300 referred conference and journal papers cited more than 145,000 times with an H-index of 174. He has applied for more than 300 patents. Dr. Yu was the Editor-in-Chief of *ACM Transactions on Knowledge Discovery from Data* (2011–2017) and IEEE TRANSACTION ON KNOWLEDGE AND DATA ENGINEERING (2001–2004).

Zheng Wang is currently an Associate Professor with the University of Leeds, Leeds, U.K. His research interests include parallel computing, compilation, and systems security.