ECAI 2023
K. Gal et al. (Eds.)

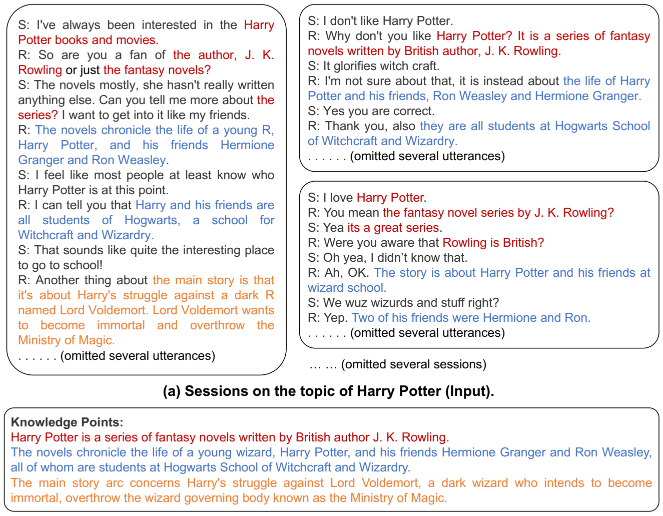
doi:10.3233/FAIA230620

Read Key Points: Dialogue-Grounded Knowledge Points Generation with Multi-Level Salience-Aware Mixture

Sen Zhang[a;†], Baokui Li[a;†], Wangshu Zhang[b], Changlin Yang[b], Yicheng Chen[b], Sen Hu[b], Teng Xu[b] and Jiwei Li [a;*]

[a]Zhejiang University, Hangzhou, China
[b]Ant Group, Hangzhou, China

Abstract. Knowledge-grounded dialogue (KGD) has become increasingly essential for online services, enabling individuals to obtain desired information. While KGD contains knowledge information, most knowledge points are fragmented and repeated in dialogues, making it difficult for users to quickly grasp complete and key information from a collection of sessions. In this paper, we propose a novel task of dialogue-grounded knowledge points generation (DialKPG) to condense a collection of sessions on a topic into succinct and complete knowledge points. To enable empirical study, we create TopicDial and OpenDial corpus based on two existing knowledge-grounded dialogue corpus FaithDial and OpenDialKG by a Three-Stage Annotation Framework, and establish a novel approach for DialKPG task, namely MSAM (Multi-Level Salience-Aware Mixture). MSAM explicitly incorporates salient information at the token-level, utterance-level, and session-level to better guide knowledge points generation. Extensive experiments have verified the effectiveness of our method over competitive baselines. Furthermore, our analysis shows that the proposed model is particularly effective at handling long inputs and multiple sessions due to its strong capability of duplicated elimination and knowledge integration.



S: I've always been interested in the Harry Potter books and movies.
R: So are you a fan of the author, J. K. Rowling or just the fantasy novels?
S: The novels mostly, she hasn't really written anything else. Can you tell me more about the series? I want to get into it like my friends.
R: The novels chronicle the life of a young R, Harry Potter, and his friends Hermione Granger and Ron Weasley.
S: I feel like most people at least know who Harry Potter is at this point.
R: I can tell you that Harry and his friends are all students of Hogwarts, a school for Witchcraft and Wizardry.
S: That sounds like quite the interesting place to go to school!
R: Another thing about the main story is that it's about Harry's struggle against a dark R named Lord Voldemort. Lord Voldemort wants to become immortal and overthrow the Ministry of Magic.
. . . . . . (omitted several utterances)

S: I don't like Harry Potter.
R: Why don't you like Harry Potter? It is a series of fantasy novels written by British author, J. K. Rowling.
S: It glorifies witch craft.
R: I'm not sure about that, it is instead about the life of Harry Potter and his friends, Ron Weasley and Hermione Granger.
S: Yes you are correct.
R: Thank you, also they are all students at Hogwarts School of Witchcraft and Wizardry.
. . . . . . (omitted several utterances)

S: I love Harry Potter.
R: You mean the fantasy novel series by J. K. Rowling?
S: Yea its a great series.
R: Were you aware that Rowling is British?
S: Oh yea, I didn't know that.
R: Ah, OK. The story is about Harry Potter and his friends at wizard school.
S: We wuz wizurds and stuff right?
R: Yep. Two of his friends were Hermione and Ron.
. . . . . . (omitted several utterances)

… … (omitted several sessions)

(a) Sessions on the topic of Harry Potter (Input).

Knowledge Points:
Harry Potter is a series of fantasy novels written by British author J. K. Rowling.
The novels chronicle the life of a young wizard, Harry Potter, and his friends Hermione Granger and Ron Weasley, all of whom are students at Hogwarts School of Witchcraft and Wizardry.
The main story arc concerns Harry's struggle against Lord Voldemort, a dark wizard who intends to become immortal, overthrow the wizard governing body known as the Ministry of Magic.

(b) Knowledge Points for Sessions (Output).

Figure 1: The task overview for DialKPG. The input contains a collection of sessions on a topic, duplicated information can be found across multiple sessions and fragment of the same knowledge can be found in a single session or across multiple sessions. The output is a list of succinct and complete natural language knowledge points contained in the sessions above.

1 Introduction

Knowledge-grounded dialogue (KGD) enables users to ask questions and acquire natural and informative responses quickly. With the increase in online services, KGD has become essential for various purposes, including open-domain dialogues [2, 3], information-seeking conversations [5, 20, 32] and conversational recommender systems [21], where the content of them is based on an entity [20, 21, 32] or a topic [2, 3, 5]. While KGD contains knowledge information, most knowledge points are fragmented and repeated in dialogues due to the chat-based nature, making it difficult for users to quickly grasp complete and key information from a collection of sessions.

The existing task of knowledge point extraction only obtains the connection between entities [8, 27], and the dialogue summarization only summarizes the general content of the dialogue [33, 34]. Neither of the existing works can generate complete knowledge points based on dialogue well. We consider knowledge spread over dialogues can be condensed into a list of more concise texts, which is able to help users quickly read the critical points of dialogues. To this end, we propose a novel task of dialogue-grounded knowledge points generation (DialKPG) to condense a collection of sessions on a topic into a list of succinct and complete natural language knowledge points, as shown in Figure 1.

It is worth mentioning that the DialKPG task has the following two unique challenges. **Duplicated Elimination:** Different dialogues on the same topic involve similar questions, and the answers to these questions refer to repetitive knowledge. The DialKPG task aims to identify these duplicates to generate concise knowledge. This challenge is not trivial as existing related tasks assume that there is a one-to-one correspondence between the utterance and the reference, which might not be true in practice. **Knowledge Integration:** The contents of various parts of a complete knowledge point appear fragmented in the utterances of the same dialogues or different dialogues, and it is necessary to integrate these dialogue utterances to obtain complete knowledge. Those properties make existing related solutions unsuitable for DialKPG task.

To enable empirical study of the DialKPG task, we propose a

[*] Corresponding Author. Email: jiwei_li@zju.edu.cn.
[†] Equal contribution.

three-stage annotation framework to create TopicDial and OpenDial corpus by collecting dialogues and reference knowledge points on a topic from the FaithDial [3] and OpenDialKG [21] corpus. Then, we conduct a comprehensive experiment on created datasets that compares a collection of extractive and abstractive solutions and propose an approach, MSAM (Multi-Level Salience-Aware Mixture), for the DialKPG task that achieves state-of-the-art. Specifically, MSAM is a Transformer-based encoder-decoder model equipped with the Multi-Session Information Integration Module and Multi-Level Salience-Aware Module, which guides knowledge generation by explicitly incorporating salience information of sessions at the token-level, utterance-level, and session-level. Our experiments demonstrate the effectiveness of each module of MSAM and it outperforms other solutions for inputs of different lengths and numbers of sessions.

Our contribution of the paper is threefold:

- We propose a novel task of DialKPG, which takes a collection of sessions on a topic as input and makes a list of succinct and complete natural language knowledge points (Section 3).
- We propose a three-stage annotation framework to create TopicDial and OpenDial corpus by collecting dialogues and reference knowledge points on a topic from the FaithDial [3] and OpenDialKG [21] corpus (Section 4).
- We conduct a comprehensive experiment on TopicDial and OpenDial corpus that compares extractive and abstractive solutions and propose a novel approach, MSAM, for the DialKPG task, which implements duplicated elimination and knowledge integration and achieves state-of-the-art (Section 5 and Section 6).

2 Related Works

Dialogue Summarization. In recent years, several previous research efforts explored dialogue summarization in the context of meetings, customer service conversations, and doctor-patient communication. In the meeting scenario [33], the dialogue summarization serves as the meeting minutes, allowing both attendees and non-attendees to easily review and comprehend the key topics discussed. The dialogue summarization in the customer service scenario [34] needs to capture the core of each topic, which is usually a conversation between the user and customer service to solve one or more problems. The dialogue summarization in the doctor-patient scenario [9] differs from the above scenarios. It is not to obtain an inductive summary but to have deterministic demands. DialKPG task is similar to the dialogue summarization task in obtaining critical information from sessions between more than one user. But the dialogue summarization is the general content of the sessions, which is different from trying to obtain every specific knowledge point.

Multi-Document Summarization. The task of multi-document summarization (MDS) aims to generate a summary that combines dispersed information originally spread across given multiple documents. It enjoys a wide range of real-world applications, including summarization of news [4], Wikipedia articles generation [14], and scientific publications [18]. It is essential to model cross-document relations in MDS [12, 15], which help recognize the redundant and salient content from long documents to guide the summary generation. Although the DialKPG task also extracts vital information from multiple sources, the input contains duplicate content needed to be eliminated and integrated.

Knowledge Extraction. In order to extract knowledge from the text corpus to construct the Knowledge Bases (KBs), knowledge extraction has three subtasks: named entity recognition [11, 29], attribute extraction [6, 31], and relation extraction [8, 27], which can separately extract entities and relations from raw texts and link them to KBs. Although this task extracts the knowledge information from the input, its output is structured information. Different from it, the DialKPG task is to generate complete and concise unstructured text.

3 Task Overview

Let D denote a dataset of dialogues on individual topics $\{t_1, t_2, ..., t_{|D|}\}$, (e.g. movie, bowling). For every topic t, we define a collection of sessions $Dialogue_t = \{d_i\}_{i=1}^{|Dialogue_t|}$, where d_i is a complete knowledge-grounded dialogue, which consists of multi-turn of utterance $d_i = \{u_1^{r1}, u_2^{r2} ... u_n^{rn}\}$, in which r_i represents the speaker role of the i-th utterance $u_i^{r_i}$, and $u_i^{r_i}$ is the sequence of tokens $u_i^{r_i} = (w_1, ..., w_n)$.

In this task, we focus on two-party dialogues. More concretely, there are $r_i \in \{S, R\}$, where S and R represent the roles of speakers: Knowledge Seeker and Knowledge Responder, respectively.

Given a collection of sessions on a topic t, the DialKPG task is to generate succinct and complete natural language knowledge points $y = \{y_1, y_2 ... y_m\}$ with m sequence of words from $Dialogue_t$. The definitions of knowledge point and knowledge are interchangeable. These terms refer to factual and descriptive text that can be obtained from a professional book or a wiki repository.

4 Dataset Construction

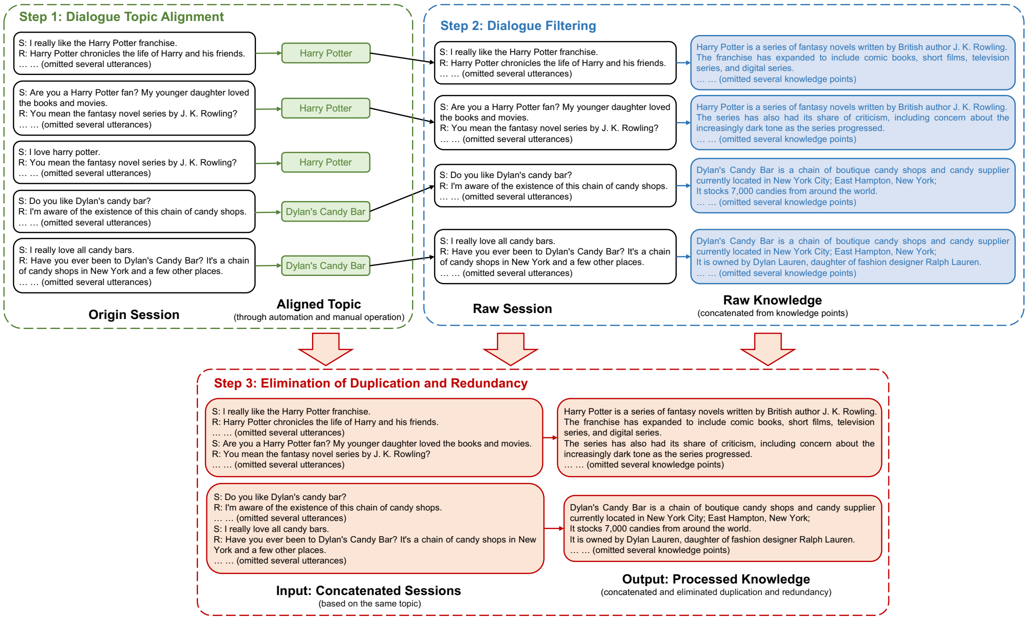
In order to evaluate the DialKPG task, in the following we build datasets TopicDial and OpenDial based on FaithDial [3] and OpenDialKG [21], respectively, two existing knowledge-grounded dialogue datasets, but the task is the opposite. Using the construct process of TopicDial as an example, we first describe the three-stage annotation framework to collect all sessions on the same topic and gold-standard reference knowledge points. And then we build our benchmark datasets TopicDial and OpenDial.

4.1 A Three-Stage Annotation Framework

Construct TopicDial corpus based on knowledge-driven dialogue dataset FaithDial is challenging because the FaithDial corpus does not have topic fields and contains references that are not fully utilized. Consequently, modifications are necessary to satisfy the proposed task. To this end, we design a three-stage annotation framework, where schematic procedure is depicted in Figure 2.

Step 1: Dialogue Topic Alignment. In this step, we attach a topic for each session and its corresponding reference knowledge in the original corpus. Because the FaithDial, a corpus modified from the original Wizard of wikipedia (WoW) [2], drop the topic field, alignment using the WoW corpus is required. Specifically, there are three alignment methods as follows: (1) match utterances from the FaithDial sessions with those from the WoW sessions based on similarity to obtain the topic field of the current session in the WoW corpus; (2) use all topic words of the WoW corpus in turn to match the utterances or references from the FaithDial to determine the topic of the current session; (3) if a session corresponds to multiple topics, the operator will select the most suitable topic. The percentage of aligned sessions based on automation (method 1 and method 2) is 97.33%. In addition, we use manual operation (method 3) to accomplish the rest and sample all aligned results to ensure the quality of alignment.

Step 2: Dialogue Filtering. The raw sessions are formed by reordering the sessions based on the topic obtained in the first step for each session, where sessions on the same topic are grouped. To alleviate the uneven distribution in quantity of sessions across different


Step 1: Dialogue Topic Alignment
Harry Potter
Harry Potter
Harry Potter
Dylan's Candy Bar
Dylan's Candy Bar
Origin Session
Aligned Topic
(through automation and manual operation)
Step 2: Dialogue Filtering
Raw Session
Raw Knowledge
(concatenated from knowledge points)
Step 3: Elimination of Duplication and Redundancy
Input: Concatenated Sessions
(based on the same topic)
Output: Processed Knowledge
(concatenated and eliminated duplication and redundancy)

S: I really like the Harry Potter franchise.
R: Harry Potter chronicles the life of Harry and his friends.
… … (omitted several utterances)

S: Are you a Harry Potter fan? My younger daughter loved the books and movies.
R: You mean the fantasy novel series by J. K. Rowling?
… … (omitted several utterances)

S: I love harry potter.
R: You mean the fantasy novel series by J. K. Rowling?
… … (omitted several utterances)

S: Do you like Dylan's candy bar?
R: I'm aware of the existence of this chain of candy shops.
… … (omitted several utterances)

S: I really love all candy bars.
R: Have you ever been to Dylan's Candy Bar? It's a chain of candy shops in New York and a few other places.
… … (omitted several utterances)

S: I really like the Harry Potter franchise.
R: Harry Potter chronicles the life of Harry and his friends.
… … (omitted several utterances)

Harry Potter is a series of fantasy novels written by British author J. K. Rowling.
The franchise has expanded to include comic books, short films, television series, and digital series.
… … (omitted several knowledge points)

S: Are you a Harry Potter fan? My younger daughter loved the books and movies.
R: You mean the fantasy novel series by J. K. Rowling?
… … (omitted several utterances)

Harry Potter is a series of fantasy novels written by British author J. K. Rowling.
The series has also had its share of criticism, including concern about the increasingly dark tone as the series progressed.
… … (omitted several knowledge points)

S: Do you like Dylan's candy bar?
R: I'm aware of the existence of this chain of candy shops.
… … (omitted several utterances)

Dylan's Candy Bar is a chain of boutique candy shops and candy supplier currently located in New York City; East Hampton, New York;
It stocks 7,000 candies from around the world.
… … (omitted several knowledge points)

S: I really love all candy bars.
R: Have you ever been to Dylan's Candy Bar? It's a chain of candy shops in New York and a few other places.
… … (omitted several utterances)

Dylan's Candy Bar is a chain of boutique candy shops and candy supplier currently located in New York City; East Hampton, New York;
It is owned by Dylan Lauren, daughter of fashion designer Ralph Lauren.
… … (omitted several knowledge points)

S: I really like the Harry Potter franchise.
R: Harry Potter chronicles the life of Harry and his friends.
… … (omitted several utterances)
S: Are you a Harry Potter fan? My younger daughter loved the books and movies.
R: You mean the fantasy novel series by J. K. Rowling?
… … (omitted several utterances)

Harry Potter is a series of fantasy novels written by British author J. K. Rowling.
The franchise has expanded to include comic books, short films, television series, and digital series.
The series has also had its share of criticism, including concern about the increasingly dark tone as the series progressed.
… … (omitted several knowledge points)

S: Do you like Dylan's candy bar?
R: I'm aware of the existence of this chain of candy shops.
… … (omitted several utterances)
S: I really love all candy bars.
R: Have you ever been to Dylan's Candy Bar? It's a chain of candy shops in New York and a few other places.
… … (omitted several utterances)

Dylan's Candy Bar is a chain of boutique candy shops and candy supplier currently located in New York City; East Hampton, New York;
It stocks 7,000 candies from around the world.
It is owned by Dylan Lauren, daughter of fashion designer Ralph Lauren.
… … (omitted several knowledge points)

Figure 2: Overview of our three-stage annotation framework for the DialKPG task.

topics, sessions are restricted to a maximum of 10 per topic. Then, the raw knowledge is obtained by concatenating the knowledge points referenced by each utterance for the filtered sessions.

Step 3: Elimination of Duplication and Redundancy. The filtered sessions and knowledge related to the same topic are concatenated separately. The concatenated knowledge points may contain duplicated information since both the utterances in a single session and multiple sessions on the same topic refer to the same knowledge point. In addition to that, not all content in knowledge points of the FaithDial corpus is completely referenced, making it contain redundant information. To address these issues, we combine similar knowledge points according to semantic similarity (BERT embeddings and word overlap) and remove the unused knowledge span by manual operation to eliminate duplication and redundancy. The validation step ensures that our reference knowledge can be inferred by the enriched input sessions well.

4.2 *Dataset Statistics*

Using the proposed Three-Stage Annotation Framework, we create the TopicDial and OpenDial corpus. Overall, TopicDial comprises a total of 5,649 sessions consisting of 50,761 utterances on 1,633 topics, and OpenDial contains a total of 30,410 sessions consisting of 215,134 utterances on 11,536 topics. Table 1 reports statistics for each dataset split. An input example of the TopicDial corpus and its reference knowledge is given in Figure 1.

5 Methodology

In this paper, a transformer-based encoder-decoder model is employed for the DialKPG task. The model MSAM, shown in Figure 3, encapsulates multi-session information integration and multi-level salience-aware mixture to generate knowledge from a collection of session on a topic. We perform multi-task end-to-end training, only one forward propagation when inference. During training, the model jointly learns to predict the salience score at the token-level, utterance-level, and session-level and fuse this salience information to guide the knowledge generation. During inference, MSAM predicts and fuses the salience with the encoder outputs and uses this fusion salience to guide the decoder to generate knowledge.

Table 1: Dataset statistics of the TopicDial and OpenDial.

Dataset	TopicDial			OpenDial		
	Train	Valid	Test	Train	Valid	Test
Turns	36809	6851	7101	178078	18447	18609
Sessions	4094	764	791	25099	2656	2655
Topics	968	322	343	8075	1730	1731
Avg. Sessions per Topic	4.23	2.37	2.31	3.11	1.54	1.53
Avg. Knowledge points per Topic	9.46	6.23	6.21	8.74	4.43	4.96
Avg. Tokens for Responder	20.29	21.76	20.86	14.24	14.15	14.27
Avg. Tokens for Seeker	17.25	16.65	16.49	12.26	12.25	12.27
Avg. Tokens for Knowledge	15.21	25.13	25.66	6.94	6.93	6.89
Turns per Session	9	9	9	7	7	7

5.1 *Problem Formulation*

Our assumption comes from an intuition that is explicitly leveraging salience helps the model pay more attention to the key content and generate succinct and informative knowledge. The problem can be formulated as follows:

$$P(y|x) = \prod_{k=1}^{|y|} p(y_k|y_{<k}, x, MoE(x)), \tag{1}$$

where x is the sequence of input in the source sessions on a topic, y is the sequence of the corresponding knowledge, and the fused salience allocation $MoE(x)$ will be defined in later. The conditional probability $P(y|x)$ is calculated by RHS, where each token prediction is conditioned on the previously decoded tokens, the input tokens from the source sessions, and the allocation of fused salience of the input.

5.2 *Multi-Session Information Integration*

In order to integrate the multi-session information and predict the salience degrees of input tokens at the token-level, utterance-level, and session-level, we first construct the encoder input sequence $x = Dialogue_t = \{d_i\}_{i=1}^{|Dialogue_t|}$ by adding a special token at the beginning of each session and utterance on a topic: $\hat{d}_i = \langle session\rangle, \langle utterance\rangle, r_1, w_{11}, w_{12}, ..., \langle utterance\rangle, r_2, w_{21},$

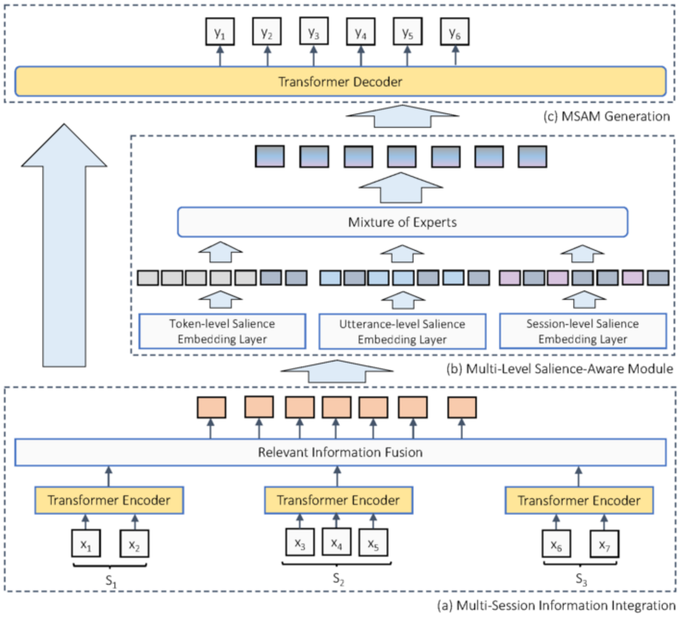

y1
y2
y3
y4
y5
y6
Transformer Decoder
(c) MSAM Generation
Mixture of Experts
Token-level Salience Embedding Layer
Utterance-level Salience Embedding Layer
Session-level Salience Embedding Layer
(b) Multi-Level Salience-Aware Module
Relevant Information Fusion
Transformer Encoder
Transformer Encoder
Transformer Encoder
x1
x2
x3
x4
x5
x6
x7
S1
S2
S3
(a) Multi-Session Information Integration

Figure 3: The overview of our model consists of three main parts: a Multi-Session Information Integration, a Multi-Level Salience-Aware Module, and an MSAM Generation module. Different colored squares in the middle part mean their respective representation.

$w_{22}, ..., \langle /s \rangle$, where r_j and w_{jk} represent the role of speaker and the k-token in j-th utterance defined in Section 3. Then we obtain the last-layer hidden states of each special token and other tokens as session representation, utterance representations, and token representations:

$$[\mathbf{h}_i^s, \mathbf{h}_{i1}^u, \mathbf{h}_{i1}^t, \mathbf{h}_{i2}^t, ..., \mathbf{h}_{im}^u, ..., \mathbf{h}_{in}^t] = Encoder(\hat{d}_i), \tag{2}$$

where $\hat{d}_i$ is the modified input sequence, and $\mathbf{h}_i^s, \mathbf{h}_{ij}^u, \mathbf{h}_{ik}^t$ are the contextualized embeddings of the i-th session, j-th utterance, and k-th token, respectively, with $j \in \{1, ..., M\}$ and $k \in \{1, ..., N\}$.

Then, we concatenate the representation of each session to get the representation of multi-session:

$$\mathbf{H} = [Encoder(\hat{d}_1); ...; Encoder(\hat{d}_n)]. \tag{3}$$

5.3 Multi-Level Salience-Aware Module

To predict the multi-level salience information, we feed the multi-session representation into Token-Level Salience-Aware, Utterance-Level Salience-Aware, and Session-Level Salience-Aware modules, respectively, and the obtained salience information at three levels are fused by the Mixture of Experts module. In the following, we will introduce each module in detail.

5.3.1 Token-Level Salience-Aware

In Token-Level Salience-Aware module, we use z_i^t to represent the salience degree at the token-level of the i-th token in the input sessions, and feed the token representations into a single-layer classification head:

$$P(z_i^t = l^t | x) = Softmax(\mathbf{h}_i^t \mathbf{w}_{l^t}^t), \tag{4}$$

where $\mathbf{w}_{l^t}^t$ is a learnable parameter, and $l^t \in \{0, 1\}$ is the index of the salience degree at the token-level. The ground truth of z_i^t is calculated by Algorithm 1 with a threshold $\lambda = 0.1$.

Next, we map the salience degrees at the token-level to trainable embeddings:

$$f(z_i^t) = \mathbf{Embedding}(z_i^t). \tag{5}$$

Algorithm 1 Calculate the ground truth of z_i^t

```
Input: list of all tokens in session $T^s = [t_1^s, t_2^s, ..., t_n^s]$, and list of reference knowledge $K = [T_1^r, T_2^r, ..., T_M^r]$
Parameter: threshold $\lambda$
Output: list of salience degree of each token in session $L = [l_1, l_2, ..., l_n]$
Let $L$ be an empty list.
for $t\ in\ set(T^s)$ do
    $TF_t = \frac{count(t)\ in\ T^s}{n}$
    $IKF_t = \log \frac{M}{1+\sum_{i=1}^{M} In(t,K_i)}$, $In(t, K_i) = 1$ if $K_i$ contains token $t$ else 0.
end for
for $t\ in\ T^s$ do
    $S_t = TF_t \times IKF_t$
    if $S_t > \lambda$ then
        L.ADD(1)
    else
        L.ADD(0)
    end if
end for
return L
```

During inference, we use the soft estimation that calculates the expectation for the salience embedding to predict the salience degrees:

$$f(z_i^t) = \sum_{l^t=0}^{1} \mathbf{Embedding}(z_i^t = l^t) P(z_i^t = l^t | x). \tag{6}$$

Finally, the salience allocation at the token-level is defined as $\delta(x) = [f(z_1^t), ..., f(z_{|x|}^t)]$.

5.3.2 Utterance-Level Salience-Aware

In Utterance-Level Salience-Aware module, we use z_j^u to represent the salience degree at the utterance-level of the j-th utterance in the input sessions, and feed the utterance representations into a single-layer classification head:

$$P(z_j^u = l^u | x) = Softmax(\mathbf{h}_j^u \mathbf{w}_{l^u}^u), \tag{7}$$

where $\mathbf{w}_{l^u}^u$ is a learnable parameter, and $l^u \in \{0, 1, 2\}$ is the index of the salience degree at the utterance-level. We calculate ROUGE-L F1 score between each utterance and corresponding reference knowledge to represent salience at the utterance-level, and set the best threshold $\mu_1 = 0.025$ and $\mu_2 = 0.060$ for three salience degrees.

Next, we map the salience degrees at the utterance-level to trainable embeddings $g(z^u_j)$ similar to (5) and (6). Finally, we define o^u_i as the utterance index for the i-th token, so that the salience allocation at the utterance-level is defined as $\zeta(x) = [g(z^u_{o^u_1}), ..., g(z^u_{o^u_{|x|}})]$.

5.3.3 Session-Level Salience-Aware

In Session-Level Salience-Aware module, we use z^s_k to represent the salience degree at the session-level of the k-th session in the input sessions, and feed the session representations into a single-layer classification head:

$$P(z^s_k = l^s|x) = Softmax(\mathbf{h}^s_k \mathbf{w}^s_{l^s}), \tag{8}$$

where $\mathbf{w}^s_{l^s}$ is a learnable parameter, and $l^s \in \{0, 1, 2\}$ is the index of the salience degree at the session-level. We calculate ROUGE-L F1 between each session and corresponding reference knowledge to represent salience at the session-level, and set the best threshold $\nu_1 = 0.01$ and $\nu_2 = 0.20$ for three salience degrees.

Next, we map the salience degrees at the session-level to trainable embeddings $h(z^s_k)$ similar to (5) and (6).

Finally, we define o^s_i as the session index for the i-th token, so that the salience allocation at the session-level is defined as $\eta(x) = [h(z^s_{o^s_1}), ..., h(z^s_{o^s_{|x|}})]$.

5.3.4 Mixture of Experts

To fuse the salience information at the three levels, we use a gate [26]. In this work, Token-Level Salience-Aware, Utterance-Level Salience-Aware, and Session-Level Salience-Aware modules are treated as three experts. Inspired by [22], we use the gating function $G = \{\alpha(x), \beta(x), \gamma(x)\}$ to calculate the ratio of information preservation based on a matching heuristics between $\mathbf{H}$ in (3) and the salience allocations $\delta(x), \zeta(x), \eta(x)$, respectively. The gating function G is calculated as follows:

$$\begin{aligned}
\tilde{H_1} &= ReLU([H, \delta(x), H - \delta(x), H \odot \delta(x)]^{L\times 4h}\ W_1^{4h\times h}),\\
\tilde{H_2} &= ReLU([H, \zeta(x), H - \zeta(x), H \odot \zeta(x)]^{L\times 4h}\ W_2^{4h\times h}),\\
\tilde{H_3} &= ReLU([H, \eta(x), H - \eta(x), H \odot \eta(x)]^{L\times 4h}\ W_3^{4h\times h}),\\
G &= Softmax([\tilde{H_1}; \tilde{H_2}; \tilde{H_3}]^{L\times 3h}\ W_G^{3h\times 3}),
\end{aligned} \tag{9}$$

where W_1, W_2, W_3 and W_G are learnable parameters, L is the length of input sequence, h is the size of hidden states, and $\odot$ represents element-wise multiplication.

The output of Mixture of Experts (MoE) is defined as the linear combination of three experts as follows:

$$MoE(x) = \alpha(x) \odot \delta(x) + \beta(x) \odot \zeta(x) + \gamma(x) \odot \eta(x). \tag{10}$$

After incorporating salience information at three levels by (9) and (10), we obtain the final fused representation $MoE(x)$, which can be used for guiding knowledge generation.

5.4 MSAM Generation

In order to use the fused salience information to guide the knowledge generation, we add the fused salience embedding to the encoder hidden state of each token as the key state in the cross-attention layer on the decoder side. The cross-attention is formulated as:

$$CrossAttn(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d_k}})V, \tag{11}$$

where the attention query $Q = \mathbf{h}^{decoder}$ is the hidden state of the decoder, attention key $K = \mathbf{h}^{encoder} + MoE(x)$ is the sum of the hidden state of the encoder and the fused salience embedding, and attention value $V = \mathbf{h}^{encoder}$ is the hidden state of the encoder. Adding salience information to cross-attention scores explicitly, the model can better perceive knowledge-related content at the token-level, utterance-level and session-level to generate the knowledge.

5.5 Learning Objectives

In training, we perform multi-task learning to let MSAM learn to predict the salience allocation at the three levels and generate the knowledge points simultaneously.

For salience prediction, we apply the averaged cross-entropy loss on each predicted token, utterance, and session:

$$\begin{aligned}
\mathcal{L}^t_{cls} &= -\frac{1}{N}\sum_{i=1}^{N} logP(z^t_i|x),\\
\mathcal{L}^u_{cls} &= -\frac{1}{M}\sum_{j=1}^{M} logP(z^u_j|x),\\
\mathcal{L}^s_{cls} &= -\frac{1}{|D_t|}\sum_{k=1}^{|D_t|} logP(z^s_k|x),
\end{aligned} \tag{12}$$

where N, M, and $|D_t|$ are the number of the tokens, utterances, and sessions in input.

For knowledge generation, we feed the ground-truth salience allocation into decoder and use the averaged cross-entropy loss on each predicted token as below:

$$\mathcal{L}_{gen} = -\frac{1}{|y|}\sum_{i=1}^{|y|} logp(y_i|y_{<i}, x, MoE(x)). \tag{13}$$

We combine the above loss functions together with different coefficients ω_g, ω_t, ω_u, and ω_s as shown in:

$$\mathcal{L} = \omega_g\mathcal{L}_{gen} + \omega_t\mathcal{L}^t_{cls} + \omega_u\mathcal{L}^u_{cls} + \omega_s\mathcal{L}^s_{cls}, \tag{14}$$

where the hyperparameter ω_g=1.0, ω_t=0.5, ω_u=1.0, and ω_s=0.5.

6 Experiment

Metrics and Evaluation. Several metrics are needed to provide a multifaceted measure of performance, including ROUGE (ROUGE-1, ROUGE-2 and ROUGE-L) [13], BLEU (BLEU-4) [24], and BARTScore [28], which evaluate the recall rate, accuracy, and faithfulness of the knowledge generation, respectively. All results are the mean values of three runs with different random seeds.

Baselines. We test both extractive and abstractive baselines to examine the feasibility and explore the challenges of DialogKPG task. **Extractive method** extracts salient utterances from input sessions as the output knowledge. In addition to a simple rule-based baseline (Origin and Oracle) that filters the original utterances, we evaluate unsupervised (TextRank) and supervised (BertSum) models to understand how well extracting utterances without rewriting copes with the task. **Origin-All:** Use the original input sessions directly as the output knowledge. **Origin-Reply:** Use the answers in the original input sessions directly as the output knowledge. **Oracle [23]:** Use a greedy algorithm to extract the utterances in sessions that maximize the ROUGE scores to the ground knowledge. **TextRank [19]:** It is an unsupervised extractive method that converts utterances in sessions into graphs and extracts top-ranked utterances by a graph-based ranking algorithm. **BertSum [16]:** It is a supervised extractive method which fine-tunes BERT to extract utterances by solving multiple sentence-level classifications. We also use a series of the pre-

Table 2: Automatic evaluation results on the TopicDial and OpenDial datasets. Methods are categorized into three groups: extractive, abstractive, and our methods. **Bold** represents best result under the Extractive or Abstractive setting and **Bold*** represents best result.

Model		TopicDial					OpenDial				
		R-1	R-2	R-L	BLEU	B-S	R-1	R-2	R-L	BLEU	B-S
Extractive	Origin-All	43.90	15.91	38.62	7.90	-2.63	27.46	11.66	26.98	6.27	**-2.34**
	Origin-Reply	50.78	20.22	43.88	10.77	**-2.60**	29.47	13.75	28.78	8.35	-2.71
	Oracle	**52.44**	**21.69**	**45.08**	**11.67**	-2.64	**54.02**	**31.70**	**52.71**	**16.17**	-3.16
	TextRank	38.42	13.45	33.22	6.99	-3.71	29.57	13.42	28.76	7.9	-2.81
	BertSum	38.02	14.07	32.85	4.26	-3.69	45.29	25.57	44.03	12.83	-3.12
Abstractive	BART	**55.25**	26.19	**49.11**	13.73	-2.65	67.21	**55.84**	67.12	31.70	-1.88
	T5	54.23	25.22	48.17	11.78	-2.67	60.73	47.02	60.62	23.54	-2.32
	Pegasus	54.38	26.41	48.44	13.26	-2.68	64.31	51.01	64.20	26.87	-2.17
	LED	54.83	26.05	49.08	13.77	-2.68	65.91	54.04	67.12	29.16	-1.97
	FiD	54.52	**26.70**	49.03	**15.64**	**-2.62**	**67.54**	55.63	**67.45**	**33.12**	**-1.84**
Ours	MSAM	**56.19***	**28.36***	**50.48***	**17.93***	**-2.58***	**68.30***	**56.58***	**68.17***	**34.98***	**-1.81***

trained language models[1] to explore **abstractive methods** that directly generate the knowledge given the input sessions. **BART [10]:** It is an encoder-decoder Transformer model with denoising seq2seq pre-training. **T5 [25]:** This model is pre-trained on a multi-task mixture of unsupervised and supervised tasks and for which each task is converted into a text-to-text format. **PEGASUS [30]:** It uses gap-sentence generation for pre-training an encoder-decoder Transformer on abstractive summarization tasks. **LED [1]:** To address the limitation that transformer-based models are unable to process long sequences, this model is equipped with an attention mechanism that scales linearly with sequence length. **Fusion-in-Decoder (FiD) [7]:** It is built on top of the pre-trained generative model T5, which processes each session by the encoder independently, and the decoder performs attention over the concatenation of the resulting representations of all the sessions.

Implementation details. Our Implementation is based on Huggingface Transformers library[2], and we choose BART-base backbone. We fine-tune our model using PyTorch distributed data-parallel (DDP) training on 4×A100 GPUs (40GB) with a batch size of 16, resulting in an effective batch size of 64. All models were trained for 50 epochs using AdamW optimizer with learning rate of 3e-5 and a learning rate schedule with 200 warmup steps and linear decay. For inference, we use beam search with beam size of 4, and the minimum and maximum decoding length is 100 and 512 respectively.

6.1 Main Results

Table 2 shows the results on the TopicDial and OpenDial datasets, and we discuss the performance of extractive model and abstractive model based on three metrics respectively.

Extractive: The Origin-All rule method, which directly uses the original input, works well for the TopicDial dataset, indicating that the original utterances contain a large amount of knowledge content. Furthermore, Origin-Reply, which uses the reply in the original input, performs better than Origin-All. This is expected because, in most cases, the knowledge content is in the answers. However, this rule method is not as effective as the Oracle method, indicating that selecting input directly as knowledge results in duplicated knowledge and redundant content. Additionally, two extractive summarization baselines perform poorly. TextRank, the unsupervised method, performs worst, indicating that it is difficult to distinguish utterances containing knowledge. The supervised summarization method, BertSum, performs better than TextRank, but worse than Origin, indicating that while a supervised summarization method can extract more knowledge content, its ability to summarize knowledge is limited. For the OpenDial dataset, BertSum performs better than Origin, which is understandable because OpenDial has less knowledge content than TopicDial. It is obvious that using extractive methods cannot solve the two challenges of the DialKPG task: eliminating duplicates and integrating knowledge.

Abstractive: For both datasets, all five abstractive methods achieve high performance. This indicates that the abstractive approach is more suitable for this task because it generates concise and comprehensive knowledge. Despite BART and T5 being powerful pre-training models that generate fluent knowledge with high ROUGE scores, they lack coherence to the original sessions, reflected in their low BELU score and BARTScore. The FiD model, which incorporates all session information, produces knowledge with comprehensive references and performs well on faithfulness. Furthermore, our proposed model MSAM, which implements duplicated elimination and knowledge integration, achieves the best performance for all the evaluation metrics. Compared to the baselines, MSAM recalls more concise knowledge points and generates content that is more consistent with the original sessions.

Table 3: Results of model w/ or w/o Multi-Session Information Integration (MSII) and Salience-Aware (SA) modules.

MSII	SA	TopicDial		OpenDial	
		R-L	BLEU	R-L	BLEU
w/o	w/o	49.11	13.73	67.12	31.70
	token-level	50.08	15.36	67.93	33.03
	utterance-level	50.01	**15.77**	67.89	**33.11**
	session-level	49.98	15.21	67.74	32.88
	linear-mix	49.83	15.32	67.81	32.89
	MoE	**50.16**	15.71	**68.09**	33.21
w/	w/o	50.12	15.73	68.01	33.12
	token-level	50.28	17.52	68.11	34.63
	utterance-level	50.43	17.90	68.14	34.91
	session-level	50.33	17.39	67.88	34.45
	linear-mix	50.32	17.84	68.08	34.88
	MoE	**50.48***	**17.93***	**68.17***	**34.98***

6.2 Ablation Study

In this section, we conducted the ablation studies on the Multi-Session Information Integration (MSII) and Salience-Aware (SA) modules using the same setting as in the previous section.

Multi-Session Information Integration (MSII). We first investigate the effectiveness of information integration by removing the

[1] They are implemented with TextBox2.0.

[2] https://github.com/huggingface/transformers

Table 4: An example of knowledge generated by BART, FiD, and MSAM. Texts with the same color represent the same piece of knowledge points, and texts with the black color indicate what does not appear in the Gold.

BART	They are usually carniverous, furry, and felid. When kept as pets, they tend to be called house cats or simply cats when there is no need to distinguish them from other felines and felids. The domestic cat, also known as ”Felis silvestris catus”, is an organism with purebred parents of two different breeds, varieties, or populations. . . (omitted several utterances)
FiD	The domestic cat, also known as ”Felis silvestris catus”, is a small, furry, carnivorous mammal, usually meat-eating, often sneezing, or swelling. When kept as a pet, it is often called house cat. Cats can hear high frequency or faint sounds that human ears cannot detect, such as the sounds made by small animals, such as mice. . . . (omitted several utterances)
MSAM	The domestic cat (”Felis silvestris catus”) is a small, furry, carnivorous mammal. When kept as indoor pets, they are often called house cats or simply cats when there is no need to distinguish them from other felines and felids. They can hear high frequency or faint sounds that human ears cannot hear, such as the sounds made by small animals such as mice. . . . (omitted several utterances)
Gold	The domestic cat (”Felis silvestris catus” or ”Felis catus”) is a small, typically furry, carnivorous mammal. They are often called house cats when kept as indoor pets or simply cats when there is no need to distinguish them from other felids and felines. Cats can hear sounds too faint or too high in frequency for human ears, such as those made by mice and other small animals. . . . (omitted several utterances)

MSII module. In this setting, the model takes the concatenated original sessions directly as input but does not encode and integrate each session separately. The results in Table 3 show a drop in performance by 1.01/2.00 and 0.89/1.42 points in terms of R-L/BLEU of TopicDial and OpenDial without the MSII module. This indicates that MSII module is essential for integrating the multi-session input.

Salience-Aware (SA). We then examine the effectiveness of the proposed SA module from six perspectives in Table 3. First, we observe that using different levels of SA improve performance. The utterance-level SA show the most significant improvement (0.90/2.04 and 0.77/1.41 points) when only one level of SA was used. Second, when the multi-level of SA is used, our proposed MoE is more efficient (improve by 1.05/1.98 and 0.97/1.51 points) than simple linear combinations (improve by 0.72/1.59 and 0.69/1.19 points). These results indicate that using a single level of SA is helpful, but using the MoE module for fusion is even more beneficial.

6.3 *Qualitative Analysis*

We present a case study in Table 4 with a representative example to illustrate the advantage of MSAM. In this case, BART tends to generate fragmented knowledge and redundant content that does not appear in Gold and omits part of knowledge. FiD is able to combine the fragmented knowledge into one, but some redundant content still remains. It is worth mentioning that our proposed MSAM performs well, generating complete and concise knowledge, and indicating its suitability for the DialKPG task.

Table 5: Performance of the models in the short/long inputs and single/multiple sessions scenarios on the TopicDial dataset.

Input	Model	R-1	R-2	R-L	BLEU
Short	T5	**55.63**	25.72	49.23	12.55
	FiD	55.46	**26.94**	49.79	**16.14**
	BART	56.15	26.48	**49.97**	13.40
	MSAM	**56.86***	**28.61***	**51.00***	**18.51***
Long	T5	44.17	21.63	40.57	6.27
	FiD	47.75	25.01	43.53	**12.04**
	BART	**49.87**	**25.53**	**45.58**	10.21
	MSAM	**51.43***	**26.54***	**46.75***	**13.77***
Session	**Model**	**R-1**	**R-2**	**R-L**	**BLEU**
Single	T5	55.96	25.96	49.33	12.77
	FiD	55.65	**27.11**	49.82	**15.94**
	BART	**56.51**	26.57	**50.18**	13.57
	MSAM	**57.07***	**28.91***	**51.13***	**18.54***
Multiple	T5	50.34	23.54	45.56	9.57
	FiD	51.99	25.78	47.24	**14.97**
	BART	**52.86**	**25.92**	**47.75**	11.77
	MSAM	**54.23***	**27.12***	**49.02***	**16.58***

6.4 *Performance in Different Scenarios*

The inputs are classified as either short or long based on whether the total length of the inputs is greater than 512 tokens, and as either a single input session or multiple input sessions depending on the number of input sessions. The results in the Table 5 indicate that the MSAM model performs best against various inputs, with the lowest performance penalty when transitioning from shorter to longer inputs and from a single input session to multiple input sessions.

6.5 *Human Evaluation*

In addition to automatic evaluation, we further conducted human evaluation to assess the quality of knowledge generated by the models. We randomly chose 50 samples from the TopicDial and OpenDial test set. Five annotators were presented with the knowledge generated by three models (BART, FiD, and MSAM) and Oracle, and asked to select the best and worst one based on three criteria: informativeness (Inf.), coherence (Coh.), and conciseness (Con.). Then, we computed the performance of the models using the Best-Worst Scaling [17]. Table 6 illustrates that MSAM outperforms other methods in informativeness and coherence due to its ability to eliminate duplicates and integrate knowledge. Conversely, Oracle performs best in coherence but worst in both informativeness and conciseness among all the tested methods. This is expected, as extracting utterances directly from sessions can result in redundant and duplicated content, as well as the omission of knowledge.

Table 6: Best-Worst Scaling on human evaluation.

Model	TopicDial			OpenDial		
	Inf.	Coh.	Con.	Inf.	Coh.	Con.
Oracle	-0.62	**+0.51**	-0.68	-0.64	**+0.53**	-0.63
BART	+0.06	-0.48	+0.17	+0.14	-0.47	+0.12
FiD	+0.18	-0.36	+0.20	+0.15	-0.42	+0.14
MSAM	**+0.38**	+0.33	**+0.31**	**+0.35**	+0.36	**+0.37**

7 Conclusion

In this work, we proposed a novel task of dialogue-grounded knowledge points generation (DialKPG), which generates knowledge points from a collection of sessions on a topic. To conduct empirical study, we developed a three-stage annotation framework and created the TopicDial and OpenDial datasets for the DialKPG task. A novel approach called multi-level salience aware mixture (MSAM) has been proposed for implementing duplicate elimination and knowledge integration. Using the created datasets TopicDial and OpenDial, the effectiveness of MSAM over strong baselines has been verified. Furthermore, the proposed MSAM achieves state-of-the-art.

Ethics Statement

Our solution can aid in condensing a collection of sessions on a topic into concise and complete knowledge points. Our datasets are derived from the publicly available FaithDial and OpenDialKG datasets, which to our knowledge, do not contain any harmful content. If this method is being used to process sensitive data, it is recommended that users adhere to privacy-preserving policies.

Acknowledgements

We would like to thank the referees for their comments, which helped improve this paper considerably. This research was supported by Ant Group through Ant Research Intern Program.

References

[1] Iz Beltagy, Matthew E. Peters, and Arman Cohan, 'Longformer: The long-document transformer', *arXiv:2004.05150*, (2020).

[2] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston, 'Wizard of wikipedia: Knowledge-powered conversational agents', in *International Conference on Learning Representations*, (2019).

[3] Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M. Ponti, and Siva Reddy, 'FaithDial: A Faithful Benchmark for Information-Seeking Dialogue', *Transactions of the Association for Computational Linguistics*, **10**, 1473–1490, (12 2022).

[4] Alexander R. Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R. Radev. Multi-news: a large-scale multi-document summarization dataset and abstractive hierarchical model, 2019.

[5] Song Feng, Hui Wan, Chulaka Gunasekara, Siva Patel, Sachindra Joshi, and Luis Lastras, 'doc2dial: A goal-oriented document-grounded dialogue dataset', in *EMNLP 2020*. Association for Computational Linguistics, (November 2020).

[6] Rayid Ghani, Katharina Probst, Yan Liu, Marko Krema, and Andrew Fano, 'Text mining for product attribute extraction', *SIGKDD Explor. Newsl.*, **8**(1), 41–48, (jun 2006).

[7] Gautier Izacard and Edouard Grave, 'Leveraging passage retrieval with generative models for open domain question answering', in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 874–880, Online, (April 2021). Association for Computational Linguistics.

[8] Zhanming Jie, Jierui Li, and Wei Lu, 'Learning to reason deductively: Math word problem solving as complex relation extraction', in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5944–5955, Dublin, Ireland, (May 2022). Association for Computational Linguistics.

[9] Kundan Krishna, Sopan Khosla, Jeffrey P. Bigham, and Zachary Chase Lipton, 'Generating soap notes from doctor-patient conversations using modular summarization techniques', in *Annual Meeting of the Association for Computational Linguistics*, (2020).

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

[11] Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li, 'Unified named entity recognition as word-word relation classification', in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 10965–10973, (2022).

[12] Wei Li, Xinyan Xiao, Jiachen Liu, Hua Wu, Haifeng Wang, and Junping Du, 'Leveraging graph to improve abstractive multi-document summarization', in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6232–6243, Online, (July 2020). Association for Computational Linguistics.

[13] Chin-Yew Lin, 'ROUGE: A package for automatic evaluation of summaries', in *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, (July 2004). Association for Computational Linguistics.

[14] Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. Generating wikipedia by summarizing long sequences, 2018.

[15] Yang Liu and Mirella Lapata, 'Hierarchical transformers for multi-document summarization', in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5070–5081, Florence, Italy, (July 2019). Association for Computational Linguistics.

[16] Yang Liu and Mirella Lapata. Text summarization with pretrained encoders, 2019.

[17] Jordan J Louviere, Terry N Flynn, and Anthony Alfred John Marley, *Best-worst scaling: Theory, methods and applications*, Cambridge University Press, 2015.

[18] Yao Lu, Yue Dong, and Laurent Charlin, 'Multi-XScience: A large-scale dataset for extreme multi-document summarization of scientific articles', in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8068–8074, Online, (November 2020). Association for Computational Linguistics.

[19] Rada Mihalcea and Paul Tarau, 'TextRank: Bringing order into text', in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 404–411, Barcelona, Spain, (July 2004). Association for Computational Linguistics.

[20] Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M. Khapra, 'Towards exploiting background knowledge for building conversation systems', in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pp. 2322–2332, (2018).

[21] Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba, 'OpenDialKG: Explainable conversational reasoning with attention-based walks over knowledge graphs', in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 845–854, Florence, Italy, (July 2019). Association for Computational Linguistics.

[22] Lili Mou, Rui Men, Ge Li, Yan Xu, Lu Zhang, Rui Yan, and Zhi Jin, 'Natural language inference by tree-based convolution and heuristic matching', in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 130–136, Berlin, Germany, (August 2016). Association for Computational Linguistics.

[23] Ramesh Nallapati, Feifei Zhai, and Bowen Zhou, 'Summarunner: A recurrent neural network based sequence model for extractive summarization of documents', in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, p. 3075–3081, (2017).

[24] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, 'Bleu: a method for automatic evaluation of machine translation', in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, (July 2002). Association for Computational Linguistics.

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

[26] Noam Shazeer, *Azalia Mirhoseini, *Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean, 'Outrageously large neural networks: The sparsely-gated mixture-of-experts layer', in *International Conference on Learning Representations*, (2017).

[27] Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng, 'Document-level relation extraction with adaptive focal loss and knowledge distillation', in *Findings of ACL*, (2022).

[28] Weizhe Yuan, Graham Neubig, and Pengfei Liu, 'BARTScore: Evaluating generated text as text generation', in *Advances in Neural Information Processing Systems*, eds., A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, (2021).

[29] Zheng Yuan, Chuanqi Tan, Songfang Huang, and Fei Huang, 'Fusing heterogeneous factors with triaffine mechanism for nested named entity recognition', in *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 3174–3186, Dublin, Ireland, (May 2022). Association for Computational Linguistics.

[30] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization, 2019.

[31] Guineng Zheng, Subhabrata Mukherjee, Xin Luna Dong, and Feifei Li, 'Opentag: Open attribute value extraction from product profiles', in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining*, KDD '18, p. 1049–1058, New York, NY, USA, (2018). Association for Computing Machinery.

[32] Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black, 'A dataset for document grounded conversations', in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, (2018).

[33] Chenguang Zhu, Ruochen Xu, Michael Zeng, and Xuedong Huang, 'A hierarchical network for abstractive meeting summarization with cross-domain pretraining', *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, (November 2020).

[34] Yicheng Zou, Lujun Zhao, Yangyang Kang, Jun Lin, Minlong Peng, Zhuoren Jiang, Changlong Sun, Qi Zhang, Xuanjing Huang, and Xiaozhong Liu. Topic-oriented spoken dialogue summarization for customer service with saliency-aware topic modeling, 2020.