Dynamic Entity Memory Network for Dialogue Relational Triplet Extraction

Anonymous ACL submission

Abstract

Relational triplet extraction (RTE) is a crucial task in information extraction and has aroused extensive attention. Although advanced studies on RTE have achieved great progress, they are still insufficient for supporting practical applications, such as dialogue system and information retrieval. In this paper, we focus on relational triplet extraction in dialogue scenarios and introduce a new task named dialogue relational triplet extraction (DRTE). Instead of being treated as static texts like sentences or documents, dialogues should be regarded as dynamic ones generated with the progress of conversations. Thus, it imposes three important challenges, including extracting triplets in real-time with incomplete dialogue context, discovering cross-utterance relational triplets, and perceiving the transition of dialogue topics. To tackle these challenges, we propose a Dynamic Entity Memory Network (DEMN). Specifically, the key of our approach is an attentional context encoder and an entity memory network. The attentional context encoder learns dialogue semantics utterance by utterance and dynamically captures salient contexts for each utterance. The entity memory network is devised to store the entities extracted from previous utterances and for cross-utterance triplets extraction. Meanwhile, it also tracks topic transitions in real-time and forgets the semantics of trivial entities. To verify the effectiveness of our model, we manually build three datasets based on KdConv benchmark. Extensive experimental results demonstrate that our model achieves state-of-the-art performances.

1 Introduction

Relational triplet extraction (RTE) task is an important task in natural language processing field, which aims to identify entities and their relations from unstructure text and orginize them in the form of $\langle$subject, relation, object$\rangle$. As a crucial task beneficial to many applications such as automatic knowledge base construction and question answering, it has attracted widespread attention. Existing studies deal with this task with different paradigms, including table filling (Miwa and Sasaki, 2014; Bekoulis et al., 2018), sequence labeling (Zheng et al., 2017; Wei et al., 2020; Liu et al., 2020), sequence generation (Zeng et al., 2018b; Sui et al., 2020; Zeng et al., 2020), and so on.

Dialogue

Utterance 1: Have you seen the film Se7en?
Utterance 2: Yes, I have. It was released in 1995.
...
Utterance 9: Who is the director of this film?
Utterance 10: David finch. Have you heard of him?
Utterance 11: Yes, he is an American. Do you know where he was born?
Utterance 12: He was born in Denver, Colorado, USA.
...
Utterance 17: Do you know what his representative works are?
Utterance 18: There are Fight Club, Gone Girl and Se7en. Have you seen these films?
Utterance 19: I've seen Se7en. Which actor do you like best in this film?
Utterance 20: I like Kevin Spacey. Do you know where he was born?
...

Topic Transition

Se7en → David Fincher → Se7en → Kevin Spacey

Relational Triplets

Subject	Relation	Object
Se7en	Release Date	1995
Se7en	Director	David Fincher
David Fincher	Nationality	USA
David Fincher	Birth Place	Denver, Colorado, USA
David Fincher	Masterpiece	Se7en
...	...	...

Figure 1: An example of Dialogue Relational Triplet Extraction (DRTE) task. The entities related to different topics are marked in red.

Although these studies have achieved great progress, they generally focus on constructing statics knowledge bases by extracting triplets from sentences or documents such as news and Wikipedia articles, while lacking attention to dialogues. To fill this blank, recent studies (Yu et al., 2020; Chen et al., 2020; Xue et al., 2021) explore dialogue relation extraction task and propose graph-based models to deal with it. However, they mainly extract the relations between pre-defined speakers and speaker-related arguments rather than general knowledges such as $\langle$*Se7en*, *Director*, *David Fincher*$\rangle$ shown in

Figure 1. Besides, these models still treat dialogues as flat long texts and neglect the dynamics of them.

To solve the above problems, we introduce a novel task named dialogue relational triplets extraction (DRTE) task, which aims to dynamically discover general knowledge triplets with the progress of dialogues. The dynamic characteristic of dialogue imposes three pivotal challenges for DRTE task. **First**, utterances of each dialogue are generated in real-time, hence posing a key challenge on how to accurately identify entities and relations with incomplete dialogue, especially when partial components of triplets have not yet appeared. **Second**, utterances are usually short and casual, which leads to plenty of cross-utterance triplets. And some triplets even span more than 10 utterances, such as $\langle Se7en, Director, David\ Fincher \rangle$ in Figure 1. Therefore, how to properly pair the long-distance entities and predict their relation type is an important challenge. **Third**, dialogues generally have more complex topic transitions, how to adapt to this unique logical structure is a key challenge for discovering triplets.

Facing the aforementioned challenges, we propose a Dynamic Entity Memory Network (DEMN). Specifically, we first devise an attentional context encoder to learn the semantics of dialogue utterance by utterance. When our model receives a real-time utterance, this mechanism first utilizes Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) to capture its local semantics, and then adopts the attention mechanism to learn its contextual semantics. Furthermore, we utilize utterance-level LSTM to track the latent topic transition, and devise an entity memory network with forgetting gate for discovering the long-distance triplets without disturbances from entities unrelated to the current topic. To verify the effectiveness of our model, we make a comprehensive and comparative analysis on three datasets, and the results demonstrate that our model achieves state-of-the-art performances. In summary, our contributions are three-fold:

- We introduce dialogue relational triplet extraction (DRTE) task, which is valuable and crucial for downstream tasks but remains under-investigated.
- We propose a dynamic entity memory network (DEMN). By devising an attentional context encoder and an entity memory network, our model can effectively adapt to the dynamic characteristic of dialogues and accurately extract cross-utterance triplets.
- We manually build three datasets based on KdConv benchmark. Extensive experiments are conducted to verify that our model achieves state-of-the-art performances.

2 Related Work

Extracting relational triplets from unstructure text is an important task in information extraction. Previous researches can be mainly categorized into two types, including relation extraction and joint entity and relation exraction.

Relation extraction task aims to predict the relation between any two pre-defined entities according to the given text. Early studies (Mintz et al., 2009; Zeng et al., 2014) effort on sentence-level relation extraction and propose various approaches to alleviate noisy data from distant supervision, such as multi-instance learning (Riedel et al., 2010; Zeng et al., 2015), reinforcement learning (Feng et al., 2018; Zeng et al., 2018a; Qin et al., 2018b), and adversarial training (Qin et al., 2018a; Wu et al., 2017). Although these approaches can effectively classify relations, they fail to deal with cross-sentence relations which limits their application scenarios. To solve this problem, recent studies focus on document-level relation extraction (Yao et al., 2019) and dialogue relation extraction (Yu et al., 2020), which aim to predict relations via semantics of multiple sentences. And plenty of graph based methods (Nan et al., 2020; Li et al., 2020; Xue et al., 2021; Chen et al., 2020) are proposed to adequately model interactions between entities and the context. But these methods assum that entities are pre-defined, which suffers from error propagation problem in practice.

To solve this problem, some studies (Gupta et al., 2016; Zheng et al., 2017) are dedicated to identify entities and their relations in a joint manner. Considering complex relation structures, a variety of neural networks are proposed to extract overlapped triplets, including sequence-to-sequence models (Zeng et al., 2018b; Nayak and Ng, 2020; Ye et al., 2021), sequence labeling models Liu et al. (2020); Wei et al. (2020), token pair linking model (Wang et al., 2020), and reinforcement learning models (Takanobu et al., 2019; Xiao et al., 2020).

However, recent studies generally regard sentences, documents, or dialogues as static text, which fail to adapt the dynamic characteristic of dialogues. To handle this issues, this paper introduce dialogue relational triplets extraction (DRTE) task which aims to identify entities and their relations in real-time for constructing dynamic knowledge graph. To achieve this goal, we propose a Dynamic Entity Memory Network.

3 Methodology

3.1 Problem Formulation

Given a dialogue $U = \left\{u_1, u_2, ..., u_{|U|}\right\}$ with $|U|$ utterances, dialogue relational triplet extraction (DRTE) task aims to identify the collection of triplets $T = [s_i, r_i, o_i]_{i=1}^{|T|}$, where s_i, o_i, and r_i represent the subject, object, and relation type of the i-th triplet, respectively. To deal with this task, we need to recognize the collection of entities $E = \left\{e_1, e_2, ..., e_{|E|}\right\}$ from the given dialogue and predict the relation r between any two entities. Note that, each entity is extracted from the dialogue content, and the relation type r is selected from a pre-defined set $\mathcal{R} = \left\{\mathcal{R}_1, \mathcal{R}_2, ..., \mathcal{R}_{|\mathcal{R}|}\right\}$.

3.2 Framework

There are three pivotal challenges of DRTE task should be tackle, including learning the dynamic context semantics in real-time, discovering the cross-utterance triplets, and tracking the topic transition. To solve these issues, we propose a dynamic entity memory network (DEMN) mainly consisting of an attentional context encoding layer, an entity memory network, and a triplet extraction layer. The overall framework of DEMN is illustrated in Figure 2. Considering the dynamic nature of dialogues, we perform the utterance encoding, entity recognition, triplet extraction and entity memory utterance by utterance. Concretely, we first devise an attentional context encoding layer to learn the isolated semantics and context semantics of each utterance. Based on the fusion of these two semantics, we utilize a token-pair binary classifier for entity recognition. After that, we adopt a supervised multi-head attention mechanism to discovering the relations between any two entity and obtain the inter-utterance and intra-utterance triplets. Finally, the entities of current utterance are used to update the entity memory network, while the semantics of current utterance is used to track the topic transition and weaken the trivial entities.

3.3 Attentional Context Encoding

To dynamically capture the semantics of the real-time utterances, we divide the encoding layer into three parts, including isolated semantics encoding, context semantics encoding and semantics fusion.

Given the t-th utterance, we first utilize BERT to encode the isolated semantics without considering the historical dialogue. Formally, we tokenize the utterance with the WordPiece vocabulary (Wu et al., 2016) and obtain the input sequence $u_t = \left\{x_{[\text{CLS}]}, x_{1,t}, x_{2,t}, ..., x_{|u_t|,t}, x_{[\text{SEP}]}\right\}$, where [CLS], [SEP], and $|u_t|$ denotes the beginning token, the end token, and the utterance length, respectively. The initial representation $\mathbf{x}_{i,t}$ of each token, which is fed into BERT, is constructed by summing its word embedding, position embedding and segment embedding. We take the output of the last Transformer block in BERT as the isolated semantics $H_t^s = \left\{\mathbf{h}^s_{[\text{CLS}]}, \mathbf{h}^s_{1,t}, ..., \mathbf{h}^s_{|u_t|,t}, \mathbf{h}^s_{[\text{SEP}]}\right\}$.

Meanwhile, we adopt the scaled dot-product attention mechanism to access the historical information pool and obtain the context semantics. Concretely, given the isolated semantics H_t^s of the t-th utterance and the historical information pool C_t at the t-th step, the context semantics H_t^c can be calculated as follows:

$$H_t^c = \text{softmax}\left(\frac{H_t^s W_s \cdot (C_t W_c)^T}{\sqrt{d_c}}\right) C_t, \quad (1)$$

where $W_s \in \mathbb{R}^{d_h \times d_c}$ and $W_t \in \mathbb{R}^{d_h \times d_c}$ are model parameters, d_h denotes the dimension of BERT, and d_c is the middle dimension of the dot-product attention.

Finally, we fuse the isolated semantics and the context semantics as follows:

$$H_t^f = \tanh\left(H_t^s + H_t^c\right). \quad (2)$$

The final semantics H_t^f is used to update the historical information pool and extract triplets. When encoding the first utterance, the history pool is empty and the final semantics H_1^f is equivalent to the isolated semantics H_1^c. After each utterance encoding, we push the final semantics H_t^f into the historical information pool to update it:

$$C_{t+1} = \left[C_t; H_t^f\right]. \quad (3)$$

3.4 Entity Memory Network

3.4.1 Memory Updating

Based on the semantics of the given utterances, we first identify the entities existing in them and update the entity memory network with these entites.

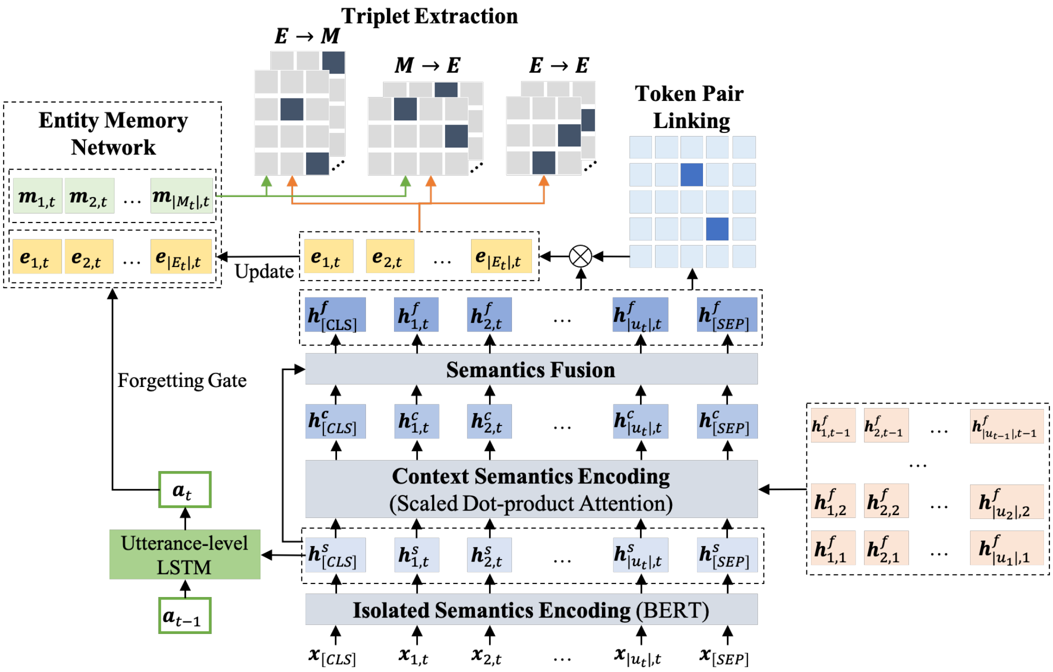


Figure 2: The framework of Dynamic Entity Memory Network (DEMN).

To follow the dynamic nature of dialogue and the principle that entities will not cross utterances, we perform entity recognition utterance by utterance. Furthermore, due to the existence of nested entities, such as 'Denver, Colorado, USA' and 'USA' in Figure 1, we formalize the entity recognition task as a token pair linking task (Wang et al., 2020). Formally, we project the final semantics $H_t^f = \left\{\mathbf{h}_{[\text{CLS}]}^f, \mathbf{h}_{1,t}^f, ..., \mathbf{h}_{|u_t|,t}^f, \mathbf{h}_{[\text{SEP}]}^f\right\}$ to two semantic subspaces, corresponding to the start and end of the entity respectively. And the probability that two tokens indicate the boundary of an entity can be calculated via a binary classifier:

$$\mathbf{s}_{i,t} = \mathbf{h}_{i,t}^f W_s, \quad \mathbf{v}_{i,t} = \mathbf{h}_{i,t}^f W_g, \tag{4}$$

$$\alpha_{i,j,t} = \sigma\left(\mathbf{s}_{i,t} \cdot (\mathbf{v}_{j,t})^T\right), \tag{5}$$

where $\sigma(*)$ represents the sigmoid function, $W_s \in \mathbb{R}^{d_h \times d_e}$ and $W_v \in \mathbb{R}^{d_h \times d_e}$ are model parameters, and d_e represents the middle dimension of the token pair linking.

During training, we aim to maximize the likelihood probability of the gold annotations as follows:

$$p(y_t|u_t) = \prod_{i=1}^{|u_t|}\prod_{j=1}^{|u_t|} p\left(y_{i,j,t}|x_{i,t}, x_{j,t}\right), \tag{6}$$

$$p\left(y_{i,j,t}|x_i, x_j\right) = \begin{cases} \alpha_{i,j,t}, & if \quad y_{i,j,t} = 1 \\ 1 - \alpha_{i,j,t}, & if \quad y_{i,j,t} = 0 \end{cases}, \tag{7}$$

where $y_{i,j,t} = 1$ denotes the fact that the phrase $\{x_{i,t}, ..., x_{j,t}\}$ of the t-th utterance is an entity, while $y_{i,j,t} = 0$ denotes the corresponding phrase is not an entity. During testing, the entity $\{x_{i,t}, ..., x_{j,t}\}$ is extracted if $\alpha_{i,j,t}$ is higher than a given entity threshold γ.

We take the averaged hidden representation between the start and end tokens of each entity as its semantics. And the entity memory is updated via appending the extracted entities of each utterance to it. When the memory slot is full, we discard the entity with the weakest semantics so that new entities can be added.

3.4.2 Memory Forgetting

To track the topic transition and weaken the semantics of trivial entities, we devise a memory forgetting mechanism. Since it is difficult to obtain the direct supervision information of the topic transition, we first utilize an utterance-level LSTM to discover the latent core topic. After that, we design a forgetting gates to attenuate the semantics of entities that are not related to the current topic. Formally, at the t-th step, the semantics $\mathbf{h}_{[\text{CLS}],t}^f$ of the t-th utterance is fed into the utterance-level

LSTM, and the hidden representation $\mathbf{a}_t$ reflecting the current dialogue topic can be distilled as follows:

$$\mathbf{a}_t = \mathrm{LSTM}\left(\mathbf{h}^f_{[\mathrm{CLS}],t}, \mathbf{a}_{t-1}\right). \tag{8}$$

Afterwards, the trivial semantics of the entity memory can be filtered via the forgetting gate:

$$\mathbf{g}_{i,t} = \sigma\left([\mathbf{a}_t; \mathbf{m}_{i,t}]W_g + b_g\right), \tag{9}$$

$$\mathbf{m}_{i,t+1} = \mathbf{g}_{i,t} \odot \mathbf{m}_{i,t}, \tag{10}$$

where $\mathbf{m}_{i,t}$ denotes the hidden representation of the i-th entity memory slot at the t-th step, $\odot$ denotes the element-wise multiplication, $W_g \in \mathbb{R}^{d_h \times d_h}$ and $b_g \in \mathbb{R}^{d_h}$ are model parameters. The forgetting gate $g_{i,t} \in [0,1]^{d_h}$ controls the amount of information flowing from each entity memory slot and updates the entity memory M_t to M_{t+1}.

3.5 Triplet Extraction

To identify triplets accurately and avoid duplicate entity pairing, we design an inter-utterance triplet extraction module and an intra-utterance triplet extraction module.

Formally, given the t-th utterance, we can obtain the collection of entities $E_t = \left\{e_{1,t}, ..., e_{|E_t|,t}\right\}$ extracted from it and the entity memory $M_t = \left\{\mathbf{m}_{1,t}, \mathbf{m}_{2,t}, ..., \mathbf{m}_{|M_t|,t}\right\}$ consisting of the historical entities. The intra-utterance triplet extraction module only predicts the relations between any two entities from E_t, while the inter-utterance triplet extraction module detects the relations between the entity from E_t and the entity from M_t.

For each module, we adopt the supervised multi-head attention mechanism (Liu et al., 2020) to predict the relations. Given two entities e_i and e_j, we project them to different relation subspaces and calculate their correlation intensity under each subspace as follows:

$$\mathbf{q}^l_i = \mathbf{e}_i W^l_q, \quad \mathbf{k}^l_j = \mathbf{e}^l_j W^l_k, \tag{11}$$

$$\beta^l_{i,j} = \sigma\left(\frac{\mathbf{q}^l_i \cdot (\mathbf{k}^l_j)^T}{\sqrt{d_r}}\right), \tag{12}$$

where $\beta^l_{i,j}$ represents the probability that $(e_i, \mathcal{R}_l, e_j)$ is identifies as a triplet, and d_r is the dimension of each subspace. The representation $\mathbf{q}^l_i$ denotes the semantics of e_i as the subject under the relation r_l, while $\mathbf{k}^l_i$ is the semantics of e_j as the object under the same relation.

During training, we separately maximize the likelihood probability of the gold inter-utterance triplets and the gold intra-utterance triplets as follows:

$$p_{E\to E}\left(z_t|E_t\right) = \prod_{i=1}^{|E_t|}\prod_{j=1}^{|E_t|}\prod_{l=1}^{|\mathcal{R}|} p\left(z^l_{i,j,t}|e_{i,t}, e_{j,t}\right), \tag{13}$$

$$p_{E\to M}\left(z_t|E_t, M_t\right) = \prod_{i=1}^{|E_t|}\prod_{j=1}^{|M_t|}\prod_{l=1}^{|\mathcal{R}|} p\left(z^l_{i,j,t}|e_{i,t}, m_{j,t}\right), \tag{14}$$

$$p_{M\to E}\left(z_t|M_t, E_t\right) = \prod_{i=1}^{|E_t|}\prod_{j=1}^{|M_t|}\prod_{l=1}^{|\mathcal{R}|} p\left(z^l_{j,i,t}|m_{j,t}, e_{i,t}\right), \tag{15}$$

$$p\left(z^l_{i,j,t}|*_{i,t}, *_{j,t}\right) = \begin{cases} \beta^l_{i,j,t}, & if \quad z^l_{j,i,t} = 1 \\ 1-\beta^l_{i,j,t}, & if \quad z^l_{j,i,t} = 0 \end{cases}, \tag{16}$$

where $E \to E$ denotes that both the suject and object are from E_t. Meanwhile, $E \to M$ denotes that the subject is from E_t and the object is from M_t, and the meaning of $M \to E$ is opposite to that. During testing, we extract the triplet if the corresponding $\beta^l_{i,j}$ is higher than the given relation threshold λ.

3.6 Joint Learning

To synchronously learn the entity recognition and triplet extraction and make them mutually improve, we combine the binary cross-entropy loss functions of the them to form the entire loss objective of our model:

$$\mathcal{L}(\theta) = \mathcal{L}_E + \mathcal{L}_{E\to E} + \mathcal{L}_{E\to M} + \mathcal{L}_{M\to E}, \tag{17}$$

$$\mathcal{L}_E = -\sum_{t=1}^{|U|}\sum_{i=1}^{|u_t|}\sum_{j=1}^{|u_t|} p\left(y_{i,j,t} = \eta|x_{i,t}, x_{j,t}\right), \tag{18}$$

$$\mathcal{L}_{E\to E} = -\sum_{t=1}^{|U|}\sum_{i=1}^{|E_t|}\sum_{j=1}^{|E_t|}\sum_{l=1}^{|\mathcal{R}|} p\left(z^l_{i,j,t} = \eta|e_{i,t}, e_{j,t}\right), \tag{19}$$

$$\mathcal{L}_{E\to M} = -\sum_{t=1}^{|U|}\sum_{i=1}^{|E_t|}\sum_{j=1}^{|M_t|}\sum_{l=1}^{|\mathcal{R}|} p\left(z^l_{i,j,t} = \eta|e_{i,t}, m_{j,t}\right), \tag{20}$$

$$\mathcal{L}_{M\rightarrow E} = -\sum_{t=1}^{|U|}\sum_{i=1}^{|E_t|}\sum_{j=1}^{|M_t|}\sum_{l=1}^{|\mathcal{R}|} p\left(z_{j,i,t}^{l} = \eta | m_{j,t}, e_{i,t}\right), \tag{21}$$

where $\eta \in [0, 1]$ represents the gold annotation. The optimization problem in Eq. (17) can be solved by using any gradient descent approach. In this paper, we adopt the AdamW (Loshchilov and Hutter, 2017) approach.

4 Experiments

4.1 Datasets

We construct three datasets based on KdConv (Zhou et al., 2020) dataset, which is a Chinese multi-topic knowledge-driven conversation dataset and covers three domains including film, music, and travel. Each sample in KdConv consists of multi-turn dialogue and its corresponding knowledge triplets. Since there are some problems in the original dataset, we reconstruct it and obtain three datasets named 'Film', 'Music', and 'Travel'. Specifically, we first merge the relation types with the same meaning, such as '主要作品' (major works) and '代表作品' (representative works). After that, we correct the triplets whose subject or object could not be extracted from utterances. Finally, we supplement some triplets missing from the original dataset. The statistics of the corrected datasets are shown in Table 1. Particularly, the number of topics in each sample is more than 2 on average, and even up to 13. Meanwhile, 93% of triplets span multiple sentences, which increases the difficulty of dialogue triplet extraction.

Dataset	Film	Music	Travel
# Dialogues	1500	1500	1500
# Triplets	34475	19342	14280
# Intra-Utters Triplets	2217	1529	630
# Inter-Utters Triplets	32258	17813	13650
Avg. # Tokens per Dialogue	474.7	304.6	347.4
Max. # Tokens per Dialogue	901	628	812
Avg. # Utters	24.4	16.6	16.1
Avg. # Entities	20.0	12.8	10.4
Avg. # Topics	2.9	2.4	2.1
Max. # Topics	13	8	4
Avg. Dist	5.0	4.0	4.1
Max. Dist	29	19	14
# Relations	196	230	6

Table 1: Statistics of datasets. '#', 'Avg. #', 'Max. #' and 'Utters' denotes the number, average number, maximum, and utterances, respectively. 'Dist' is the number of utterances between subject and object in triplet. Following Zhou et al. (2020), we treat the distinct subjects as topics.

Hyper-parameters	Datasets			Range of values
	Film	Music	Travel	
d_c	384	384	384	[64, 128, 256, 384, 512]
d_e	256	256	128	[64, 128, 256, 384, 512]
d_l	256	256	256	[64, 128, 256, 384, 512]
d_r	128	64	64	[32, 64, 128, 256]
γ	0.5	0.5	0.8	[0.1, 0.2,..., 0.9]
λ	0.9	0.8	0.9	[0.1, 0.2,..., 0.9]

Table 2: Configurations of DEMN. d_c, d_e, d_l, d_r, γ, and λ denote the dimensions of the scaled dot-product attention, token pair linking, utterance-level LSTM, relation subspace, entity threshold, relation threshold, respectively.

4.2 Experimental Settings

We adopt the BERT-base model (Devlin et al., 2019) with the hidden size of 768 for the isolated semantics encoding. The essential hyperparameters of our model and the range of values tried per hyperparameterthe are listed in Table 2. We select these hyperparameters according to the F1-score of triplet extraction on the development sets. During training, we use AdamW for optimization with the weight decay of 0.01 and the warmup rate of 0.1. The fine-tuning rate for BERT and the learning rate for training other parameters are set to 1e-5 and 5e-4, respectively. Meanwhile, we set the number of epochs, batch size, and dropout rate to 100, 8, and 0.2, respectively. Based on the above setting, we run our model on a RTX 3090Ti GPU.

4.3 Evaluation

We utilize precision, recall, and F1-score to evaluate the performances on dialogue relational triplet extraction. Specifically, a predicted triplet is correct only if the relation type and the whole spans of two entities are all the same as the golden annotation. For reproducibility, we report the average and standard deviation of testing results over 5 runs with different random seeds. At each run, we select the testing result corresponding to the best performance on development set.

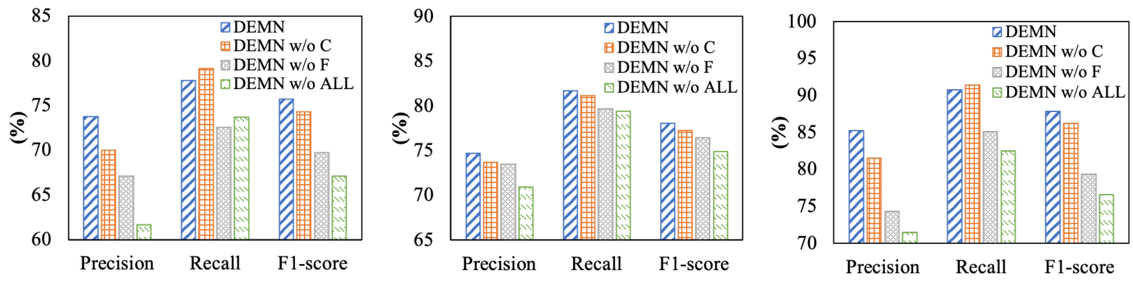
4.4 Comparison Methods

To comprehensively analyze the advantages of our model, we compare it with joint methods and pipeline methods. It is worth noting that each comparison method adopts BERT as the encoder for ensuring the fairness.

First, three advanced joint entity and relation extraction models are selected as the comparison methods. We concatenate all the utterances as a long sequence and feed it into these methods.

Methods	Film			Music			Travel		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
CasRel	71.23 ±1.41	73.56 ±0.61	72.36 ±0.44	74.91 ±2.29	69.48 ±1.52	72.06 ±0.24	64.73 ±4.55	56.65 ±0.89	60.34 ±1.48
TPLinker	68.25 ±1.02	70.82 ±0.29	69.51 ±0.67	66.89 ±0.93	69.86 ±0.87	68.33 ±0.07	81.46 ±0.63	**91.27** ±1.44	86.08 ±0.29
SPN	69.63 ±0.51	73.27 ±1.06	71.40 ±0.77	**77.21** ±1.18	69.25 ±0.34	73.01 ±0.34	85.02 ±1.35	87.51 ±0.30	86.23 ±0.87
TPBC + ATLOP	71.13 ±1.31	72.74 ±1.55	71.91 ±0.09	70.31 ±1.54	54.04 ±0.54	61.10 ±0.23	84.41 ±1.94	68.62 ±1.99	75.66 ±0.43
TPBC + SSAN	**76.05** ±0.37	61.12 ±0.08	67.77 ±0.18	66.70 ±2.48	76.23 ±2.31	71.09 ±0.41	**85.76** ±0.38	67.13 ±0.27	75.31 ±0.03
DEMN (ours)	73.75 ±0.24	**77.79** ±0.23	**75.72** ±0.23	74.72 ±0.65	**81.65** ±0.41	**78.03** ±0.54	85.20 ±0.76	90.72 ±0.82	**87.87** ±0.02

Table 3: Experimental results.



(a) Film (b) Music (c) Travel

Figure 3: Results on ablation study.

- CasRel (Wei et al., 2020) first identfies subjects from text, and then devises multiple relation-specific taggers to extract objects for each subject under each relation type.

- TPLinker (Wang et al., 2020) formulates triplet extraction task as a token pair linking problem. For each possible token pairs, this model utilizes a handshaking tagging scheme to predict whether they indicate the boundary of an entity or the association between subject and object.

- SPN (Sui et al., 2020) transforms triplet extraction task into a set prediction problem and proposed a non-autoregressive decoder with bipartite matching loss function to generate all triplets.

Besides, we also select two document-level relation extraction models for conducting pipeline methods. In the first stage, we utilize the token-pair binary classifier (TPBC) of our model to obtain the collection of entities. In the second stage, we adopt relation extraction models to predict the relation between any two entites. The details are described as follows:

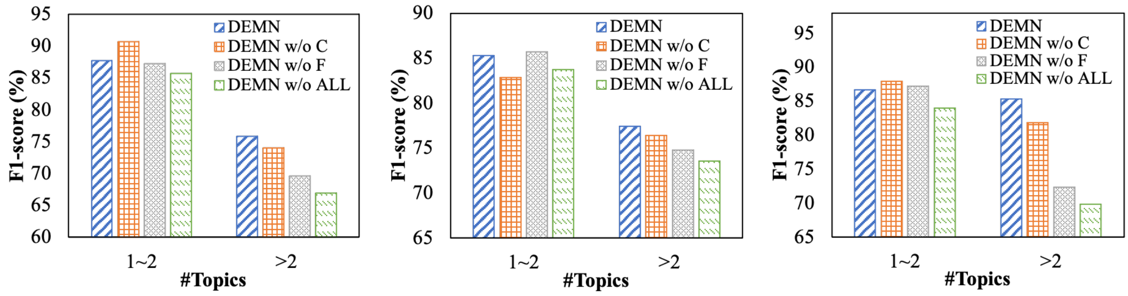
- ATLOP (Zhou et al., 2021) is a document-level relation extraction model. It designs a localized context pooling technique which utilizes the pre-trained attention to discover relevant context for entity pairs.

- SSAN (Xu et al., 2021) utilizes an extension of self-attention mechanism to model co-occurrence and coreference entity structure exhibited in document-level texts.

4.5 Results

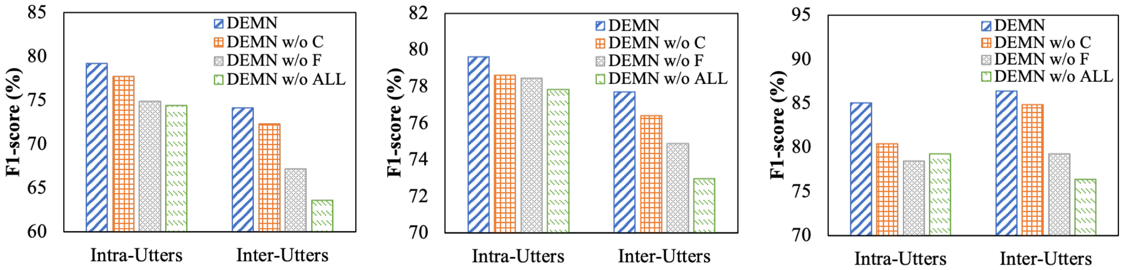
The results on dialogue relational triplet extraction are shown in Table 3. According to the results, our model consistently obtains state-of-the-art performances on three datasets. Compared with the best baseline model, our model outperforms CaseRel by 3.36% F1-score on Film dataset and is higher than SPN by 5.01% and 2.64% F1-score on Music and Travel datasets, respectively. Besides, the joint extraction models achieves better performance than pipeline models, which verifiess that joint learning can make the entity extraction and relation classification mutual promotion.

Specially, all the comparison methods concatenate utterances into a long text and feed it into BERT for encoding, which can fully exploit the semantics of previous utterances and subsequent utterances for entity extraction and relation detection. However, they does not consider the effect of topic transition between utterances, which leads to mismatch and omission of triplets and harms their performances. Conversely, although our model can only use historical information to encodes semantics utterance by utterance, the entity memory network of our model can effectively track dialogue topics and accurately discover triplets. It is



(a) Film (b) Music (c) Film

Figure 4: Results on different dialogue types according to the number of topics.



(a) Film (b) Music (c) Travel

Figure 5: Results on intra-utterance and inter-utterance triplet extraction.

worth noting that our model can flexibly adapt the dynamic characteristic of dialogues and process utterances generated in real-time.

4.6 Ablation Study

To further investigate the origination of the improvement of DEMN, we conduct three ablation experiments, including 'DEMN w/o C', 'DEMN w/o F', and 'DEMN w/o ALL'. Specifically, 'DEMN w/o C' does not use context semantics captured by dot-product attention mechanism, while 'DEMN w/o F' discards utterance-level topic tracking mechanism and forgetting gate. And 'DEMN w/o ALL' abandons these two parts.

According to Figure 3-5, we can analyze the ablation results from three perspectives. First, we display the overall performances in Figure 3. Compared with 'DEMN w/o C', our model achieves significant improvements on precision, which verify the importance of context semantics in reducing mismatch. Besides, First,

5 Conclusion

In this paper, we introduced a novel task named dialogue relational triplet extraction (DRTE) and proposed a dynamic entity memory network (DEMN). To adapt the dynamic characteristic of dialogue, we mainly devised an attentional context encoder and an entity memory network. Specifically, the attentional context encoder learn the semantics of the given dialogue utterance by utterance, which can flexibly and efficiently understand the utterances generated in real time. Furthermore, the entity memory network with a forgetting gate mechanism maintains the entities extracted from previous utterances for discovering the long-distance triplets without disturbances from entities unrelated to the current topic. To verify the effectiveness of our model, we constructed three datasets. Extensive experiments show that our model achieves state-of-the-art performances.

References

Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2018. Joint entity recognition and relation extraction as a multi-head selection problem. *Expert Syst. Appl.*, 114:34–45.

Hui Chen, Pengfei Hong, Wei Han, Navonil Majumder, and Soujanya Poria. 2020. Dialogue relation extraction with document-level heterogeneous graph attention networks. *CoRR*, abs/2009.05092.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT 2019*, pages 4171–4186.

Jun Feng, Minlie Huang, Li Zhao, Yang Yang, and Xiaoyan Zhu. 2018. Reinforcement learning for relation classification from noisy data. In *AAAI 2018*, pages 5779–5786.

Pankaj Gupta, Hinrich Schütze, and Bernt Andrassy. 2016. Table filling multi-task recurrent neural network for joint entity and relation extraction. In *COLING 2016*, pages 2537–2547.

Bo Li, Wei Ye, Zhonghao Sheng, Rui Xie, Xiangyu Xi, and Shikun Zhang. 2020. Graph enhanced dual attention network for document-level relation extraction. In *COLING 2020*, pages 1551–1560.

Jie Liu, Shaowei Chen, Bingquan Wang, Jiaxin Zhang, Na Li, and Tong Xu. 2020. Attention as relation: Learning supervised multi-head self-attention for relation extraction. In *IJCAI 2020*, pages 3787–3793.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *ACL 2009*, pages 1003–1011.

Makoto Miwa and Yutaka Sasaki. 2014. Modeling joint entity and relation extraction with table representation. In *EMNLP 2014*, pages 1858–1869.

Guoshun Nan, Zhijiang Guo, Ivan Sekulic, and Wei Lu. 2020. Reasoning with latent structure refinement for document-level relation extraction. In *ACL 2020*, pages 1546–1557.

Tapas Nayak and Hwee Tou Ng. 2020. Effective modeling of encoder-decoder architecture for joint entity and relation extraction. In *AAAI 2020*, pages 8528–8535.

Pengda Qin, Weiran Xu, and William Yang Wang. 2018a. DSGAN: generative adversarial training for distant supervision relation extraction. In *ACL 2018*, pages 496–505.

Pengda Qin, Weiran Xu, and William Yang Wang. 2018b. Robust distant supervision relation extraction via deep reinforcement learning. In *ACL 2018*, pages 2137–2147.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *ECML 2010*, volume 6323 of *Lecture Notes in Computer Science*, pages 148–163.

Dianbo Sui, Yubo Chen, Kang Liu, Jun Zhao, Xiangrong Zeng, and Shengping Liu. 2020. Joint entity and relation extraction with set prediction networks. *CoRR*, abs/2011.01675.

Ryuichi Takanobu, Tianyang Zhang, Jiexi Liu, and Minlie Huang. 2019. A hierarchical framework for relation extraction with reinforcement learning. In *AAAI 2019*, pages 7072–7079.

Yucheng Wang, Bowen Yu, Yueyang Zhang, Tingwen Liu, Hongsong Zhu, and Limin Sun. 2020. Tplinker: Single-stage joint extraction of entities and relations through token pair linking. In *COLING 2020*, pages 1572–1582.

Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, and Yi Chang. 2020. A novel cascade binary tagging framework for relational triple extraction. In *ACL 2020*, pages 1476–1488.

Yi Wu, David Bamman, and Stuart J. Russell. 2017. Adversarial training for relation extraction. In *EMNLP 2017*, pages 1778–1783.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Ya Xiao, Chengxiang Tan, Zhijie Fan, Qian Xu, and Wenye Zhu. 2020. Joint entity and relation extraction with a hybrid transformer and reinforcement learning based model. In *AAAI 2020*, pages 9314–9321.

Benfeng Xu, Quan Wang, Yajuan Lyu, Yong Zhu, and Zhendong Mao. 2021. Entity structure within and throughout: Modeling mention dependencies for document-level relation extraction. In *AAAI 2021*, pages 14149–14157.

Fuzhao Xue, Aixin Sun, Hao Zhang, and Eng Siong Chng. 2021. Gdpnet: Refining latent multi-view graph for relation extraction. In *AAAI 2021*, pages 14194–14202.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. Docred: A large-scale document-level relation extraction dataset. In *ACL 2019*, pages 764–777.

Hongbin Ye, Ningyu Zhang, Shumin Deng, Mosha Chen, Chuanqi Tan, Fei Huang, and Huajun Chen. 2021. Contrastive triple extraction with generative transformer. In *AAAI 2021*, pages 14257–14265.

Dian Yu, Kai Sun, Claire Cardie, and Dong Yu. 2020. Dialogue-based relation extraction. In *ACL 2020*, pages 4927–4940.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *EMNLP 2015*, pages 1753–1762.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *COLING 2014*, pages 2335–2344.

Daojian Zeng, Haoran Zhang, and Qianying Liu. 2020. Copymtl: Copy mechanism for joint extraction of entities and relations with multi-task learning. In *AAAI 2020*, pages 9507–9514.

Xiangrong Zeng, Shizhu He, Kang Liu, and Jun Zhao. 2018a. Large scaled relation extraction with reinforcement learning. In *AAAI 2018*, pages 5658–5665.

Xiangrong Zeng, Daojian Zeng, Shizhu He, Kang Liu, and Jun Zhao. 2018b. Extracting relational facts by an end-to-end neural model with copy mechanism. In *ACL 2018*, pages 506–514.

Suncong Zheng, Feng Wang, Hongyun Bao, Yuexing Hao, Peng Zhou, and Bo Xu. 2017. Joint extraction of entities and relations based on a novel tagging scheme. In *ACL 2017*, pages 1227–1236.

Hao Zhou, Chujie Zheng, Kaili Huang, Minlie Huang, and Xiaoyan Zhu. 2020. Kdconv: A chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation. In *ACL 2020*, pages 7098–7108.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-level relation extraction with adaptive thresholding and localized context pooling. In *AAAI 2021*, pages 14612–14620.