EARL: Informative Knowledge-Grounded Conversation Generation with Entity-Agnostic Representation Learning

Hao Zhou[1,2], Minlie Huang[1]*, Yong Liu[3], Wei Chen[3], Xiaoyan Zhu[1]
[1]The CoAI group, DCST, State Key Lab of Intelligent Technology and Systems,
[1]Beijing National Research Center for Information Science and Technology,
[1]Tsinghua University, Beijing 100084, China
[2]Pattern Recognition Center, WeChat AI, Tencent Inc., China
[3]Sogou Inc., Beijing 100084, China
`tuxchow@gmail.com, aihuang@tsinghua.edu.cn`

Abstract

Generating informative and appropriate responses is challenging but important for building human-like dialogue systems. Although various knowledge-grounded conversation models have been proposed, these models have limitations in utilizing knowledge that infrequently occurs in the training data, not to mention integrating unseen knowledge into conversation generation. In this paper, we propose an Entity-Agnostic Representation Learning (EARL) method to introduce knowledge graphs to informative conversation generation. Unlike traditional approaches that parameterize the specific representation for each entity, EARL utilizes the context of conversations and the relational structure of knowledge graphs to learn the category representation for entities, which is generalized to incorporating unseen entities in knowledge graphs into conversation generation. Automatic and manual evaluations demonstrate that our model can generate more informative, coherent, and natural responses than baseline models.

1 Introduction

Generating informative and appropriate responses is vital for the success of human-like dialogue systems. To this end, there has been a rising tendency in enhancing conversation models with external knowledge recently, which is well-known as the knowledge-grounded conversation model (Ghazvininejad et al., 2018; Zhou et al., 2018; Dinan et al., 2019). Several studies incorporate unstructured texts, such as web pages (Ghazvininejad et al., 2018) and Wikipedia articles (Dinan et al., 2019), as the external knowledge to generate informative responses. Some work introduces structured knowledge, *e.g.* the knowledge graph (Zhou et al., 2018) to generate knowledge enhanced conversations.

KG	**# Entities**	**# Triples**	**# Relations**
Freebase	40M	637M	35,000
Wikidata	18M	66M	1,632
ConceptNet	8M	21M	36

Table 1: Statistics of some widely used knowledge graphs (KG, Knowledge Graph; M, million).

Prior studies adopt either pre-trained knowledge graph embeddings (Zhou et al., 2018), *e.g.* TransE (Bordes et al., 2013), word embeddings (Wu et al., 2019), or adjacency matrix (Tuan et al., 2019) to model entities and relations in knowledge graphs and incorporate them to conversation generation. These models face two major challenges when applied to introduce large-scale knowledge graphs. **First**, there is a significant gap in representations between knowledge and text (Buitelaar and Cimiano, 2008; Zhou et al., 2018), which requires model training to apply knowledge in conversation generation based on different knowledge representations. However, the training corpus of knowledge-grounded conversations only contains a small subset of entities for applying knowledge, while the large-scale untrained entities are difficult to be utilized due to the gap between their representations. **Second**, it is extremely challenging to represent millions of entities and triples of large-scale knowledge graphs (see Table 1) by these methods in practice, for instance, the adjacency matrix requires $|\mathcal{V}| \times |\mathcal{L}| \times |\mathcal{V}|$ computational resources ($\mathcal{V}$, $\mathcal{L}$ denote the set of entities and relations, respectively).

To address these issues, we propose EARL, an Entity-Agnostic Representation Learning method to incorporate knowledge graphs into informative conversation generation, which can be easily integrated into existing conversation frameworks, such as Seq2Seq (Sutskever et al., 2014), HRED (Serban et al., 2015), and Transformer (Vaswani et al., 2017). The **intuition** lies in that knowledge graphs have sparse entities but dense relations, *e.g.* ConceptNet (Speer et al., 2017) contains over 8 million

* Corresponding author: Minlie Huang.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 2383–2395
November 7–11, 2021. ©2021 Association for Computational Linguistics

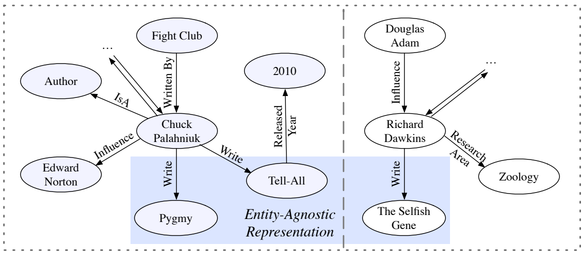
entities while only 36 relations, as shown in Table 1. EARL learns entity-agnostic representations for nodes in the knowledge graph (see Section 3.2 for more details) based on the context information of the conversation and the structure information of the knowledge graph, which does not parameterize the specific representation for each entity like prior methods. Thus it alleviates the problems mentioned above and is more suitable for applying large-scale knowledge graphs.

Specifically, EARL addresses the issues mentioned above in three ways: (1) A delexicalization step replaces the entities in the conversation history with mask tokens, which makes it entity-agnostic to the conversation context thus generalized for unseen entities. (2) A knowledge interpreter is proposed to model the generalized representation of an entity by the structure information of knowledge graphs and the context information of the conversation, which allows our method to generate informative responses with the unseen knowledge graph during inference. (3) EARL learns the relation embeddings for conversation generation while it does not need to store representations for millions of entities, making it scalable to apply large-scale knowledge graphs. Figure 1 shows conversation samples generated by EARL and the prior knowledge-grounded baseline, where EARL (the last line) can inject the unseen entities (the white nodes) coherently. In contrast, the prior baseline model wrongly utilizes the seen knowledge graph (the grey nodes) in conversation generation given the unseen entities as input.

To summarize, our contributions are as follows:

- This work is the first attempt to utilize knowledge graphs without parameterizing specific entity representations in conversation generation, which can be easily integrated to existing conversation frameworks.

- Automatic and manual evaluations show that EARL can generate informative responses with both seen knowledge graphs and unseen knowledge graphs in two benchmark datasets. Ablation studies demonstrate the influence of different mechanisms and conversation frameworks.



Context	**Do you know any good books by Chuck Palahniuk ?**
Baseline	Yes , he wrote Pygmy .
EARL	Chuck Palahniuk wrote Tell-All .
Context	**Do you know any good books by Richard Dawkins ?**
Baseline	Yes , he wrote the book Pygmy. Have you read that ?
EARL	Richard Dawkins wrote The Selfish Gene.

Figure 1: Conversation samples generated with well-trained knowledge graphs (left) and unseen knowledge graphs (right). Grey nodes are well-trained entities, and white nodes are unseen nodes in the training data. Entities in the blue rectangle share the same entity-agnostic representation (see Section 3.2 for more details).

2 Related Work

2.1 Open-domain Conversation Models

Recently, Sequence-to-Sequence (Seq2Seq) models (Sutskever et al., 2014; Bahdanau et al., 2014) have been applied to large-scale open-domain conversation generation, including neural responding machine (Shang et al., 2015), hierarchical recurrent models (Serban et al., 2015), and many others (Sordoni et al., 2015; Li et al., 2016; Shao et al., 2017). Some models are proposed to improve the content quality of generated responses by copy mechanisms, diversified beam search algorithms, and various techniques (Shao et al., 2017; Li et al., 2016; Mou et al., 2016; Gu et al., 2016). However, the lack of background information or related knowledge results in significantly degenerated conversations, where the text is bland and strangely repetitive (Holtzman et al., 2020). Other studies, aiming to generate informative responses, incorporate external knowledge into conversation generation, including unstructured texts (Ghazvininejad et al., 2018; Long et al., 2017), and structured knowledge graphs (Han et al., 2015; Xu et al., 2017; Zhou et al., 2018).

2.2 Knowledge Graph Enhanced Conversation Models

Some prior works introduce high-quality structured knowledge graph for conversation generation. Zhu et al. (2017) presented an end-to-end knowledge grounded conversation model using a copy network (Gu et al., 2016). A large-scale commonsense knowledge graph is introduced to open-domain conversation generation by graph attention mechanisms in (Zhou et al., 2018). Moon et al. (2019) proposed a knowledge graph walker to select relevant entities of the knowledge graph to improve the performance of retrieval-based conversation models. The adjacency matrix (Tuan et al., 2019) is introduced to modeling the dynamic knowledge graph in conversation generation. However, these studies adopt pre-trained knowledge graph embeddings (Zhou et al., 2018), word embeddings (Wu et al., 2019), or the adjacency matrix (Tuan et al., 2019) to represent knowledge triples, making them not applicable for large-scale and unseen knowledge graphs. By contrast, our model addresses this issue by representing knowledge entities based on the context and the structure information of the knowledge graph, making our model entity-agnostic and able to incorporate large-scale and unseen knowledge graphs into conversation generation.

3 Model

3.1 Task Definition

Our problem is formulated as follows: Given a context $X = (x_1, x_2, \cdots, x_n)$, which is the word sequence of a conversation history $H = (U_1, U_2, \cdots, U_c)$, and knowledge graphs $G = \{g_1, g_2, \cdots, g_{|G|}\}$, the goal is to generate the response $Y = (y_1, y_2, \cdots, y_m)$ by estimating the probability: $P(Y|X,G) = \prod_{t=1}^{m} P(y_t|y_{<t}, X, G)$. The graphs are retrieved from a knowledge base using the words in the context as queries. As Zhou et al. (2018), each graph contains one-hop triples as $g_i = \{\tau_1^i, \tau_2^i, \cdots, \tau_{|g_i|}^i\}$, and each triple (subject, relation, object) is represented as $\tau_j^i = (subj^i, rel_j^i, obj_j^i)$.

3.2 Entity-Agnostic Representation Learning

EARL consists of three modules: an encoder to convert the context to the hidden representations, a knowledge interpreter to represent each subject and object entity based on the context and structure information, and a decoder to generate a token or select an entity from the knowledge graph determined by a knowledge selector. The overview of EARL is presented in Figure 3.

Instead of parameterizing specific representations for entities of knowledge graphs as used in prior studies (Zhou et al., 2018; Wu et al., 2019; Tuan et al., 2019), EARL learns entity-agnostic representations conditioning on the context information of the conversation and the structure information of the knowledge graph. Entity-agnostic representations are defined as category representations for entities sharing the same context and structure information, including two major circumstances. One is caused by the one-to-many mapping property of knowledge graphs (Fan et al., 2014; Xiao et al., 2016), where a subject has multiple objects with the same relation. As shown in the left knowledge graph in Figure 1, *(Chuck Palahniuk, Write, Pygmy)* and *(Chuck Palahniuk, Write, Tell-All)* have the one-to-many mapping property, and EARL learns the same category representation for *Pygmy* and *Tell-All*, which is suitable for the dialogue con-

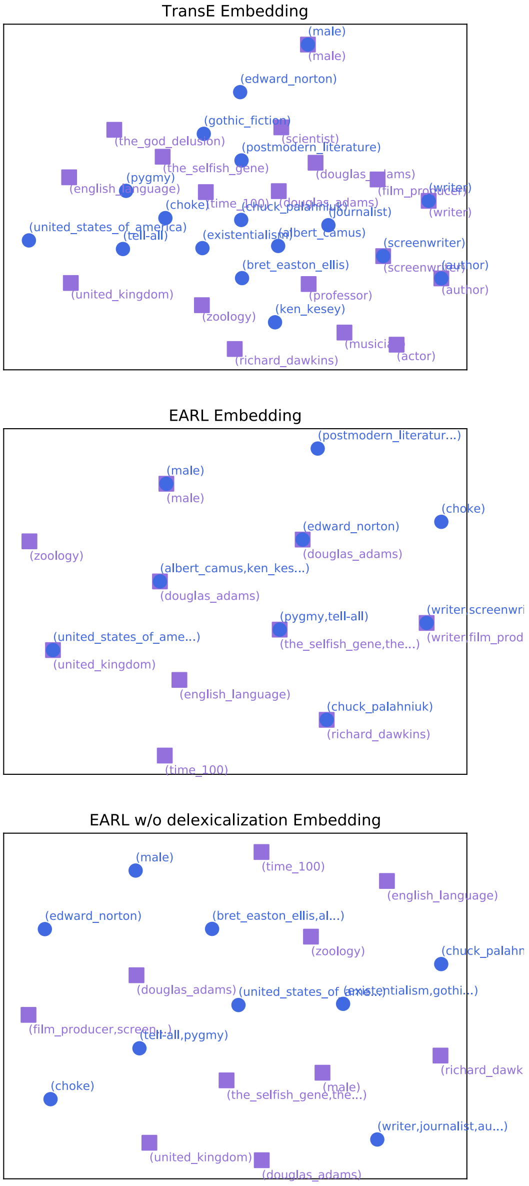


Figure 2: t-SNE projection of TransE, EARL, and EARL without delexicalization embeddings, where blue circles and purple squares represent entities from two knowledge graphs in Figure 1.

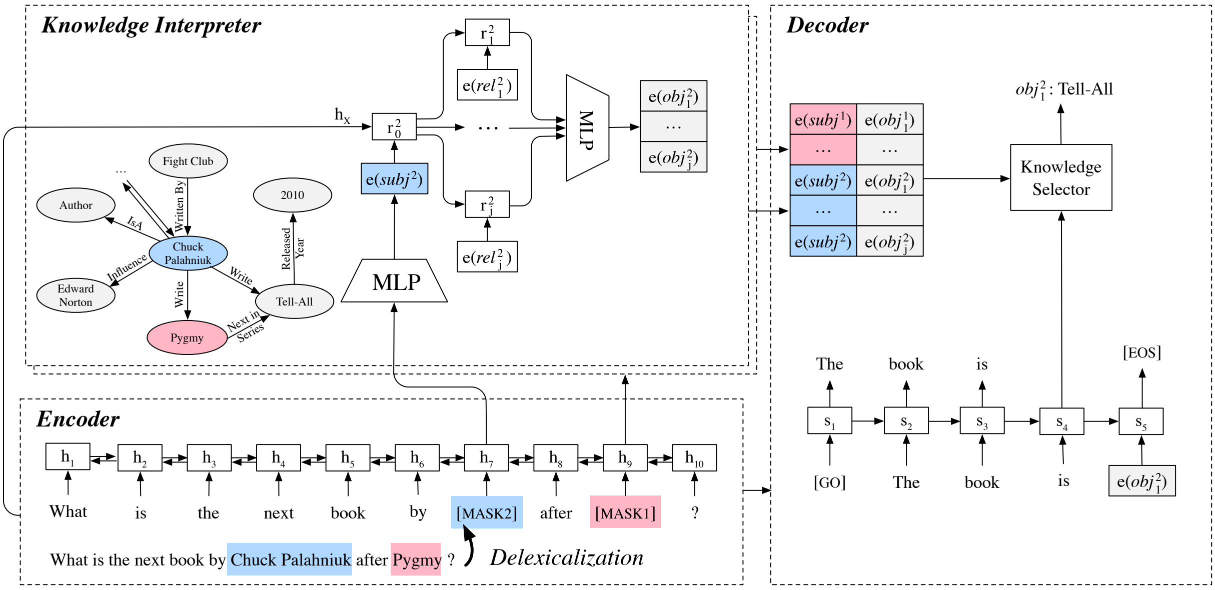


Figure 3: Overview of EARL. The blue and red content denote two subject entities in the context, and the grey content represent the object entities in the knowledge graph. Entities are represented by the knowledge interpreter and stored in the memory module of the decoder, where $subj^i$ and obj_j^i denote the ith subject entity and the jth object entity corresponding to the ith subject entity, respectively.

text inquiring about books by *Chuck Palahniuk*. The other is the circumstance that the same dialogue context (after delexicalization) grounding on different knowledge graphs, where the object is connected to the subject with the same relation. For instance, the conversations in Figure 1 share the same context but different knowledge graphs, where *The Selfish Gene* in the right graph, *Pygmy* and *Tell-All* in the left graph possess the same relational structure *(subj, Write, obj)* in knowledge graphs, leading to the same category representations for these entities.

The visualizations of entity embeddings of two knowledge graphs in Figure 1 by EARL and TransE are provided in Figure 2. EARL learns different representations for entities in different conversation context or with different structure of knowledge graphs but the same representation for entities sharing the same context and relational structure information (*Pygmy*, *Tell-All*, and *The Selfish Gene*). However, there is a gap between TransE embeddings of these entities, making it difficult to utilize unseen entities.

3.3 Encoder

In the encoder, we propose a delexicalization step before encoding the context, which replaces entities in the context with a token [MASKi]. i denotes the reverse order of entities in the context, designed to allow our model to concentrate on the newest entities mentioned in the context. The delexicalization process makes EARL entity-agnostic for conversation context, which enables our model to extend to unseen entities in knowledge graphs.

After the delexicalization step, the context X is fed to a bi-directional encoder f_θ to get the hidden representation $\mathbf{H} = (\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_n)$ and $\boldsymbol{h}_X$, which are defined as follows:

$$\mathbf{H} = f_\theta(X), \tag{1}$$

$$\boldsymbol{h}_X = pooling(\mathbf{H}), \tag{2}$$

where f_θ can be implemented by Transformer (Vaswani et al., 2017) or the gated recurrent unit (GRU, Cho et al. 2014).

3.4 Knowledge Interpreter

After obtaining the hidden representations of the context, knowledge interpreter is designed to represent each entity in the knowledge graph based on the context and the structure information. For each subject entity $subj^i$ mentioned in the context, we retrieved the corresponding knowledge graph g_i, where each object entity obj_j^i can be connected to the central entity (subject) with relation rel_j^i. In order to ensure our model to be agnostic to entities, we don't learn embeddings for each entity. By contrast, we represent the mentioned entity $subj^i$ with the hidden representations of the context, and model the related object entity obj_j^i by reasoning through the structure information of the knowledge graph g_i [1]. This process is defined as follows:

$$\boldsymbol{e}(subj^i) = \mathbf{MLP}(\boldsymbol{h}_{subj^i}), \tag{3}$$

$$\boldsymbol{e}(obj_j^i) = \mathbf{MLP}(\boldsymbol{r}_j^i), \tag{4}$$

$$\boldsymbol{r}_0^i = \mathbf{GRU}(\boldsymbol{h}_X, \boldsymbol{e}(subj^i)), \tag{5}$$

$$\boldsymbol{r}_j^i = \mathbf{GRU}(\boldsymbol{r}_0^i, \boldsymbol{e}(rel_j^i)), \tag{6}$$

where $\mathbf{MLP}$ represents the multi-layer perceptron layer, $\boldsymbol{e}(subj^i)$, $\boldsymbol{e}(rel_j^i)$, $\boldsymbol{e}(obj_j^i)$ denote the embedding of the subject entity, the relation, and the object entity, respectively.

Although aforementioned methods is able to represent the relevant entities related to the context, it cannot represent entities, which are not mentioned in the context or not connected to the subject entity in the context with any path in the knowledge graph. In this case, we resort to represent the entity i with N_{r_i} relations connected to it by graph attention based on the hidden state $\boldsymbol{h}_X$ of the context, which is formulated as follows:

$$\boldsymbol{e}(subj^i) = \sum_{n=1}^{N_{r_i}} \alpha_n [\boldsymbol{h}_X; \boldsymbol{e}(rel_n^i)], \tag{7}$$

$$\boldsymbol{e}(obj^i) = \mathbf{MLP}(\boldsymbol{e}(subj^i)), \tag{8}$$

$$\alpha_n = \frac{\exp(\beta_n)}{\sum_{j=1}^{N_{r_i}} \exp(\beta_j)}, \tag{9}$$

$$\beta_n = \boldsymbol{e}(rel_n^i)^\top \tanh(\mathbf{W_h} \boldsymbol{h}_X), \tag{10}$$

where $\boldsymbol{e}(rel_n^i)$ denotes the embedding of the relation n connected to the entity i, $\boldsymbol{e}(subj^i)$ and $\boldsymbol{e}(obj^i)$ are two representations of a same entity i, serving as the subject and object entity embeddings used in the decoding process.

3.5 Decoder

The decoder g_θ is a unidirectional neural network with the attention mechanism (Bahdanau et al., 2014; Vaswani et al., 2017) conditioning on the hidden representation of the context $\mathbf{H}$, which updates its state as follows:

$$\boldsymbol{s}_t = g_\theta(\boldsymbol{e}(y_{t-1}), \boldsymbol{s}_{<t}, \mathbf{H}). \tag{11}$$

In order to generate related entities from knowledge graphs during decoding, a knowledge selector is designed to allow the decoder to select object entities from knowledge graphs or words from the vocabulary. Inspired by Tu et al., 2016, we also introduce a coverage mechanism to facilitate the decoder to avoid generating repetitive entities. The decoding process is formulated as follows:

$$g_t = \text{sigmoid}(\boldsymbol{v}_s^\top \boldsymbol{s}_t), \tag{12}$$

$$P_g(y_t = w_g) = \text{softmax}(\mathbf{W_g} \boldsymbol{s}_t), \tag{13}$$

$$P_e(y_t = obj_j^i) = \frac{\exp(\gamma_t^{i,j})\ cov_t^{i,j}}{\sum_{x=1}^{|subj|} \sum_{y=1}^{|obj|} \exp(\gamma_t^{x,y})\ cov_t^{x,y}}, \tag{14}$$

$$\gamma_t^{i,j} = [\boldsymbol{e}(subj^i); \boldsymbol{e}(obj_j^i)]^\top \mathbf{W_e} \boldsymbol{s}_t, \tag{15}$$

$$cov_t^{i,j} = \begin{cases} 0, \text{if} \quad obj_j^i \in \{y_{<t}\} \\ 1, \text{otherwise} \end{cases}, \tag{16}$$

$$P(y_t) = \begin{bmatrix} (1-g_t) P_g(y_t = w_g) \\ g_t P_e(y_t = obj_j^i) \end{bmatrix}, \tag{17}$$

$$\boldsymbol{e}(y_t) = \begin{cases} \boldsymbol{e}(w_g), \text{if} \quad y_t = w_g \\ \boldsymbol{e}(obj_j^i), \text{if} \quad y_t = obj_j^i \end{cases}, \tag{18}$$

where $g_t \in [0, 1]$ is a scalar to balance the choice between an entity obj^i and a generic word w_g, P_g / P_e is the distribution over generic words / entities respectively, and $P(y_t)$ is the final word decoding distribution.

3.6 Loss Function

The loss function is the cross entropy between the predicted token distribution $P(y_t)$ and the ground-truth distribution p_t in the training corpus. Additionally, we apply supervised signals on the knowledge selector to teacher-force the selection of entities or generic words. The loss on one sample $< X = (x_1, x_2, \cdots, x_n), Y = (y_1, y_2, \cdots, y_m) >$is defined as:

$$\mathcal{L}(\theta) = -\frac{1}{m} \sum_{t=1}^{m} \boldsymbol{p}_t \log(P(y_t)) - \lambda \sum_{t=1}^{m} (\frac{q_t}{\alpha} \log(g_t) + \frac{1-q_t}{\beta} \log(1-g_t)), \tag{19}$$

where $\boldsymbol{p}_t$ is the one-hot vector of the ground-truth y_t, g_t is the probability of selecting an entity word or a generic word, $q_t \in \{0, 1\}$ is the true choice of an entity word or a generic word in Y, α and β are the number of entity words and the number

[1] This method can be straightforward extended to represent the object entity, which is connected to the subject entity in L hops as $path^j = (subj^i, rel_1^j, rel_2^j, \cdots, rel_L^j, obj^j)$. Due to the length limit, we leave it as future work.

of generic words in a batch, respectively. The second term is used to supervise the probability of selecting an entity word or a generic word.

4 Experiments

4.1 Datasets

We adopt two knowledge graph enhanced conversation generation datasets in our experiments:

The DuConv dataset[2]: a knowledge graph enhanced conversation dataset in Chinese proposed by Wu et al. (2019). It has 29,858 dialogues and 270,399 utterances in the domain of Movies. DuConv constructs the knowledge graph with the information crawled from a movie website as the external knowledge, which contains 3,598,246 fact triples over 143,627 entities and 45 relations. However, only the training data is released with the knowledge information, which contains 19,858 dialogues. After filtering the noisy data, we randomly split the corpus into the train (80%), validation (10%), and test sets (10%). The test set consists of the seen test set (5%) and the unseen test set (5%), where the former contains the knowledge graphs that appeared during the training process, and the latter contains the knowledge graphs, of which the subject entities and most of the object entities are unseen in the training process. The statistics is shown in Table 2.

The OpenDialKG dataset[3]: a knowledge graph enhanced conversation dataset in English proposed by Moon et al. (2019). It has 15,673 dialogues and 91,209 utterances in four domains, including Movies, Books, Sports, and Music. OpenDialKG uses the Freebase (Bast et al., 2014) knowledge graph as the external knowledge, which contains 1,190,658 fact triples over top 100,813 entities and 1,358 relations. However, the released data only consists of 13,776 dialogues, which contains some noisy data, *e.g.* empty utterances in the dialogue. After filtering the noisy data, we randomly split the corpus in the same way as DuConv. The statistics is presented in Table 2.

4.2 Baselines

We chose several suitable baselines:

- **Seq2Seq**: a sequence to sequence (Seq2Seq) model (Sutskever et al., 2014) implemented by Recurrent Neural Network (RNN) (Mikolov et al., 2010), which is widely used in open-domain conversation systems.

Dataset	Conversations			Knowledge Graphs	
DuConv	Training		14,845	Entity	12,909
	Validation		1,800	Relation	39
	Test	Seen	900	Triple	113,959
		Unseen	900		
OpenDialKG	Training		10,583	Entity	100,717
	Validation		1,200	Relation	1,380
	Test	Seen	600	Triple	1,172,552
		Unseen	600		

Table 2: Statistics of datasets and knowledge graphs.

- **DIALOGPT**: a pre-trained dialogue model (Zhang et al., 2020; Wang et al., 2020) based on transformers, which is widely adopted in dialogue generation.

- **MemNet**: a knowledge-grounded model adapted from (Ghazvininejad et al., 2018), of which the memory units store word embeddings of knowledge triples.

- **PostKS**: a knowledge-grounded model selecting knowledge by prior and posterior distributions proposed by Wu et al. (2019), where we adopt word embeddings, instead of the RNN knowledge encoder, to represent knowledge triples.

- **CopyNet**: a copy network model (Zhu et al., 2017), which represents knowledge triples by word embeddings, and can copy words from knowledge triples or generate words from the vocabulary.

- **CCM**: a knowledge graph enhanced conversation model proposed by Zhou et al. (2018), which represents knowledge graphs using graph attention mechanisms based on the pre-trained TransE (Bordes et al., 2013) embeddings.

4.3 Implementation Details

We used Tensorflow(Abadi et al., 2016) and Pytorch(Paszke et al., 2017) to implement our model and baselines. We chose RNN, implemented by GRU, as the framework for EARL to make a fair comparison with baseline models, as most of them (Zhou et al., 2018; Wu et al., 2019) are implemented by GRU. The encoder/decoder, f_θ/g_θ, has 2-layer BiGRU/GRU structures with 512 hidden cells for each layer and uses different parameters. DIALOGPT is initialized by pre-trained parameters (Zhang et al., 2020; Wang et al., 2020) and

[2] https://ai.baidu.com/broad/subordinate?dataset=duconv
[3] https://github.com/facebookresearch/opendialkg

finetuned in downstream datasets. Following prior studies (Zhou et al., 2018; Wu et al., 2019), we adopted greedy search as the decoding objective. The λ in the loss function is set to 0.1 by manual tuning. The word embedding size is set to 300. The vocabulary size is limited to 19,000/30,000 for EARL and 24,000/56,000 for baselines in OpenDialKG/DuConv datasets respectively. The TransE embedding size of entities and relations is set to 100.

We used the stochastic gradient descent (SGD) algorithm with mini-batch. The batch size and learning rate are set to 100 and 0.5, respectively. The model was run at most 20 epochs, and the training stage of each model took about one day on a GPU machine. We selected the model performing best in the validation set to evaluate in test sets. Our code is available at: `https://github.com/thu-coai/earl`.

4.4 Automatic Evaluation

Metrics[4]**:** We chose **Entity** (Zhou et al., 2018) to evaluate the ability of generating informative responses by calculating the number of entities per response. **Distinct-n** and Perplexity (**PPL**) (Serban et al., 2015) are adopted to evaluate the ability of generating diverse responses and the probability of generating ground-truth responses. We computed **Precision**, **Recall**, and **F1** scores between generated entities and ground-truth entities per response to evaluate the relevance of generated entities.

Results: As shown in Table 3, EARL obtains the best performances in most metrics on all the test sets, demonstrating that EARL can generate more informative, relevant, and diverse responses than baseline models based on both trained and untrained knowledge graphs. Specifically, EARL achieves the highest number of entities per generated response, which is nearly two times higher than the second-highest score obtained by CCM, indicating that EARL is able to generate more informative responses. Besides, EARL outperforms all the baselines in the Precision, Recall, and F1 metrics, showing that entities selected by EARL are more relevant to the ground-truth entities. Furthermore, the Distinct-3/4 scores of EARL are also higher than the baselines' scores, demonstrating that EARL can generate more diverse responses. DIALOGPT achieves the best performance in Perplexity, due to the pre-trained process and the large-scale parameters. However, it performs worse than knowledge-grounded models in metrics except for Perplexity, without the ability of utilizing the relevant knowledge. For perplexity, EARL outperforms most knowledge-grounded baselines except for CopyNet, as CopyNet learns the embeddings for each entity during training, while EARL does not parameterize any entities.

Compared to the seen test set, most of the baselines perform worse in Precision, Recall, and F1 scores on the unseen test set, leading to irrelevant entities generated, as it contains knowledge graphs and entities that do not appear during training. The decrease of Precision, Recall, and F1 becomes larger from DuConv to OpenDialKG, as the size of knowledge graphs increases (see Table 2). However, EARL achieves comparable performances on the unseen test set, even most scores are slightly higher than those on the seen test set, indicating that EARL can utilize the unseen entities in knowledge graphs during the inference process. As we provide the pre-trained TransE embeddings of the knowledge graphs in the unseen test set to CCM to build a strong baseline, its performance on the unseen test set does not decrease as other baselines, albeit still worse than the performance of EARL.

4.5 Manual Evaluation

In order to better understand the quality of generated responses from the content and knowledge perspectives, we resorted to manual evaluation through crowdsourcing. 400 contexts were randomly sampled from four test sets (100 samples for each test set) for manual annotation. We conducted the pairwise comparison between the response generated by EARL and the one by a baseline for the same context. In total, there are 1,200 pairs since we chose three baselines, which achieve top performances in automatic evaluation. For each response pair, three judges were hired to give a preference between the two responses in terms of the following two metrics. The tie was allowed. Notice that system identifiers were masked during annotation.

Metrics: We adopted two widely used metrics, **Appropriateness** and **Informativeness** as proposed in (Zhou et al., 2018). Appropriateness measures the quality of the generated response at the content level (whether the response is appropriate in relevance, coherence, and adequacy). In-

[4] BLEU (Papineni et al., 2002) is not adopted due to its low correlation with human judgment, as proposed by Liu et al. (2016).

Dataset	Model	Entity	Precision	Recall	F1	Distinct-3	Distinct-4	PPL
DuConv Seen Test Set	Seq2Seq	0.068	0.020	0.013	0.015	0.128	0.201	20.54
	DIALOGPT	0.141	0.054	0.036	0.041	0.078	0.125	**9.94**
	MemNet	0.195	0.084	0.062	0.068	0.179	0.278	19.88
	PostKS	0.131	0.051	0.036	0.040	0.135	0.232	25.30
	CopyNet	0.650	0.399	0.396	0.379	0.255	0.378	15.63
	CCM	0.655	0.376	0.392	0.365	0.239	0.350	20.71
	EARL	**1.269**	**0.435**	**0.478**	**0.422**	**0.379**	**0.519**	17.00
DuConv Unseen Test Set	Seq2Seq	0.062	0.020	0.014	0.016	0.128	0.200	19.45
	DIALOGPT	0.133	0.047	0.042	0.049	0.079	0.127	**10.50**
	MemNet	0.195	0.074	0.048	0.055	0.175	0.269	19.37
	PostKS	0.110	0.054	0.034	0.040	0.126	0.212	24.13
	CopyNet	0.684	0.339	0.342	0.324	0.249	0.365	13.13
	CCM	0.686	0.421	0.445	0.410	0.247	0.364	17.41
	EARL	**1.310**	**0.457**	**0.525**	**0.455**	**0.383**	**0.520**	14.02
OpenDialKG Seen Test Set	Seq2Seq	0.160	0.043	0.026	0.031	0.114	0.166	23.14
	DIALOGPT	0.231	0.094	0.065	0.071	0.275	0.391	**8.43**
	MemNet	0.226	0.060	0.041	0.046	0.157	0.229	22.13
	PostKS	0.190	0.048	0.030	0.035	0.172	0.259	25.39
	CopyNet	0.335	0.176	0.116	0.132	0.214	0.302	19.81
	CCM	0.759	0.212	0.182	0.183	0.251	0.328	24.98
	EARL	**1.712**	**0.268**	**0.357**	**0.287**	**0.336**	**0.421**	21.17
OpenDialKG Unseen Test Set	Seq2Seq	0.138	0.030	0.020	0.022	0.102	0.147	20.69
	DIALOGPT	0.157	0.073	0.049	0.055	0.249	0.354	**10.39**
	MemNet	0.148	0.039	0.026	0.029	0.137	0.205	20.07
	PostKS	0.206	0.042	0.025	0.029	0.143	0.223	24.84
	CopyNet	0.285	0.157	0.104	0.117	0.179	0.258	17.75
	CCM	0.760	0.257	0.223	0.221	0.259	0.334	21.94
	EARL	**1.630**	**0.322**	**0.410**	**0.336**	**0.349**	**0.426**	18.49

Table 3: Automatic evaluation in four test sets, where "Unseen" denotes the test set contains unseen entities in knowledge graphs.

Dataset	Model	App.	Inf.
DuConv Seen Test Set	EARL vs. MemNet	**0.649**	**0.933**
	EARL vs. CopyNet	**0.714**	**0.625**
	EARL vs. CCM	**0.645**	0.531
DuConv Unseen Test Set	EARL vs. MemNet	**0.629**	**0.953**
	EARL vs. CopyNet	**0.650**	**0.702**
	EARL vs. CCM	0.553	**0.569**
OpenDialKG Seen Test Set	EARL vs. MemNet	0.556	**0.924**
	EARL vs. CopyNet	**0.679**	**0.871**
	EARL vs. CCM	0.566	**0.746**
OpenDialKG Unseen Test Set	EARL vs. MemNet	**0.615**	**0.931**
	EARL vs. CopyNet	**0.722**	**0.913**
	EARL vs. CCM	**0.615**	**0.755**

Table 4: Manual evaluation in Appropriateness (App.), and Informativeness (Inf.). The score is the percentage that EARL wins its competitor after removing "Tie" pairs, where **bold** represents EARL is significantly better (sign test, p-value < 0.05) than the baseline.

formativeness measures the quality of the generated response at the knowledge level (whether the response provides new information and relevant knowledge in response to the context).

Annotation Statistics: We calculated the Fleiss' kappa (Fleiss, 1971) to measure inter-rater consistency. Fleiss' kappa for Appropriateness and Informativeness is 0.57 and 0.49, respectively, denoting the "Moderate agreement" of the annotations. We also calculated the agreements of human annotators. For Appropriateness, the percentage of the pairs that at least two judges gave the same label (2/3 agreement[5]) amounts to 97.5%, and the percentage for 3/3 agreement is 58.3%. For Informativeness, the percentage for at least 2/3 agreement is 95.7%, and that for 3/3 agreement is 51.0%.

[5] 2/3 means 2 out of 3 annotators assign the same label to an annotation item.

Results: The results are shown in Table 4. The score is the percentage that EARL wins a baseline after removing "Tie" pairs. EARL outperforms all the baselines in terms of both metrics on all the test sets, where EARL achieves significantly better performances (sign test, p-value < 0.05) in most cases. EARL has over 90% chances to win MemNet in Informativeness on all the test sets, as MemNet cannot utilize the knowledge triples stored in the memory efficiently, leading to generic or irrelevant responses. For Appropriateness, the probabilities that EARL wins MemNet are a little lower than that of Informativeness, as the generic or high-frequency responses generated by MemNet are usually fluent in grammar. Compared to MemNet, CopyNet performs slightly better in Informativeness while worse in Appropriateness since generating more informative entities brings about difficulties in fluent and coherent conversation generation. CCM performs best among all the baselines because it can introduce the knowledge graph information taking the pre-trained TransE embeddings as input. However, its performances are still worse than EARL, especially in the unseen test sets, as the usage of knowledge graphs and entities is not finetuned during the training process, which leads to the gap of performance between the seen and unseen test sets.

Noticeably, the probabilities that EARL wins baselines achieve higher in the unseen test set, as utilizing untrained knowledge graphs are relatively difficult for baselines. However, this problem is alleviated by the entity-agnostic knowledge interpreter of EARL, which learns the representations of entities based on the context information and the knowledge graph structure information. EARL's better performances on the unseen test sets demonstrate EARL can utilize unseen entities in knowledge graphs and suitable for informative knowledge-grounded conversation generation.

5 Conclusion and Future Work

In this paper, we present an entity-agnostic representation learning method to incorporate knowledge graphs into informative conversation generation. It learns to represent entities using the relational structure of the knowledge graph instead of parameterizing billions of entities directly, thereby more suitable for applying large-scale unseen graphs. Automatic and manual evaluations show that EARL can generate appropriate and informative responses with both seen and unseen knowledge graphs as input.

In future work, we will explore the pre-trained knowledge-grounded conversation model based on EARL, which can incorporate the large-scale knowledge graphs with entities in multiple hops into conversation generation.

Acknowledgments

This work was supported by the National Science Foundation for Distinguished Young Scholars (with No. 62125604) and the NSFC projects (Key project with No. 61936010 and regular project with No. 61876096). This work was also supported by the Guoqiang Institute of Tsinghua University, with Grant No. 2019GQG1 and 2020GQG0005.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Hannah Bast, Florian Bäurle, Björn Buchhold, and Elmar Haußmann. 2014. Easy access to the freebase dataset. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 95–98.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems*, pages 2787–2795.

Paul Buitelaar and Philipp Cimiano. 2008. *Ontology learning and population: bridging the gap between text and knowledge*, volume 167. Ios Press.

Mia Xu Chen, Orhan Firat, Ankur Bapna, Melvin Johnson, Wolfgang Macherey, George Foster, Llion Jones, Niki Parmar, Mike Schuster, Zhifeng Chen, et al. 2018. The best of both worlds: Combining recent advances in neural machine translation. *arXiv preprint arXiv:1804.09849*.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Miao Fan, Qiang Zhou, Emily Chang, and Fang Zheng. 2014. Transition-based knowledge graph embedding with relational mapping properties. In *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*, pages 328–337.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5110–5117. AAAI Press.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor OK Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.

Sangdo Han, Jeesoo Bang, Seonghan Ryu, and Gary Geunbae Lee. 2015. Exploiting knowledge base to generate responses for natural language dialog listening agents. In *SIGDIAL*, pages 129–133.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. A simple, fast diverse decoding algorithm for neural generation. *CoRR*, abs/1611.08562.

Chia-Wei Liu, Ryan Lowe, Iulian Vlad Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132.

Yinong Long, Jianan Wang, Zhen Xu, Zongsheng Wang, Baoxun Wang, and Zhuoran Wang. 2017. A knowledge enhanced generative conversational service agent. In *DSTC6 Workshop*.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Interspeech*, volume 2, page 3.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854.

Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. 2016. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. In *26th International Conference on Computational Linguistics, Proceedings of the Conference*, pages 3349–3358.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS) Workshop Autodiff*.

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2015. Hierarchical neural network generative models for movie dialogues. *arXiv preprint arXiv:1507.04808*.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 1577–1586.

Louis Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. 2017. Generating long and diverse responses with neural conversation models. *CoRR*, abs/1701.03185.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 196–205.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: an open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 4444–4451.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–85.

Yi-Lin Tuan, Yun-Nung Chen, and Hung-yi Lee. 2019. Dykgchat: Benchmarking dialogue generation grounding on dynamic knowledge graphs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1855–1865.

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

Yida Wang, Pei Ke, Yinhe Zheng, Kaili Huang, Yong Jiang, Xiaoyan Zhu, and Minlie Huang. 2020. A large-scale chinese short-text conversation dataset. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 91–103. Springer.

Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019. Proactive human-machine conversation with explicit conversation goal. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3794–3804.

Han Xiao, Minlie Huang, and Xiaoyan Zhu. 2016. Transg: A generative model for knowledge graph embedding. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2316–2325.

Zhen Xu, Bingquan Liu, Baoxun Wang, Chengjie Sun, and Xiaolong Wang. 2017. Incorporating loose-structured knowledge into conversation modeling via recall-gate lstm. In *IJCNN*, pages 3506–3513. IEEE.

Albert Zeyer, Parnia Bahar, Kazuki Irie, Ralf Schlüter, and Hermann Ney. 2019. A comparison of transformer and lstm encoder decoder models for asr. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 8–15. IEEE.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI*, pages 4623–4629.

Wenya Zhu, Kaixiang Mo, Yu Zhang, Zhangbin Zhu, Xuezheng Peng, and Qiang Yang. 2017. Flexible end-to-end dialogue system for knowledge grounded conversation. *CoRR*, abs/1709.04264.

A Appendices

A.1 Ablation Study

In order to investigate the influence of coverage and delexicalization mechanisms, we conducted ablation tests where one of these mechanisms was removed from EARL each time, as shown in Table 5. As we can see, EARL performs best in Precision and Perplexity metrics, indicating that EARL, equipped with coverage and delexicalization mechanisms, is able to generate more relevant entities and responses compared to other alternatives.

Impact of Coverage Mechanism EARL without coverage mechanism achieves the best performance in Distinct-n scores. However, the improvement in diversity is caused by the repetitive entities generated in responses, as the number of repetitive entities per response is improved from 0.5%/4.2% to 23.9%/44.9% on the DuConv/OpenDialKG dataset after removing the coverage mechanism. Thus, adopting the coverage mechanism in the decoding process is helpful to alleviate entity repetition and generate informative responses.

Impact of Delexicalization Mechanism After removing the delexicalization mechanism, the performances in Precision and Perplexity decrease in four test sets, though the Entity score increases in the OpenDialKG dataset. The reason is that EARL without delexicalization introduces noises in the encoding process, as the word embeddings of unseen entities are not finetuned during training. Besides, it causes the gap between the entity-agnostic representations of trained entities and unseen entities, as shown in Figure 2, leading to the decrease in Precision.

Comparison of Conversation Framework To evaluate the generalization ability of EARL, we also integrated EARL into the Transformer framework. Similar to RNN-based EARL, Transformer-based EARL can generate informative responses

Dataset	Model	Entity	Precision	Recall	F1	Distinct-3	Distinct-4	PPL
DuConv Seen Test Set	Transformer	0.063	0.019	0.012	0.014	0.070	0.101	21.38
	EARL	**1.269**	**0.435**	0.478	**0.422**	0.379	0.519	**17.00**
	w/o coverage	1.159	0.409	0.462	0.407	**0.402**	**0.546**	17.09
	w/o delexical	1.245	0.413	**0.482**	0.417	0.394	0.533	17.04
	w/ transformer	1.006	0.370	0.385	0.354	0.252	0.343	17.55
DuConv Unseen Test Set	Transformer	0.057	0.027	0.016	0.019	0.070	0.100	19.99
	EARL	**1.310**	**0.457**	**0.525**	**0.455**	0.383	0.520	**14.02**
	w/o coverage	1.204	0.440	0.508	0.442	**0.417**	**0.563**	14.23
	w/o delexical	1.249	0.441	0.524	0.449	0.401	0.537	14.19
	w/ transformer	1.075	0.447	0.468	0.428	0.269	0.361	14.68
OpenDialKG Seen Test Set	Transformer	0.108	0.027	0.018	0.021	0.076	0.111	22.76
	EARL	1.712	**0.268**	0.357	0.287	0.336	0.421	**21.17**
	w/o coverage	**1.977**	0.258	0.374	0.284	**0.492**	**0.609**	22.89
	w/o delexical	1.886	0.266	**0.384**	**0.293**	0.362	0.443	21.88
	w/ transformer	1.913	0.255	0.369	0.279	0.333	0.422	22.02
OpenDialKG Unseen Test Set	Transformer	0.080	0.018	0.012	0.014	0.067	0.094	21.22
	EARL	1.630	**0.322**	0.410	0.336	0.349	0.426	**18.49**
	w/o coverage	**2.038**	0.294	**0.445**	0.329	**0.514**	**0.628**	20.25
	w/o delexical	1.853	0.318	0.444	**0.345**	0.380	0.466	19.30
	w/ transformer	1.943	0.304	0.413	0.324	0.360	0.459	19.88

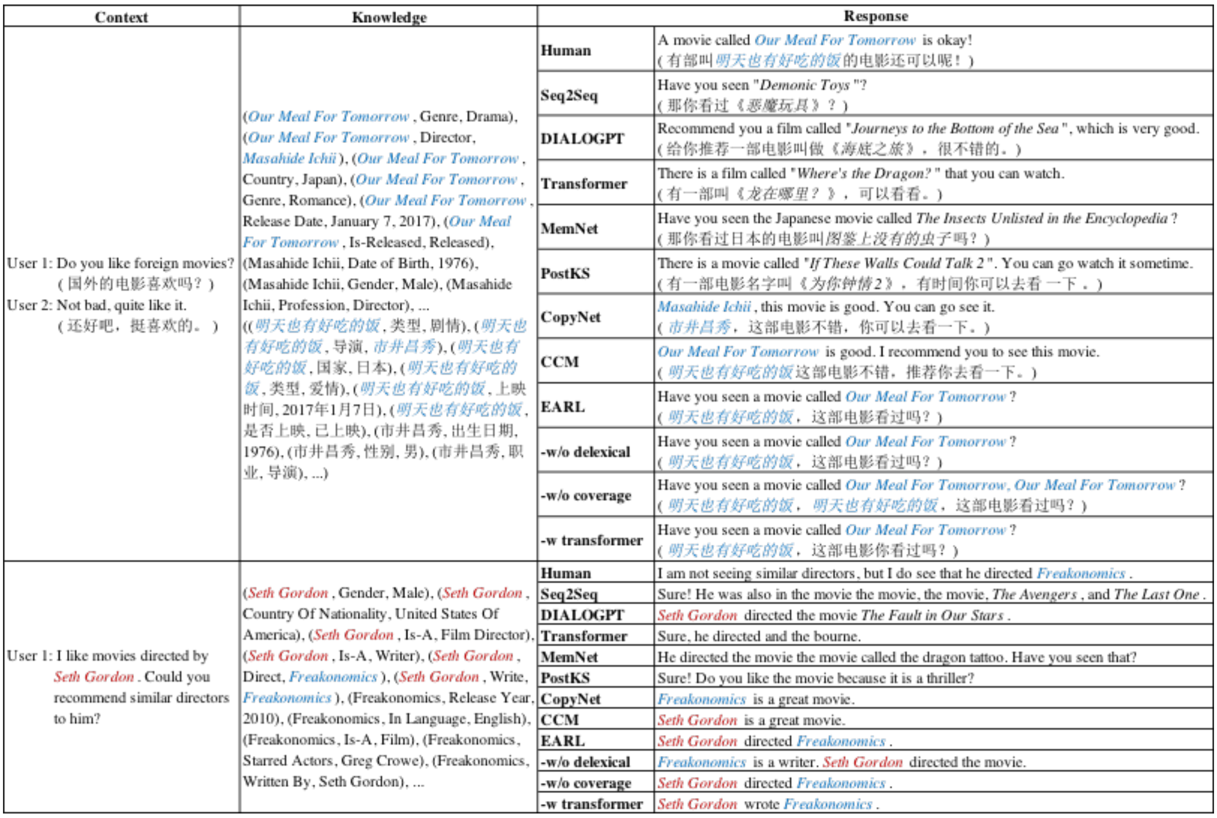
Table 5: Ablation study in four test sets, where "Unseen" denotes the test set contains unseen entities in knowledge graphs.

and outperform baselines, including the large-scale pre-trained model, DIALOGPT (see Table 3). However, it performs worse in Precision, F1, and Perplexity metrics than RNN-based EARL, which may be caused by the small datasets and model sizes (Zeyer et al., 2019; Chen et al., 2018). To make a fair comparison with baseline models, we implemented the Transformer-based EARL with 3 Transformer blocks, which is approximately equal to baseline models in model sizes. In future work, we believe the larger corpus and deeper networks may further improve the performance of EARL implemented by Transformer.

A.2 Case Study

Sample conversations are shown in Figure 4. The text in red/blue denotes the entity of the provided knowledge, which appeared in the context/response. For the first conversation, a movie in the provided knowledge called, *Our Meal For Tomorrow*, is recommended in the human response. However, Seq2Seq, DIALOGPT, and Transformer generated irrelevant movies, *Demonic Toys*, *Journeys to the Bottom of the Sea*, and *Where's the Dragon?*, without access to the provided knowledge. Although MemNet, PostKS, and CopyNet can take the knowledge as input, they also generated undesired entities, as they cannot learn a meaningful representation of the entity, *Our Meal For Tomorrow*, which has not appeared during training. CCM and our model EARL generated *Our Meal For Tomorrow* as human since they can represent the entity with the relational structure. It is noteworthy, after removing the coverage mechanism, EARL w/o coverage generated *Our Meal For Tomorrow* twice. The repetitive text undermines the quality of responses.

For the second conversation, baselines generated irrelevant content as before. Although CCM can represent entities with the relational structure, it still generated undesired content as "*Seth Gordon* is a great movie", because of the noise introduced by the word embeddings and knowledge representations of unseen entities. EARL utilized unseen knowledge more efficiently and generated "*Seth Gordon* directed *Freakonomics*" according to the knowledge "(*Seth Gordon*, Direct, *Freakonomics*)", as it learns entity-agnostic representations, which are more generalized for unseen entities. After removing the delexicalization mechanism, EARL w/o delexical generated irrelevant content, due to the noise introduced by the word embeddings of



Context	Knowledge	Response	
User 1: Do you like foreign movies? （国外的电影喜欢吗？） User 2: Not bad, quite like it. （还好吧，挺喜欢的。）	(*Our Meal For Tomorrow* , Genre, Drama), (*Our Meal For Tomorrow* , Director, *Masahide Ichii*), (*Our Meal For Tomorrow* , Country, Japan), (*Our Meal For Tomorrow* , Genre, Romance), (*Our Meal For Tomorrow* , Release Date, January 7, 2017), (*Our Meal For Tomorrow* , Is-Released, Released), (Masahide Ichii, Date of Birth, 1976), (Masahide Ichii, Gender, Male), (Masahide Ichii, Profession, Director), ... ((*明天也有好吃的饭*, 类型, 剧情), (*明天也有好吃的饭*, 导演, *市井昌秀*), (*明天也有好吃的饭*, 国家, 日本), (*明天也有好吃的饭*, 类型, 爱情), (*明天也有好吃的饭*, 上映时间, 2017年1月7日), (*明天也有好吃的饭*, 是否上映, 已上映), (市井昌秀, 出生日期, 1976), (市井昌秀, 性别, 男), (市井昌秀, 职业, 导演), ...)	**Human**	A movie called *Our Meal For Tomorrow* is okay! （有部叫*明天也有好吃的饭*的电影还可以呢！）
		Seq2Seq	Have you seen "*Demonic Toys* "? （那你看过《*恶魔玩具*》？）
		DIALOGPT	Recommend you a film called "*Journeys to the Bottom of the Sea* ", which is very good. （给你推荐一部电影叫做《*海底之旅*》，很不错的。）
		Transformer	There is a film called "*Where's the Dragon?* " that you can watch. （有一部叫《*龙在哪里？*》，可以看看。）
		MemNet	Have you seen the Japanese movie called *The Insects Unlisted in the Encyclopedia* ? （那你看过日本的电影叫*图鉴上没有的虫子*吗？）
		PostKS	There is a movie called "*If These Walls Could Talk 2* ". You can go watch it sometime. （有一部电影名字叫《*为你钟情2*》，有时间你可以去看 一下 。）
		CopyNet	*Masahide Ichii* , this movie is good. You can go see it. （ *市井昌秀*，这部电影不错，你可以去看一下。）
		CCM	*Our Meal For Tomorrow* is good. I recommend you to see this movie. （ *明天也有好吃的饭*这部电影不错，推荐你去看一下。）
		EARL	Have you seen a movie called *Our Meal For Tomorrow* ? （ *明天也有好吃的饭*，这部电影看过吗？）
		-w/o delexical	Have you seen a movie called *Our Meal For Tomorrow* ? （ *明天也有好吃的饭*，这部电影看过吗？）
		-w/o coverage	Have you seen a movie called *Our Meal For Tomorrow, Our Meal For Tomorrow* ? （ *明天也有好吃的饭*， *明天也有好吃的饭*，这部电影看过吗？）
		-w transformer	Have you seen a movie called *Our Meal For Tomorrow* ? （ *明天也有好吃的饭*，这部电影你看过吗？）
User 1: I like movies directed by *Seth Gordon* . Could you recommend similar directors to him?	(*Seth Gordon* , Gender, Male), (*Seth Gordon* , Country Of Nationality, United States Of America), (*Seth Gordon* , Is-A, Film Director), (*Seth Gordon* , Is-A, Writer), (*Seth Gordon* , Direct, *Freakonomics*), (*Seth Gordon* , Write, *Freakonomics*), (Freakonomics, Release Year, 2010), (Freakonomics, In Language, English), (Freakonomics, Is-A, Film), (Freakonomics, Starred Actors, Greg Crowe), (Freakonomics, Written By, Seth Gordon), ...	**Human**	I am not seeing similar directors, but I do see that he directed *Freakonomics* .
		Seq2Seq	Sure! He was also in the movie the movie, the movie, *The Avengers* , and *The Last One* .
		DIALOGPT	*Seth Gordon* directed the movie *The Fault in Our Stars* .
		Transformer	Sure, he directed and the bourne.
		MemNet	He directed the movie the movie called the dragon tattoo. Have you seen that?
		PostKS	Sure! Do you like the movie because it is a thriller?
		CopyNet	*Freakonomics* is a great movie.
		CCM	*Seth Gordon* is a great movie.
		EARL	*Seth Gordon* directed *Freakonomics* .
		-w/o delexical	*Freakonomics* is a writer. *Seth Gordon* directed the movie.
		-w/o coverage	*Seth Gordon* directed *Freakonomics* .
		-w transformer	*Seth Gordon* wrote *Freakonomics* .

Figure 4: Sample responses generated by all the models on the unseen test set of DuConv (upper) and OpenDialKG (lower). The *text in italic* denotes the entity, and the text in red/blue denotes the entity of the provided knowledge, which appeared in the context/response. The original text in Chinese of DuConv is presented in parentheses.

the entity, *Seth Gordon*, which causes the gap in entity-agnostic representations between seen and unseen entities.