TEGTOK: Augmenting Text Generation via Task-specific and Open-world Knowledge

Chao-Hong Tan[1]*, **Jia-Chen Gu**[1]*, **Chongyang Tao**[2], **Zhen-Hua Ling**[1]†,
Can Xu[2], **Huang Hu**[2], **Xiubo Geng**[2], **Daxin Jiang**[2]†
[1]National Engineering Research Center for Speech and Language Information Processing,
University of Science and Technology of China, Hefei, China
[2]Microsoft, Beijing, China
{chtan,gujc}@mail.ustc.edu.cn, zhling@ustc.edu.cn,
{chotao,caxu,huahu,xigeng,djiang}@microsoft.com

Abstract

Generating natural and informative texts has been a long-standing problem in NLP. Much effort has been dedicated into incorporating pre-trained language models (PLMs) with various open-world knowledge, such as knowledge graphs or wiki pages. However, their ability to access and manipulate the task-specific knowledge is still limited on downstream tasks, as this type of knowledge is usually not well covered in PLMs and is hard to acquire. To address the problem, we propose augmenting TExt Generation via Task-specific and Open-world Knowledge (TEGTOK) in a unified framework. Our model selects knowledge entries from two types of knowledge sources through dense retrieval and then injects them into the input encoding and output decoding stages respectively on the basis of PLMs. With the help of these two types of knowledge, our model can learn what and how to generate. Experiments on two text generation tasks of dialogue generation and question generation, and on two datasets show that our method achieves better performance than various baseline models.

1 Introduction

Enabling natural models to generate natural and informative sequences is a challenging yet intriguing problem of artificial intelligence and has attracted increasing attention due to its promising potentials and alluring commercial values (Bahdanau et al., 2015; Du et al., 2017; Kepuska and Bohouta, 2018; Berdasco et al., 2019; Zhou et al., 2020; Gehrmann et al., 2021). Thanks to the achievements on neural sequence modeling and pre-training technologies, current generative models are able to generate nature and fluency target sequences using either encoder-decoder architectures (Sutskever et al., 2014; Bahdanau et al., 2015; Vaswani et al., 2017) or language models (Radford et al., 2019; Brown et al., 2020; Lewis et al., 2020a) Despite these methods being the state-of-the-art frameworks for NLG, they are often provided limited knowledge to generate the desired output. Thus, the performance of text generation is still far from satisfaction in many real-world scenarios (Yu et al., 2020a).

Recently, much effort has been dedicated into incorporating traditional generative models or pre-trained language models (PLMs) with a variety of open-world knowledge, such as structural knowledge bases (e.g., ConceptNet) (Speer and Havasi, 2012; Speer et al., 2017) or unstructured documents (e.g., documents from Wikipedia) (Zhou et al., 2018c; Dinan et al., 2019). By providing the supplementary knowledge of an entity mentioned within or the background knowledge of a source text, it can help to better understand the input text and its surrounding context, and to ameliorate the informativeness of the generated text.

Although the open-world knowledge brings improvement to the generation process in most cases, its effect is still limited to the cases involving fewer entities or abstract semantics. On the other hand, the process of generating text by humans is often grounded by more than one single type of knowledge perception. In addition to world knowledge, the task-specific knowledge also acts as an important information source, and is usually not well covered in PLMs and is hard to acquire through fine-tuning. For example, in dialogue systems, what people have said or responded before can be reused as an important knowledge source, where these utterances talked before can be retained as the task-related knowledge in the mind of an interlocutor; for question generation, what part of a document makes people curious most and then ask specific questions, can often get enlightened by the existing questions raised from their corresponding passages. Intuitively, these related task-specific examples can bring additional information associated with the given source mes-

*Work done during the internship at Microsoft.
†Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2022, pages 1597 - 1609

sages and provide exemplary information for neural generative models, but this useful information source is neglected in previous studies.

On account of the above issues, we propose augmenting TExt Generation via Task-specific and Open-world Knowledge (TEGTOK). Specifically, the world knowledge is assumed to be *unstructured Wikipedia documents* that provide supplementary information of an entity mentioned within or background knowledge of an input sequence. The task-specific knowledge is a pre-built index that is domain-relevant and acts as an exemplary information source for guiding text generation. It can be flexibly adjusted according to different tasks or domains, e.g., context-response pairs in dialogue generation or passage-question pairs in question generation. Inspired by the success of dense retrieval methods for the task of open-domain question answering (Lee et al., 2019; Guu et al., 2020; Karpukhin et al., 2020), we use pre-trained encoders to convert input texts and knowledge entries into dense representation vectors and employ fast maximum inner-product search (MIPS) (Shrivastava and Li, 2014) to complete the retrieval, so as to ensure effectiveness and efficiency of knowledge selection. Finally, these two types of knowledge are injected into source text encoding and target text decoding stages respectively. By this means, our model can learn how and what to generate in a unified framework with the help of two types of knowledge.

To measure the effectiveness of our proposed framework, we evaluate it on the tasks of dialogue generation and question generation, which are both important research issues of text generation. Experimental results show that our proposed method outperforms the GPT-2 (Radford et al., 2019) and BART (Lewis et al., 2020a) baseline models, and can generate more informative texts including entities that do not appear in the input texts.

In summary, our contributions in this paper are three-fold: (1) A proposal of a general and unified text generation framework named TEGTOK that incorporates both task-specific and world knowledge through dense retrieval. (2) The proposed framework is verified on two text generation tasks.

2 Related Work

Knowledge-enhanced Text Generation. As knowledge can help to understand the input text and its surrounding context, many previous studies explored the leverage of knowledge bases (Speer and Havasi, 2012; Speer et al., 2017; Koncel-Kedziorski et al., 2019; Liu et al., 2021) or unstructured texts (Zhang et al., 2018; Zhou et al., 2018c; Dinan et al., 2019; Lewis et al., 2020b) for the text generation task, and they have demonstrated promising performance on generating informative and coherent texts. To incorporate unstructured knowledge from the web, retrieval-augmented text generation (Lewis et al., 2020b) has been widely explored. Besides, researchers also introduced the paradigm of retrieve-and-edit (Hashimoto et al., 2018; Wu et al., 2019; Ren et al., 2020) or exemplar-based decoding (Peng et al., 2019; Gupta et al., 2020) to enhance the generation processes with similar input-output pairs come from the specific task. More related works about knowledge-enhanced text generation can be referred to Yu et al. (2020b).

Dialogue Generation. The generation-based dialogue models synthesize a response with a NLG model by maximizing its generation probability given the previous conversation context. The pioneer researchers formulated the dialogue generation task as a sequence-to-sequence translation problem (Shang et al., 2015; Sordoni et al., 2015; Vinyals and Le, 2015; Serban et al., 2016, 2017) where encoder is designed for dialogue context modeling, and decoder is constructed to conduct the target response prediction. Expanded from the general dialogue generation problem, more interesting and challenging tasks relying on external knowledge have been explored to improve the anthropomorphic characteristic of dialogue systems. A line of work introduced personalized information into dialogue generation to help deliver better dialogue response such as emotion (Li and Sun, 2018; Zhou et al., 2018a; Song et al., 2019) and persona (Zhang et al., 2018; Zheng et al., 2020). In addition, to further enhance and enrich the response generation, researchers have studied grounding dialogue generation on knowledge graphs (Zhou et al., 2018b; Moon et al., 2019) or unstructured documents (Dinan et al., 2019; Zhang et al., 2018; Zhou et al., 2018c; Santhanam et al., 2020; Tan et al., 2021).

Question Generation. This task aims at generating a question from a given passage (Du et al., 2017) in an answer-aware or answer-unaware manner. In this paper, we work on the answer-unaware setting, encouraging diversity of generated

questions. Researchers have explored statistical keyword extraction techniques to select salient words from input documents, and then incorporated the extracted keywords into question generation (Cho et al., 2019; Wang et al., 2020) Recent work has applied reinforcement learning to natural question generation (Chen et al., 2020).

Different from previous text generation models that either incorporate unstructured Wikipedia knowledge or enhance the generation with exemplar cases, to the best of our knowledge, this paper makes the first attempt to retrieve and exploit both the task-specific and world knowledge for text generation in a unified framework. Our knowledge retrieval process is conducted through dense representations which can help to capture deep and latent semantics.

3 Method Formulation

The task of text generation is to output an appropriate target text given a source text as input. Given a dataset $\mathcal{D}$, an example is represented as (s, t). Specifically, s represents a source text and t represents a target text. A source text is used as a query to retrieve task-specific and world knowledge. Technically, the retrieved task-specific and world knowledge entries can be treated as two latent variables z_1 and z_2 respectively that are marginalized to get the Seq2Seq probability $p(t|s)$ via a top-*m* approximation as

$$p(t|s) = \sum_{z_1,z_2} p_1(z_1|s)p_2(z_2|s)p_\theta(t|s,z_1,z_2)$$
$$= \sum_{z_1,z_2} p_1(z_1|s)p_2(z_2|s) \prod_{t=1}^{|y|} p_\theta(t_j|s,z_1,z_2,t_{<j}), \quad (1)$$

where $z_1 \in$ top-$m(p_1(\cdot|s))$, $z_2 \in$ top-$m(p_2(\cdot|s))$, t_j and $t_{<j}$ stand for the j-th token and the first $(j-1)$ tokens of a target text t respectively, $|t|$ is the length of t, and the target text tokens are generated in an auto-regressive way. $p_1(\cdot|s)$ and $p_2(\cdot|s)$ are modeled with the retrieval probability that will be introduced in Eq. (2).

4 TEGTOK Model

Figure 1 shows the overview architecture of TEGTOK which consists of a *retriever* and a *generator*. The retriever uses the input source text as a query to retrieve the world knowledge and task-specific knowledge, the former of which is concatenated with the source text as additional background knowledge and the latter is fed into the decoder as exemplary information to guide the target text decoding. Details about each component are provided in the following subsections.

4.1 Knowledge Retriever

As shown in Figure 1(a), given a collection of a large number of knowledge entries (k_i^α), the goal of the retriever is to index all knowledge entries in a low-dimensional and continuous space, so that it can retrieve efficiently the top-*m* knowledge entries relevant to the input source text. Here, $\alpha \in$ {world knowledge (*W*), task-specific knowledge (*T*)}. Inspired by the dense passage retrieval (DPR) (Karpukhin et al., 2020), we adopt a bi-encoder architecture to derive the dense representations of the source text and each knowledge entry. Specifically, two independent pre-trained language models (i.e., BERT (Devlin et al., 2019)), $E_S^\alpha(\cdot)$ and $E_K^\alpha(\cdot)$ are employed as the encoders for the source text and the knowledge entry respectively. Furthermore, the representation of the `[CLS]` token is output as the dense representation. At retrieval-time, the retriever first maps the input source text to a vector, and then retrieves knowledge entries of which vectors are the closest to the source text vector. The similarity $s(s, k_i^\alpha)$ between the source text s and each knowledge entry k_i^α is defined using the dot product of their vectors as

$$s(s, k_i^\alpha) = E_S^\alpha(s)^\top \cdot E_K^\alpha(k_i^\alpha), i \in \{1, 2, ...\}. \quad (2)$$

Due to the significant difference between the two types of knowledge, we employ two independent retrievers for these two knowledge indexes.

World Knowledge Retriever World knowledge usually covers a wide variety of domains and has been proven effective in improving informativeness of the generated texts through providing the relevant background knowledge in open-domain text generation (Dinan et al., 2019; Zhao et al., 2020). Motivated by the success of open-domain question answering (QA) (Guu et al., 2020; Karpukhin et al., 2020; Lee et al., 2019), we assume the open-world knowledge as documents from the Wikipedia dump. Specifically, we adopt the Wikipedia dump provided in open-domain QA tasks as our open-world knowledge which is composed of over 21 millions of passages segmented from the Wikipedia pages. The goal of this retriever is to retrieve a small number of documents relevant

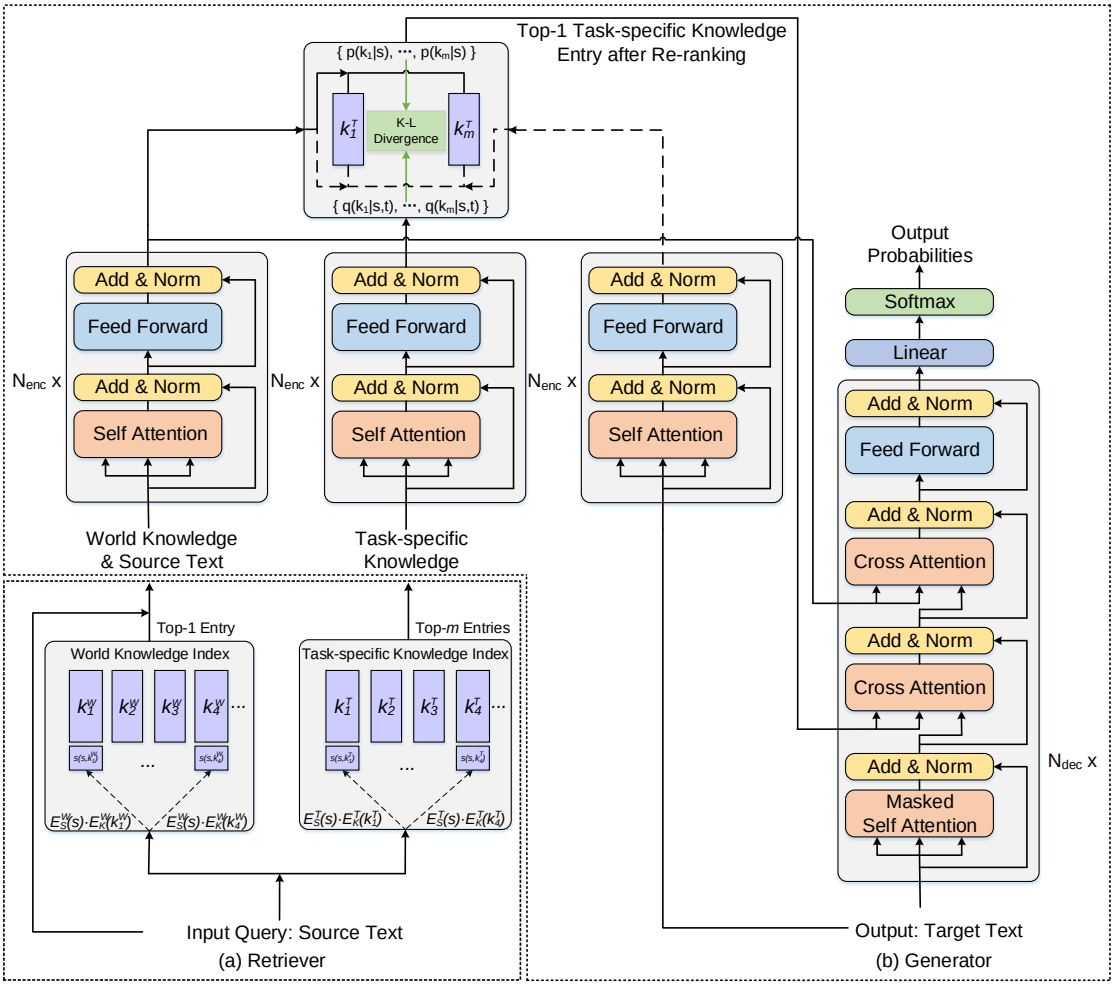

Figure 1: The overview architecture of our proposed TEGTOK model which consists of (a) a retriever and (b) a generator. Here, $E^{\alpha}_{\beta}(\cdot)$ denotes the dense representation of an input sequence, where $\alpha \in$ {world knowledge (W), task-specific knowledge (T)} and $\beta \in$ {source text (S), knowledge (K)}.

to the given source text. Meanwhile, we use the DPR model which is a pre-trained bi-encoder released by Karpukhin et al. (2020) as the world knowledge retriever in our paper, since it has achieved great performance on various knowledge-intensive tasks.[1] The retrieved top-1 Wikipedia document (k^W) is employed for augmenting source text which will be described in Section 4.2.

Task-specific Knowledge Retriever In addition to the world knowledge, it would also be desirable to obtain the relevant task-specific knowledge to guide the text generation process, since open-domain texts are often grounded by more than one single type of knowledge perception. These related task-specific examples from a pre-built index can also bring additional information associated with the given source messages and provide exemplary information for guiding the target text decoding.

Formally, given a training example represented as $(s, t^+, t_1^-, ..., t_n^-)$, where each instance contains one source text s and one matched (positive) target text t^+, along with n mismatched (negative) distractors t_i^- that are randomly sampled from the whole corpus, we can define the training objective function of the task-specific knowledge retriever as

$$L(s, t^+, t_1^-, ..., t_n^-) = -\log \frac{e^{s(s,t^+)}}{e^{s(s,t^+)} + \sum_{i=1}^{n} e^{s(s,t_i^-)}}. \quad (3)$$

At testing time, the model retrieves the top-m knowledge entries (k^T) with the highest similarities calculated by Eq. (2).

4.2 Generator

It is based on the pre-trained Transformer-based encoder-decoder architecture, BART (Vaswani et al., 2017). To incorporate both types of knowl-

[1] https://github.com/facebookresearch/DPR

edge during the source text encoding and the target text decoding stages respectively, we make several modifications as follows.

Augmented Source Text Encoder In order to incorporate the world knowledge into the source text encoding stage, we concatenate the source text with the retrieved world knowledge entry. Formally, the input sequence is organized as $\{\texttt{[BOS]}, k_1^W, ..., k_{l_{kW}}^W \texttt{[EOS]}, s_1, ..., s_{l_s}, \texttt{[EOS]}\}$, where [BOS] and [EOS] denote *begin-of-sentence* and *end-of-sentence*, $k_1^W, ..., k_{l_{kW}}^W$ and $s_1, ..., s_{l_s}$ denote the knowledge and source text tokens, and l_{kW} and l_s denote the token numbers of knowledge and source text respectively. Then the input sequence is fed into the stacked attention layers (Vaswani et al., 2017; Lewis et al., 2020a) by employing itself as *query*, *key* and *value* as

$$\mathbf{S}^{l+1} = \textsc{AttenLayer}(\mathbf{S}^l), \quad (4)$$

where $l \in \{0, ..., L-1\}$ and each AttenLayer includes operations of a self-attention layer and a feed forward layer, both of which are followed by a residual connection and a layer normalization. $\mathbf{S}^l \in \mathbb{R}^{(l_{kW}+l_s+3)\times d}$ denotes the representation of the concatenated source text and world knowledge at the l-th encoder layer, and d denotes the dimension of the embedding vector. The outputs of each encoder layer are utilized as the inputs of the next encoder layer. In each layer of encoding, the world knowledge serves as additional background and fully interacts with the source text to incorporate the relevant information into their representations through multi-head attention operations. After stacked layers of encoding, it can help to better understand the source text and return the contextualized representations, which will be further used during the decoding stage.

Task-specific Knowledge Encoder Different from the BERT-based encoding in Section 4.1 for retrieval, another encoder that is a component of the generator, is designed to encode the task-specific knowledge to derive its contextualized representations for generation. Formally, each of the retrieved top-m task-specific knowledge entries is organized as $\{\texttt{[BOS]}, k_{i,1}^T, ..., k_{i,l_i^T}^T, \texttt{[EOS]}\}, i \in \{1, ..., m\}$.[2] Then the input sequence is fed into another encoder that does not share parameters with the augmented source text encoder. Finally, we denote $\mathbf{K}_i^{T,l}$ as the representation of the i-th task-specific knowledge at the l-th encoder layer.

Task-specific Knowledge Re-ranking Since the target text cannot be foreseen at testing time, a latent variable model (Zhao et al., 2017; Lian et al., 2019; Kim et al., 2020) is introduced to select the target text by treating it as the posterior information. However, it is inefficient to calculate the prior and posterior probabilities in a large-scale dataset. Therefore, a task-specific knowledge re-ranking is designed for the top-m knowledge entries output by the knowledge retriever. In general, to further calculate the similarity between each task-specific knowledge and the target text at a fine granularity, the target text is used for re-ranking the set of retrieved task-specific knowledge entries. The target text is encoded to acquire its representation, and then combined with the representation of the augmented source text to get the posterior representation, followed by a linear transformation as

$$\mathbf{c}(s,t) = \mathbf{W}_c[\mathbf{s}_{\texttt{[BOS]}}^L; \mathbf{t}_{\texttt{[BOS]}}^{L'}] + \mathbf{b}_c, \quad (5)$$

where $\mathbf{s}_{\texttt{[BOS]}}^L$ and $\mathbf{t}_{\texttt{[BOS]}}^{L'}$ denote the outputs of the augmented source encoder and the target encoder corresponding to the [BOS] token, $\mathbf{W}_c$ and $\mathbf{b}_c$ are parameters updated during training. The similarity between this representation and the representation of each task-specific knowledge entry is calculated to obtain the probability distribution of re-ranking,

$$q_\phi(k_i^T|s,t) = \text{softmax}(\mathbf{c}(s,t) \cdot \mathbf{k}_{i,\texttt{[BOS]}}^{T,L}), \quad (6)$$

for $i \in \{1, ..., m\}$. In order to accommodate the situation where the target text is not available when testing, the prior probability is calculated as

$$p_\theta(k_i^T|s) = \text{softmax}(\mathbf{s}_{\texttt{[BOS]}}^L \cdot \mathbf{k}_{i,\texttt{[BOS]}}^{T,L}), \quad (7)$$

for $i \in \{1, ..., m\}$. Finally, two probability distributions of $q_\phi(\boldsymbol{k}^T|s,t)$ and $p_\theta(\boldsymbol{k}^T|s)$ are approximated in a way optimizing KL divergence as

$$\mathcal{L}_{kl} = \mathbb{E}_{q_\phi(\boldsymbol{k}^T|s,t)} \log \frac{q_\phi(\boldsymbol{k}^T|s,t)}{p_\theta(\boldsymbol{k}^T|s)}. \quad (8)$$

The bag-of-words (BOW) loss (Zhao et al., 2017) is introduced to facilitate the training process as

$$\mathcal{L}_{bow} = -\mathbb{E}_{k^T \sim q_\phi(\boldsymbol{k}^T|s,t)} \sum_{j=1}^{l_t} \log p(t_j|k^T), \quad (9)$$

[2] We did study concatenating the source text with the task-specific knowledge as well, but no further improvement can be achieved.

where $p(t_j|k^T)$ denotes the estimated probability of word t_j calculated by

$$p(\cdot|k^T) = \text{softmax}(\mathbf{W}_{\texttt{bow}}\mathbf{k}^{T,L}_{\texttt{[BOS]}} + \mathbf{b}_{\texttt{bow}}), \quad (10)$$

where $\mathbf{k}^{T,L}_{\texttt{[BOS]}}$ denote the outputs of the knowledge encoder corresponding to the `[BOS]` token of the selected knowledge, $\mathbf{W}_{\texttt{bow}}$ and $\mathbf{b}_{\texttt{bow}}$ are parameters updated during training.

Decoder In order to inject all the encoded information of the source text, the world knowledge and the task-specific knowledge to guide the target text decoding, two additional sub-layers are inserted into each decoder layer, which perform cross-attention over the output of the last layer of the two encoders. Particularly, after a sub-layer of masked self-attention where each token cannot attend to future tokens to avoid information leakage, the target text first attends to the output of the task-specific knowledge encoder and then attends to the output of the augmented source text encoder. Mathematically, we have

$$\begin{aligned}
\bar{\mathbf{T}}^l &= \text{LN}\Big(\mathbf{T}^l + \textsc{SelfAtten}(\mathbf{T}^l)\Big),\\
\tilde{\mathbf{T}}^l &= \text{LN}\Big(\bar{\mathbf{T}}^l + \textsc{CrossAtten}(\bar{\mathbf{T}}^l, \mathbf{K}^{T,L})\Big),\\
\hat{\mathbf{T}}^l &= \text{LN}\Big(\tilde{\mathbf{T}}^l + \textsc{CrossAtten}(\tilde{\mathbf{T}}^l, \mathbf{S}^L)\Big),\\
\mathbf{T}^{l+1} &= \text{LN}\Big(\hat{\mathbf{T}}^l + \textsc{FeedForward}(\hat{\mathbf{T}}^l)\Big),
\end{aligned} \quad (11)$$

where $l \in \{0, \cdots, L-1\}$, LN denotes the operation of layer normalization, $\mathbf{T}^l$ denotes the representation of the target text at the l-th decoder layer, $\bar{\mathbf{T}}^l$, $\tilde{\mathbf{T}}^l$ and $\hat{\mathbf{T}}^l$ are intermediate representations after each operation. In this way, the model can first learn *how to generate* and consider the retrieved task-specific knowledge as exemplary information. The model can further learn *what to say* according to the retrieved world knowledge that is used to augment the source text and enrich the exemplary information.

4.3 Learning

Given the representation of each target text token at the last decoder layer $\mathbf{T}^L = \{\mathbf{t}_j\}_{j=1}^{l_t}$ where $\mathbf{t}_j \in \mathbb{R}^d$, the probability distribution over the whole vocabulary of each target text token $\mathbf{p}_{t_j}$ can be calculated via a non-linear transformation. The learning objective of this task is to minimize the negative log-likelihood loss as

$$\mathcal{L}_{gen} = -\mathbb{E}_{k^T \sim q_\phi(\boldsymbol{k}^T|s,t)} \sum_{j=1}^{l_t} \log p(t_j|s, t_{<j}, k^T). \quad (12)$$

Finally, the parameters of our model are optimized by performing multi-task learning by minimizing the sum of all loss functions as

$$\mathcal{L}_{total} = \mathcal{L}_{gen} + \mathcal{L}_{kl} + \mathcal{L}_{bow}. \quad (13)$$

5 Experiments

We evaluated the proposed method on the tasks of dialogue generation and question generation.

5.1 Knowledge and Datasets

World Knowledge Index. For the world knowledge, all tasks and datasets shared the same English Wikipedia dump from Dec. 20, 2018 provided by Lee et al. (2019). Each Wikipedia article was split into disjoint 100-word chunks to make a total of 21M documents. Each passage was also prepended with its title, along with an `[SEP]` token.

Reddit Dataset for Dialogue Generation. To construct the task-specific knowledge index for this dataset, the Reddit dialogue corpus collected by Zhou et al. (2018b) was used. 3 millions responses were randomly sampled from the training set of the Reddit dataset. After excluding the samples used for constructing the task-specific knowledge index, the remaining dataset composed of 38.4k/10k/20k context-response pairs in the training/validation/testing sets respectively, was employed to train a generator and to evaluate the performance of our framework. Thus, there is no data overlap between that for the task-specific knowledge index and that for learning a generator.

SQuAD Dataset for Question Generation. Similarly, 45k randomly selected sentence-question pairs from the training set of the SQuAD Dataset processed by Du et al. (2017) were used to construct the task-specific knowledge index for this dataset. Also, the remaining dataset composed of 25.5k/10.5k/11.9k sentence-question pairs in the training/validation/testing sets respectively, was employed to train the generator.

5.2 Baseline Models

The following models were selected as the baseline models: **(1) RNN** (Sutskever et al., 2014) is a

Models \ Metrics	BLEU-1	BLEU-2	METEOR	ROUGE$_L$	Average	Greedy	Extrema
RNN (Sutskever et al., 2014)	7.36	2.94	7.28	10.03	0.6591	2.0585	0.3331
CVAE (Zhao et al., 2017)	7.45	2.85	7.34	9.68	0.6642	2.0853	0.3357
Transformer (Vaswani et al., 2017)	7.97	3.14	7.92	10.51	0.6693	2.0703	0.3334
GPT-2 (Radford et al., 2019)	8.43	3.04	8.33	10.65	0.6484	2.0601	0.3303
DialoGPT (Zhang et al., 2020)	7.58	3.02	7.85	10.82	0.5976	2.0774	0.3185
BART (Lewis et al., 2020a)	9.24	3.38	9.03	10.93	0.6611	2.0986	0.3355
TEGTOK	**9.71**	**3.63**	**9.53**	**11.36**	0.6522	**2.1683**	0.3362
TEGTOK w/o. WK	9.52	3.58	9.44	11.32	0.6490	2.1647	0.3361
TEGTOK w/o. TK	9.35	3.39	9.06	11.02	0.6644	2.0968	0.3371

Table 1: Performance of our method and previous methods on the test set of Reddit dataset for dialogue generation (Zhou et al., 2018b) in terms of the automated evaluation metrics. Numbers in bold denote that the improvement over the best performing baseline is statistically significant (t-test with p-value < 0.05). WK and TK denote world knowledge and task-specific knowledge respectively.

Models \ Metrics	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE$_L$
Vanilla seq2seq (Sutskever et al., 2014)	31.34	13.79	7.36	4.26	9.88	29.75
H&S (Du et al., 2017)	38.50	22.80	15.52	11.18	15.95	30.98
NQG (Du et al., 2017)	43.09	25.96	17.50	12.28	16.62	39.75
BART (Lewis et al., 2020a)	45.16	29.45	21.33	16.09	19.70	43.44
TEGTOK	**46.57**	**30.64**	**22.28**	**16.75**	**20.37**	**43.63**
TEGTOK w/o. WK	46.25	30.29	21.94	16.49	20.10	43.43
TEGTOK w/o. TK	45.63	30.02	21.88	16.56	19.79	43.61

Table 2: Performance of our method and previous methods on the test set of SQuAD dataset for question generation (Du et al., 2017) in terms of the automated evaluation metrics.

Models \ Aspects	Rel.	Flu.	Inform.	Kappa
Human	1.40	1.64	1.47	0.62
Transformer	0.86	1.07	0.71	0.42
GPT-2	1.09	1.20	0.84	0.43
BART	1.36	1.48	1.14	0.47
TEGTOK	1.44	1.51	1.23	0.46

Table 3: Human evaluation results of TEGTOK on a randomly sampled test set of the Reddit dataset. Here, Rel., Flu., and Inform. indicates relevance, fluency, and informativeness respectively.

LSTM-based sequence-to-sequence model with attention mechanism. **(2) CVAE** (Zhao et al., 2017) uses latent variables to learn a distribution over potential conversation contexts based on conditional variational autoencoders. **(3) Transformer** (Vaswani et al., 2017) uses the self-attention mechanism to build the encoder and the decoder, which has shown better performance than RNN-based Seq2Seq models in many natural language processing tasks. **(4) GPT-2** (Radford et al., 2019) is a uni-directional pre-trained language model that has shown great performance on a lot of natural language generation tasks. Following its original concatenation operation, the context and the response were concatenated with a special `[SEP]` token as input for encoding. **(5) DialoGPT** (Zhang et al., 2020) has the same architecture with GPT-2 but is trained with Reddit discussions Datasets. **(6) BART** (Lewis et al., 2020a) is a denoising autoencoder using a standard Tranformer-based neural machine translation architecture for pre-training the sequence-to-sequence models. BART is trained by corrupting text with an arbitrary noising function to reconstruct the original text.

5.3 Evaluation Metrics

To ensure all experimental results were comparable, the automated and human evaluation metrics popular used in previous work were adopted in this paper. BLEU, METEOR, ROUGE$_L$ and three embedding-based metrics including Embedding Average, Greedy Matching and Extrema Score used in Forgues et al. (2014) which can cover the weaknesses of BLEU were employed as the automated metrics. Human evaluation was also

conducted to measure the quality of the generated responses of models in terms of three independent aspects: 1) relevance (Rel.), 2) fluency (Flu.) and 3) informativeness (Inform.). Each judge was asked to give three scores for a response, each of which was ranged from 0 to 2.

5.4 Training Details

Model parameters were initialized with pre-trained weights of *bart-base* released by Wolf et al. (2020). The word embedding table was shared between the encoder and decoder. The AdamW method (Loshchilov and Hutter, 2019) was employed for optimization. The learning rate was initialized as $6.25e$-5 and was decayed linearly down to 0. The max gradient norm was clipped down to 1.0. The batch size was set to 64. The maximum length of the concatenation of open-domain knowledge and context was set to 128. The maximum length of the task-specific knowledge was set to 128. The number of task-specific knowledge entries was set to 3, achieving the best performance out of {1, 2, 3, 4, 5} on the validation set. The strategy of greedy search was performed for decoding. The maximum length of response to generate was also set to 50. All experiments were run on a single A100 GPU. The maximum number of epochs was set to 15. The validation set was used to select the best model for testing. All code was implemented in the PyTorch framework[3] and are published to help replicate our results. [4]

5.5 Evaluation Results

Automated Evaluation Table 1 and Table 2 present the evaluation results of our method and previous methods on the test sets of the Reddit dataset for dialogue generation and the SQuAD dataset for question generation respectively. Each model ran four times with identical architectures and different random initializations, and the best out of them was reported. The results show that our method outperformed all baseline models in terms of all metrics. Specifically, TEGTOK outperformed GPT-2 by 1.28% BLEU-1 and 1.20% METEOR, outperformed DialoGPT by 2.13% BLEU-1 and 1.68% METEOR, and outperformed BART by 0.47% BLEU-1 and 0.50% METEOR on the Reddit dataset. Meanwhile, TEGTOK outperformed BART by 1.41% BLEU-1 and 0.67% METEOR on the SQuAD dataset, illustrating the effectiveness of incorporating both two types of knowledge.

To further verify the effectiveness of each component in our proposed methods, ablation tests were conducted as shown in the last two rows of Table 1 and Table 2. First, the world knowledge was ablated and the results show that BLEU-1 and METEOR dropped down by 0.27% and 0.26% respectively on the Reddit dataset, along with 0.32% and 0.27% respectively on the SQuAD dataset, illustrating the effectiveness of retrieving world knowledge for text generation. On the other hand, the task-specific knowledge was ablated and only the world knowledge can be attended to during the decoding stage. The results show that BLEU-1 and METEOR dropped down by 0.24% and 0.34% respectively on the Reddit dataset, along with 0.94% and 0.58% respectively on the SQuAD dataset, illustrating the effectiveness of attending to task-specific knowledge during the decoding stage.

Human Evaluation Table 3 presents the human evaluation results on a randomly sampled test set of the Reddit dataset. 100 samples were evaluated and the order of evaluation systems were shuffled. Three judges were asked to score from 0 to 2 (2 for the best) for each human evaluation aspect and the average scores were reported. The Fleiss's kappa value (Fleiss, 1971) for each model was also reported, indicating the inter-judge moderate agreement during evaluation. In general, the results show that our method outperformed all baseline models, showing that it can generate more natural responses. Particularly, compared with BART, our method achieves the greatest improvement in terms of informativeness, illustrating the effectiveness of incorporating the task-specific and world knowledge for improving informativeness of generated texts.

5.6 Case Study

Case studies were conducted by randomly sampling an instance from the Reddit dataset in dialogue generation and an instance from the SQuAD dataset in question generation as shown in Table 4. Given the conversation context (or the passage of a question), it was used as a query to retrieve the task-specific and world knowledge in the upper block of a single case in Table 4. For case 1, as we can see that, there was no text overlap between the second task-specific knowledge entry and the conversation context, but it can be retrieved

[3]https://pytorch.org/

[4]https://github.com/lxchtan/TEGTOK

Case 1
Context: whatever happened to al qaeda? **WK**: Al-Qaeda operates as a network of Islamic extremists and Salafist jihadists. The organization has been designated as a terrorist group by the United Nations Security Council, ... The Taliban provided a safe haven for Osama bin Laden and al-Qaeda officials, allowing them to plot major terrorist attacks such as the September 11 attacks (9/11). ... **TK**: isis first iteration was al - qaeda in iraq.
Transformer: i 'm not sure what you 're talking about , but i 'm not sure if you 're referring to what you 're talking about. **GPT-2**: i think he was a member of the al qaeda branch. **DialoGPT**: they're still around. **BART**: i'm not sure. i'm sure the media is talking about the death of the leader of the country. **TEGTOK**: they're a terrorist organization in iraq plot major attacks.
Case 2
Passage: in late summer he was invited by jane stirling to visit scotland , where he stayed at calder house near edinburgh and at johnstone castle in renfrewshire , both owned by members of stirling 's family . **WK**: ... After this, in 1860 Stirling returned to Edinburgh - his address there was 4 Laverock Bank Road, Trinity, Edinburgh - which then became his permanent residence until ... **TK**: where did victoria and her family retreat to safety during a conflict in 1848?
BART: where was johnstone castle? **TEGTOK**: where did stirling stay in the summer of 1860?

Table 4: Generation results of two cases from the Reddit and SQuAD datasets respectively. We kept original texts without manual corrections. WK and TK denote world knowledge and task-specific knowledge respectively. Words in the same color are related.

through semantic relevance, which shows the effectiveness of using dense representations for knowledge retrieval. Since the given context is short and contains few informative words, it is difficult for models to generate informative responses without any external knowledge, such as the generic response generated by the Transformer model. Furthermore, our generated response can capture the relevant and important information from the retrieved knowledge, such as "*terrorist*" from the world knowledge and "*in iraq*" from the task-specific knowledge, making the generated response more informative and illustrating the effectiveness of incorporating these two types of knowledge for dialogue generation. For case 2, we can see that there was little text overlap between the world knowledge and the passage, but it could be retrieved through semantic relevance, showing the effectiveness of using dense representations for knowledge retrieval. Our generated text can capture the relevant and important information from the retrieved world knowledge, such as "*1860*" and "*Stirling*" from the world knowledge, making the generated text more informative. Furthermore, since the given passage mainly focuses on narrative descriptions, it is difficult for models to generate exemplar texts without any external knowledge, such as the "*where did ... in*" question template retrieved from the task-specific knowledge index. Again, these results illustrated the effectiveness of incorporating these two types of knowledge for question generation.

6 Conclusion

In this paper, we study retrieving relevant external knowledge for enhancing text generation. Two types of knowledge, i.e., task-specific and world knowledge, are retrieved using dense representations to ensure effectiveness and efficiency of knowledge selection, and are further incorporated into the input encoding and output decoding stages respectively, providing the supplementary information to guide text generation. Experimental results on two tasks of dialogue generation and question generation show that our method achieves better performance than baseline models and can generate more informative texts. In the future, we will explore applying this framework to more text generation tasks and other modalities such as image caption, to further verify its effectiveness and generalization.

Acknowledgements

We thank anonymous reviewers for their valuable comments.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly

learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Ana Berdasco, Gustavo López, Ignacio Díaz-Oreiro, Luis Quesada, and Luis A. Guerrero. 2019. User experience comparison of intelligent personal assistants: Alexa, google assistant, siri and cortana. In *13th International Conference on Ubiquitous Computing and Ambient Intelligence, UCAmI 2019, Toledo, Spain, December 2-5, 2019*, volume 31 of *MDPI Proceedings*, page 51. MDPI.

Tom B. Brown, Benjamin Mann, Nick Ryder, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Yu Chen, Lingfei Wu, and Mohammed J. Zaki. 2020. Reinforcement learning based graph-to-sequence model for natural question generation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Jaemin Cho, Min Joon Seo, and Hannaneh Hajishirzi. 2019. Mixture content selection for diverse sequence generation. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 3119–3129. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

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*. OpenReview.net.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1342–1352. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378–382.

Gabriel Forgues, Joelle Pineau, Jean-Marie Larchevêque, and Réal Tremblay. 2014. Bootstrapping dialog systems with word embeddings. In *Nips, modern machine learning and natural language processing workshop*, volume 2, page 168.

Sebastian Gehrmann, Tosin P. Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, et al. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. *CoRR*, abs/2102.01672.

Prakhar Gupta, Jeffrey P Bigham, Yulia Tsvetkov, and Amy Pavel. 2020. Controlling dialogue generation with semantic exemplars. *arXiv preprint arXiv:2008.09075*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.

Tatsunori B. Hashimoto, Kelvin Guu, Yonatan Oren, and Percy Liang. 2018. A retrieve-and-edit framework for predicting structured outputs. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 10073–10083.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6769–6781. Association for Computational Linguistics.

Veton Kepuska and Gamal Bohouta. 2018. Next-generation of virtual personal assistants (microsoft cortana, apple siri, amazon alexa and google home). In *IEEE 8th Annual Computing and Communication Workshop and Conference, CCWC 2018, Las Vegas, NV, USA, January 8-10, 2018*, pages 99–103. IEEE.

Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020. Sequential latent knowledge selection for knowledge-grounded dialogue. *arXiv preprint arXiv:2002.07510*.

Rik Koncel-Kedziorski, Dhanush Bekal, Yi Luan, Mirella Lapata, and Hannaneh Hajishirzi. 2019. Text Generation from Knowledge Graphs with Graph Transformers. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2284–2293, Minneapolis, Minnesota. Association for Computational Linguistics.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6086–6096. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Jingyuan Li and Xiao Sun. 2018. A syntactically constrained bidirectional-asynchronous approach for emotional conversation generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 678–683. Association for Computational Linguistics.

Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. Learning to select knowledge for response generation in dialog systems. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5081–5087. ijcai.org.

Ye Liu, Yao Wan, Lifang He, Hao Peng, and Philip S Yu. 2021. Kg-bart: Knowledge graph-augmented bart for generative commonsense reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6418–6425.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

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 Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 845–854. Association for Computational Linguistics.

Hao Peng, Ankur P. Parikh, Manaal Faruqui, Bhuwan Dhingra, and Dipanjan Das. 2019. Text generation with exemplar-based adaptive decoding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 2555–2565. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Shuo Ren, Yu Wu, Shujie Liu, Ming Zhou, and Shuai Ma. 2020. A retrieve-and-rewrite initialization method for unsupervised machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3498–3504.

Sashank Santhanam, Wei Ping, Raul Puri, Mohammad Shoeybi, Mostofa Patwary, and Bryan Catanzaro. 2020. Local knowledge powered conversational agents. *CoRR*, abs/2010.10150.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 3776–3784.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3295–3301. AAAI Press.

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.

Anshumali Shrivastava and Ping Li. 2014. Asymmetric LSH (ALSH) for sublinear time maximum inner product search (MIPS). In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2321–2329.

Zhenqiao Song, Xiaoqing Zheng, Lu Liu, Mu Xu, and Xuanjing Huang. 2019. Generating responses with a specific emotion in dialog. In *Proceedings of the 57th*

Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28-August 2, 2019, Volume 1: Long Papers, pages 3685–3695. Association for Computational Linguistics.

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. The Association for Computational Linguistics.

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, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

Robyn Speer and Catherine Havasi. 2012. Representing general relational knowledge in conceptnet 5. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 3679–3686. European Language Resources Association (ELRA).

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Chao-Hong Tan, Xiaoyu Yang, Zi'ou Zheng, Tianda Li, Yufei Feng, Jia-Chen Gu, Quan Liu, Dan Liu, Zhen-Hua Ling, and Xiaodan Zhu. 2021. Learning to retrieve entity-aware knowledge and generate responses with copy mechanism for task-oriented dialogue systems. In *Thirty-Fifth AAAI Conference on Artificial Intelligence Workshop on the Ninth Dialog System Technology Challenges (DSTC 9), AAAI 2021, Virtual Event, February 2-9, 2021*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Oriol Vinyals and Quoc V. Le. 2015. A neural conversational model. *CoRR*, abs/1506.05869.

Zhen Wang, Siwei Rao, Jie Zhang, Zhen Qin, Guangjian Tian, and Jun Wang. 2020. Diversify question generation with continuous content selectors and question type modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 2134–2143. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45. Association for Computational Linguistics.

Yu Wu, Furu Wei, Shaohan Huang, Yunli Wang, Zhoujun Li, and Ming Zhou. 2019. Response generation by context-aware prototype editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7281–7288.

Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. 2020a. A survey of knowledge-enhanced text generation. *CoRR*, abs/2010.04389.

Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. 2020b. A survey of knowledge-enhanced text generation. *arXiv preprint arXiv:2010.04389*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2204–2213. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill 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, ACL 2020, Online, July 5-10, 2020*, pages 270–278. Association for Computational Linguistics.

Tiancheng Zhao, Ran Zhao, and Maxine Eskénazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 654–664. Association for Computational Linguistics.

Xueliang Zhao, Wei Wu, Chongyang Tao, Can Xu, Dongyan Zhao, and Rui Yan. 2020. Low-resource knowledge-grounded dialogue generation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yinhe Zheng, Rongsheng Zhang, Minlie Huang, and Xiaoxi Mao. 2020. A pre-training based personalized dialogue generation model with persona-sparse data. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9693–9700. AAAI Press.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018a. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 730–739. AAAI Press.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018b. Commonsense knowledge aware conversation generation with graph attention. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4623–4629. ijcai.org.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W. Black. 2018c. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 708–713. Association for Computational Linguistics.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of xiaoice, an empathetic social chatbot. *Comput. Linguistics*, 46(1):53–93.