Lexical Knowledge Internalization for Neural Dialog Generation

Zhiyong Wu[1,2] *, **Wei Bi**[3] †, **Xiang Li**[4], **Lingpeng Kong**[1,2], **Ben Kao**[1]
[1]The University of Hong Kong, [3]Tencent AI Lab,
[2]Shanghai AI Lab, [4]East China Normal University
[1]{zywu,lpk,kao}@cs.hku.hk, [3]victoriabi@tencent.com, [4]xiangli@dase.ecnu.edu.cn

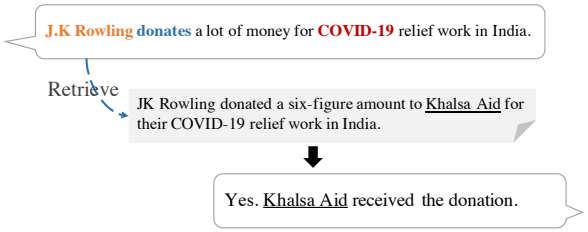
Abstract

We propose knowledge internalization (KI), which aims to complement the lexical knowledge into neural dialog models. Instead of further conditioning the knowledge-grounded dialog (KGD) models on externally retrieved knowledge, we seek to integrate knowledge about each input token internally into the model's parameters. To tackle the challenge due to the large scale of lexical knowledge, we adopt the contrastive learning approach and create an effective token-level lexical knowledge retriever that requires only weak supervision mined from Wikipedia. We demonstrate the effectiveness and general applicability of our approach on various datasets and diversified model structures.

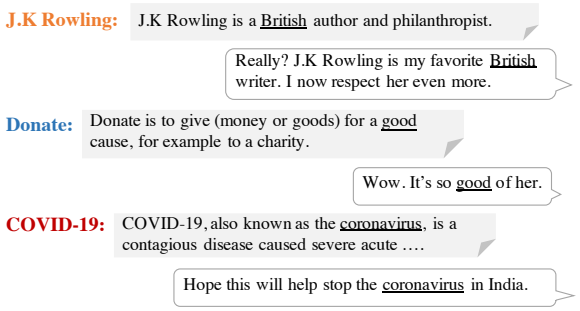
1 Introduction

Vacuous responses (Li et al., 2016; Ghazvininejad et al., 2018), such as, *I don't know*, are commonly observed in end-to-end neural dialog models (Shang et al., 2015; Sordoni et al., 2015). This is mostly because these models ignore the knowledge that resides in people's minds during a conversation. To bridge the gap, many existing works (Moghe et al., 2018; Dinan et al., 2018) have attempted to condition the dialog model on external knowledge, either a sentence or a paragraph, retrieved based on the utterance and/or previous context. This curates datasets with utterance-response-knowledge triples (see Fig 1(a)). These knowledge-grounded dialog (KGD) models, despite demonstrated effectiveness, suffer from two major problems.

First, equipping models with sentence-level knowledge alone will limit responses' informativeness and diversity. As shown in Fig 1(a), with the knowledge retrieved giving the utterance, a KGD model can relate *J.K Rowling* to *Khalsa Aid*. However, retrieval based solely on sentence embeddings will result in ignorance of lexical knowledge associated with individual tokens. In this example, the knowledge about *J.K Rowling*, *COVID-19*, *donates*, and *India*, is ignored during the retrieval, due to the semantic gaps between those lexical knowledge sentences (see Fig 1(b)) and the utterance. This makes it rather difficult (if not impossible) for the model to generate responses carrying relevant information as shown in Fig 1(b).



(a) An utterance-response-knowledge triple.



(b) Responses based on different lexical knowledge.

Figure 1: (a) An exemplary KGD data sample with an utterance (top), a response (bottom), and a sentence-level knowledge (middle). (b) A list of lexical knowledge (in grey rectangle) related to words from the utterance in (a), and the potential responses (in white speech balloon) people would make given that knowledge.

Second, retrieving knowledge for open-domain dialogs during inference incurs heavier computation, often involving similarity search over tens of millions of passages (Petroni et al., 2021). Existing systems (Zhao et al., 2020; Zheng et al., 2020) alleviate this problem relying on pre-selecting a

*The majority of this work was done while the first author was interning at Tencent AI Lab.
†Corresponding author

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 7945 - 7958

small candidate set based on TF-IDF (Schütze et al., 2008), in sacrifice of the diversity and the accuracy of the retriever. Directly conditioning the dialog model on the retrieved text, these models are easily effected by the quality of the constructed candidate set and are thus prone to errors (Dinan et al., 2018; Kim et al., 2020; Zhao et al., 2020).

In this work, we propose to complement the lexical knowledge into neural dialog models by **K**nowledge **I**nternalization (**KI**), a training approach based on contrastive learning (Hadsell et al., 2006). The central idea of KI is to integrate more fine-grained lexical knowledge about each input token internally into model parameters (e.g., word embeddings), rather than further conditioning the model on externally retrieved knowledge (e.g., directly copy and/or modify tokens from external knowledge when decoding). Our research contributions include:

- a novel training objective (KI; §3.2) that infuses lexical semantics into word representations. With the knowledge internalized into the contextualized representation of every token, a dialog model can generate informative and diverse responses *without* engaging an external knowledge retrieval module during inference time, thus making the inference more efficient (§6.1);
- an effective token-level lexical knowledge retriever (§4) trained with weak supervision to contextually align tokens in dialog corpora to their related and possibly different knowledge (Appendix C).
- a demonstration of the effectiveness and general applicability of KI with extensive experiments on diversified dialog models and on three benchmark datasets: DailyDialog (Li et al., 2017), Wizard of Wikipedia (Dinan et al., 2018), and Commonsense Reddit Dataset (Zhou et al., 2018).

2 Related Work

To address the vacuous responses problem in neural dialog models, researchers propose to ground dialogs on real world knowledge and construct new corpora that contain utterance-response-knowledge triples. Specifically, responses are grounded to external knowledge derived from different knowledge sources (Zhou et al., 2018; Liu et al., 2018; Wu et al., 2019; Dinan et al., 2018; Moghe et al., 2018; Ghazvininejad et al., 2018; Mostafazadeh et al., 2017; Meng et al., 2020; Zhang et al., 2020). Among different sources, textual knowledge (Dinan et al., 2018; Parthasarathi and Pineau, 2018; Qin et al., 2019) receives the most attention as it is easy to obtain and scale. However, the construction of knowledge-grounded datasets is costly and time-consuming. To build a more practical system without assuming a given knowledge, recent studies enhance KGD models with an extra knowledge selection component (Dinan et al., 2018; Kim et al., 2020; Zheng et al., 2020; Zhao et al., 2020).

Most existing KGD models can be viewed as *models with externalized knowledge*, where knowledge is explicitly used as part of the model input. The principle behind these models is to copy words and/or modify sentences from external knowledge when generating responses (Wu et al., 2020; Zhu et al., 2017; Zhao et al., 2019). Our KI, on the other hand, does not explicitly present knowledge to dialog models for reading and/or copying. Instead, we inject and store external knowledge into models' parameters and encourage models to elicit the encoded knowledge during generation.

The idea of knowledge internalization has also been explored in language modeling. Factual knowledge (Zhang et al., 2019; Sun et al., 2020; Liu et al., 2020), visual knowledge (Tan and Bansal, 2020) and syntactic knowledge (Kuncoro et al., 2020) have been injected into language models (LMs) and shown great promise in improving the performance of downstream tasks. KI differs from those knowledge-enhanced LMs in two aspects: (i) KI can be trained end-to-end with dialog models, while applying LMs on dialog generation often requires multiple rounds of pre-train and fine-tune. (ii) KI is lightweight that barely introduces extra parameters to the dialog model while applying LMs usually introduces hundreds of millions of extra parameters.

3 Knowledge Internalization for Neural Dialog Models

In this section, we illustrate how to train a dialog model with knowledge internalization. To infuse more fine-grained lexical knowledge to a neural dialog model, we assume a dialog corpus where each token is aligned with relevant knowledge (we will discuss the construction of such a corpus in §4). In particular, for an input sentence X in the corpus, we assume each token $x_i \in X$ is associated with a corresponding descriptive sentence K_i.

3.1 Preliminary

Given an utterance-response pair (X, Y), where $X = \{x_1, x_2, \ldots, x_n\}$ and $Y = \{y_1, y_2, \ldots, y_m\}$, neural dialog models generally minimize the negative log-likelihood loss:

$$\mathcal{L}_{\text{NLL}}(X, Y) = -\sum_{i=1}^{m} log\mathcal{P}(y_i), \qquad (1)$$

where $\mathcal{P}(y_i) = \mathcal{P}(y_i|y_{<i}, X)$ is the probability of generating the i-th response token y_i given the utterance X and other tokens generated in previous steps $y_{<i} = \{y_1, y_2, \ldots, y_{i-1}\}$. $\mathcal{P}(y_i)$ is generally modeled by a sequence-to-sequence model (Sutskever et al., 2014), which consists of an encoder and a decoder. The encoder represents X as a sequence of hidden vectors $H(X) = h_1, h_2, ..., h_n$, where each h_i is a low-dimensional representation of the token x_i. The decoder generates y_i based on $H(X)$ and $y_{<i}$, often with the attention mechanism (Bahdanau et al., 2014).

3.2 Knowledge Internalization Loss

Given a dialog corpus with token-level knowledge as discussed above, we now introduce a new training task: knowledge internalization (**KI**). In KI, we seek to boost dialog models by internalizing lexical knowledge into each token's representation. In particular, each token x_i and its associated knowledge K_i are first mapped into a low-dimensional space. We then adopt contrastive learning to shorten the distance between x_i and K_i in the space while enlarging that between x_i and other irrelevant knowledge.

Note that for each $x_i \in X$, dialog models' encoder can embed it into a contextualized representation h_i. Therefore, we only need an extra knowledge encoder to represent K_i as $g(K_i)$ (details will be given in § 4.2). After h_i and $g(K_i)$ are computed, we calculate the similarity between x_i and K_i by the inner product:

$$s(x_i, K_i) = f_1(h_i)^\top f_2(g(K_i)), \qquad (2)$$

where f_1 and f_2 are the functions that map the h_i and $g(K_i)$ into the same vector space and normalize them.

For each (x_i, K_i) pair, we randomly sample an in-batch unrelated knowledge K_i^- associated with other input sentences, where $K_i^- \neq K_i$, to construct a negative sample pair (x_i, K_i^-) in contrastive learning. Finally, the objective function of KI is defined by the contrastive loss between positive and negative sample pairs:

$$\mathcal{L}_{KI}(X) = \sum_{i=1}^{n} \max\{0, m - s(x_i, K_i) + s(x_i, K_i^-)\}, \qquad (3)$$

where m denotes the margin.

3.3 Knowledge-internalized Neural Dialog Model

We now illustrate how to deploy KI on a neural dialog model. We use a sequence-to-sequence dialog model based on Transformer (Vaswani et al., 2017) as an example. The original model is trained to minimize the negative log-likelihood loss of response tokens, i.e., $\mathcal{L}_{\text{NLL}}(X, Y)$ (see Eq. 1). We can conveniently integrate KI into the model by reusing the contextualized representations generated by the model's encoder. The training objective of a knowledge-internalized dialog model can then be formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{NLL}}(X, Y) + \lambda \mathcal{L}_{KI}(X) \qquad (4)$$

where λ is a hyperparameter. Note that the token-level knowledge is only required during training to compute $\mathcal{L}_{KI}(X)$. At the inference time, those relevant knowledge is no longer required as they have been *internalized* into model by KI, making inference more efficient.

4 Retrieval of Token-level Lexical Knowledge

In this section, we present how to train an effective retriever to mine knowledge for each token in the dialog corpora. Given a dialog utterance $X = \{x_1, x_2, \ldots, x_n\}$, the trained retriever will retrieve a relevant knowledge K_i for each token x_i in X. The constructed token-knowledge alignments can then be used to train a knowledge-internalized neural dialog model as in § 3.

4.1 Training Data Collection

To train such a retriever, we need a corpus with token-level knowledge annotated. However, to our best knowledge, no human annotated data exist and it is prohibitively expensive to build one. We therefore seek to train the retriever with distant supervision. A straight-forward solution is to align the noun words in an utterance to certain knowledge graph triples using entity linking tools (Shen et al., 2014). The problem of that is it can only cover

about 15% words in human conversations (Biber et al., 2000).

To address this issue, we propose to mine token-knowledge distant annotations from Wikipedia. In each Wiki article, the first sentence $S = \{s_1, s_2, ..., s_n\}$ is mostly declarative that gives a high-level summary on the topic of the article. Thus this sentence can used as a lexical knowledge item, denoted as K (note that K and S refer to the same sentence here). Inspired by Tan and Bansal (2020), we then further associate every token in the sentence with this knowledge item. These constructed alignments (e.g., (s_i, K)) can then be used to train a token-level knowledge retriever.

4.2 Training of Retriever

The core of the retriever's training is to learn a scoring function $r(s_i|S, K)$ to measure the relevance between a token s_i and a knowledge item K, giving s_i'context S. Similar as Eq. 2, we implement the scoring function $r(s_i|S, K)$ as the inner product between s_i'contextualized token representation $f(h_i)$ and the knowledge representation $f(g(K))$. Here, we use a pre-trained BERT (Devlin et al., 2019) model to obtain h_i; we apply another pre-trained BERT model to encode knowledge K and further generate $g(K)$ with an average-pooling operator. Two BERT models will be fine-tuned with the retriever.

Our training objective is to maximize the relevance score of aligned token-knowledge pairs while minimizing that of unaligned pairs. We also adopt the hinge loss similar as in Eq 3 by replacing x_i in the dialog corpus to s_i in the constructed token-knowledge pairs.

4.3 Mining Token-level Lexical Knowledge

Once the retriever is trained, we can use it to mine token-level lexical knowledge required in KI. We first construct a candidate knowledge base $\mathcal{K}$ that consists of 6.4 million knowledge items (first sentence) extracted from Wikipedia articles. Given a dialog utterance $X = \{x_1, x_2, \ldots, x_n\}$, we retrieve a lexical knowledge K_i for each token x_i by searching for the knowledge item that has the largest relevance score with x_i.

$$K_i = \arg\max_{K \in \mathcal{K}} r(x_i|X, K) \tag{5}$$

To improve the retrieval results, we further employ two useful strategies: (i) *Stopword Masking*, where we discard knowledge associated with stopwords; (ii) *Exact Matching*, where if an utterance token exactly matches the title of a Wikipedia article, we will directly return the first sentence of this article as the retrieval result.

The retrieval process has two properties that can significantly improve dialog corpora's knowledge coverage. First, the retrieval is *contextualized* such that a token can be aligned to different knowledge items when it occurs in different contexts. Second, the retrieval is at token-level that enables us to associate each dialog sentence with multiple knowledge items (See Appendix C).

5 Experimental Setups

In this section, we present the datasets and metrics used for evaluation.

Datasets We use three datasets from various domains (statistics in Appendix A). The first one is *DailyDialog* (Li et al., 2017), a multi-turn dialog benchmark that contains daily dialogs recorded as utterance-response pairs. However, there is no knowledge associated with the dialogs in DailyDialog, making it difficult to evaluate the informativeness of generated responses. To fully illustrate the strength of KI, we further consider two knowledge-grounded datasets: (i) *Wizard of Wikipedia (WoW)* (Dinan et al., 2018), a multi-turn dataset that contains utterance-response-knowledge triples. For each dialog, a sentence retrieved from Wikipedia is selected to guide response generation. WoW contains two test sets: Test Seen/Unseen, where the latter includes topics that never appear in Train and Valid set. (ii) *Commonsense Reddit Dataset (CRD)* (Zhou et al., 2018): a weakly knowledge-grounded single-turn dataset. Each dialog in the dataset is paired with at least one triple automatically extracted from ConceptNet (Speer et al., 2017).

Metrics We conduct both automatic evaluation and human annotations. For automatic evaluation, we evaluate the generated responses from three perspectives [1]:

• *Appropriateness*: we employ *Perplexity* (PPL), corpus-level BLEU-4 (Papineni et al., 2002) and ROUGE-l (Lin, 2004).

• *Diversity*: the ratio of distinct uni/bi-grams in all generated texts, i.e., Distinct-1/2 (Li et al., 2016).

[1]For PPL and %safe, smaller is better, while for all other metrics, larger is better.

• *Informativeness*: For WoW, we consider wikiF1 (Dinan et al., 2018), the overlapping F1 between the generated response and the grounded knowledge. For CRD, we calculate entity score (Ent.) (Zhou et al., 2018), the average number of entities per response. To further measure the likelihood of generating safe responses, we define *%safe*: the percentage of responses that contains "I'm not sure" or "I don't know". [2] We also report the accuracy of knowledge selection (ACC) following Zheng et al. (2020).

We further perform human annotations by randomly sampling 200/200/300/300 examples from WoW Test Seen/WoW Test Unseen/ CRD/DailyDialog, respectively. We recruit 5 annotators from a commercial annotation company to rate each response on a scale of 1-5 for its *appropriateness* (Zhang et al., 2020; Zheng et al., 2020) and *informativeness* (Young et al., 2018; Zhu et al., 2019). The former measures whether the topic of the response fits that of the utterance, while the latter evaluates whether a response provides new information. A response is scored 1 if it is not appropriate/informative at all, 3 if part of the response is appropriate/informative, 5 if it is highly related to utterance and context or it can provide rich information to deepen the discussion. 2 and 4 are for decision dilemmas.

6 Experiments

We evaluate the performance of KI by comparing it with three sets of baselines:

1. We first investigate the effectiveness and general applicability of KI by applying KI on conventional dialog models that are randomly initialized and trained with utterance-response pairs only.
2. We then investigate whether KI can complement or even further improve the state-of-the-art KGD model's performance.
3. As discussed in §2, although LMs differ from KI in many aspects, they also capture knowledge in their parameters. We thus compare KI with LMs to investigate its effectiveness in encouraging informative and appropriate responses.

All model structures and training setups are given in Appendix B.

[2]Upon manual inspection, we find that these two are the most common safe responses generated.

6.1 vs. Conventional Dialog Models

We first deploy KI on two representative neural dialog models that do not directly condition on any external knowledge: (i) *Seq2Seq*: a LSTM-based (Hochreiter and Schmidhuber, 1997) sequence-to-sequence model with the attention mechanism (Bahdanau et al., 2014); (ii) *Transformer* (Vaswani et al., 2017): an encoder-decoder architecture relying solely on the attention mechanisms.

Effectiveness As shown in Table 1's Setup 1 (rows 1-8), dialog models with KI consistently outperform their counterparts without KI on almost all the metrics across the datasets used. We want to point out the advantage of KI from two perspectives:

(1) Promoting informativeness. We first observe that applying KI can significantly improve the wikiF1 and Ent. scores. Unlike KGD models that can generate informative responses by explicitly copying words from given knowledge, models discussed here are not provided with any external knowledge during testing (thus copy mechanism is not applicable for them). This suggests that the improvement in informativeness should be attributed to the effectiveness of KI in injecting knowledge into models' parameters. The *Info.* scores from human evaluation in Table 2 can also substantiate our findings.

(2) Promoting diversity and reducing occurrence of safe response. Compared with the plain models, models with KI can significantly improve the Distinc-1/2 scores on all the test sets (sometimes doubled, even tripled). We also see a significant reduction of safe responses by the gap in %safe scores. Those improvements are powered by the rich lexical knowledge used in KI (see Appendix C).

Efficiency Besides the improvements in responses' quality, KI is also very efficient during inference. We report the decoding speed of Transformer and Transformer+KI in Table 3. As we can see, KI does not incur any extra computation during inference.

6.2 vs. KGD

We then apply KI on *DiffKS* (Zheng et al., 2020) [3]: a state-of-the-art model that uses a knowledge-aware decoder to generate a response based on

[3]github.com/chujiezheng/DiffKS

Setup 1: Models without externalized knowledge (trained with utterance-response pairs)

Row	Model	DailyDialog					CRD					
		PPL	BLEU-4	ROUGE-l	Distinc-1/2	%safe	PPL	Ent.	BLEU-4	ROUGE-l	Distinc-1/2	%safe
1	**Seq2Seq**	28.94	3.84	14.22	2.85/11.74	2.50	55.54	1.32	2.59	10.58	1.13/4.47	41.81
2	**Seq2Seq+KI**	29.35	4.65	14.64	3.36/14.10	2.70	47.32	2.26	2.90	11.13	1.86/7.37	35.08
3	**Transformer**	23.37	2.65	12.97	1.48/5.10	7.14	35.86	2.99	2.12	11.88	2.01/7.40	23.90
4	**Transformer+KI**	19.72	6.13	17.48	4.39/21.88	0.53	28.50	3.29	3.01	11.92	3.24/17.81	8.05

Row	Model	WoW Test Seen						WoW Test Unseen					
		PPL	wikiF1	BLEU-4	ROUGE-l	Distinc-1/2	%safe	PPL	wikiF1	BLEU-4	ROUGE-l	Distinc-1/2	%safe
5	**Seq2Seq**	77.50	6.15	1.94	10.09	1.81/5.48	53.02	144.64	6.11	1.47	10.78	2.58/10.25	36.06
6	**Seq2Seq+KI**	67.69	9.59	2.25	12.45	4.99/17.32	36.24	122.46	7.09	1.62	11.23	3.12/12.05	37.98
7	**Transformer**	48.91	6.83	2.02	11.29	1.95/4.44	83.69	93.92	5.43	1.48	10.08	1.43/3.27	84.67
8	**Transformer+KI**	46.68	10.69	2.85	12.84	5.66/18.68	35.18	93.02	7.13	1.82	11.23	3.82/12.98	41.62

Setup 2: Models with externalized knowledge (trained with utterance-response-knowledge triples)

Row	Model	WoW Test Seen						WoW Test Unseen					
		ACC	wikiF1	BLEU-4	ROUGE-l	Distinc-1/2	%safe	ACC	wikiF1	BLEU-4	ROUGE-l	Distinc-1/2	%safe
9	**DiffKS**	25.30	67.06	5.73	17.48	9.61/37.29	5.10	19.72	64.77	4.60	15.75	3.83/12.15	7.36
10	**DiffKS+KI**	26.24	74.23	6.14	17.82	9.96/39.61	6.34	21.08	72.03	5.11	16.97	4.10/13.38	8.26

Table 1: Automatic evaluation results for models with internalized knowledge (trained with utterance-response pairs), and models with externalized knowledge (trained with utterance-response-knowledge triples).

utterance and the knowledge retrieved from a set of candidates. In the empirical study, DiffKS has outperformed many KGD models like CCM (Zhou et al., 2018) [4] and ITDD (Li et al., 2019). We enhance DiffKS by applying KI on its context encoder. The rest of the model remains unchanged.

Table 1 Rows 9-10 show that DiffKS with KI improves ACC over the plain DiffKS model. The reason is that with the injection of token-level knowledge, DiffKS can better understand the utterance, which leads to more accurate knowledge selection and thus less noisy external knowledge. As a result, we observe clear gains on overlapping-based metrics (BLEU and ROUGE). These results emphasize the importance of more fine-grained knowledge in KGD. Human evaluation results (Table 2) also suggest that KI can help KGD models in generating more informative and appropriate responses.

6.3 vs. Pre-trained Language Models

We follow previous practice (Rothe et al., 2020) to replace the Transformer's encoder with LMs and keep the decoder the same as the *Transformer* above. [5] We consider two baselines: (i) *Bert2Rnd*: Initializing Transformer's encoder with a pre-trained BERT, which has been shown capturing rich factual knowledge during pre-training (Petroni et al., 2019; Jiang et al., 2020). (ii) *Ernie2Rnd*: Initializing the encoder with ERNIE 2.0 (Sun et al., 2020), a knowledge-enhanced BERT which is pre-trained with novel objectives that injecting lexical, syntactic, and semantic knowledge into its parameters (Zhang et al., 2019).

From Table 4, we see that parameters of LM-based models (Bert2Rnd and Ernie2Rnd) are more than three times than that of the Transformer baseline. But they do not seem to help improve informativeness (based on wikiF1, BLEU-4, and Info.) of responses. This indicates that although pre-trained LMs can encode knowledge in their parameters, eliciting the encoded knowledge for response generation is difficult when we only have utterance-response pairs for training. Another reason might be that previously learned knowledge is forgotten due to catastrophic forgetting (McCloskey and Cohen, 1989). Comparing with knowledge-enhanced LMs, KI is more lightweight and more effective.

In addition, we observe that introducing LMs can significantly improve responses' diversity as KI does. However, according to Appr. metric and upon manual examination, we find that although the generated responses are diverse, they are often inconsistent with the context or hallucinating non-existing facts (e.g., "Yea, Canada is the largest country in the US."). These are known issues for LMs as discussed in Dou et al. (2021); Shuster et al. (2021); Chen et al. (2020).

We also apply KI on Bert2Rnd/Ernie2Rnd, but we do not observe significant improvements as when applied on randomly initialized models. This could be due to the fact that we implement KI using knowledge from Wikipedia, which is already part of LMs' training corpora. We leave it as future work to investigate how to use KI to elicit knowledge from LMs better (e.g., use adapters (Xu et al.,

[4] Comparison with CCM is in Appendix D

[5] We keep the hidden state dimension of decoder consistent with the LMs to enable encoder-decoder attention.

Model	DailyDialog		CRD		WoW Seen		WoW Unseen		Model	WoW Seen		WoW Unseen	
	Appr.	Info.	Appr.	Info.	Appr.	Info.	Appr.	Info.		Appr.	Info.	Appr.	Info.
Transformer	3.65(1.27)	2.15(1.08)	3.65(1.27)	3.15(1.08)	3.0(1.3)	3.2(1.2)	2.9(1.2)	3.3(1.1)	**DiffKS**	3.6(1.0)	3.6(1.0)	3.7(0.9)	3.7(1.0)
Transformer+KI	4.22(1.27)	3.51(1.11)	4.22(1.27)	3.51(1.11)	3.7(0.9)	3.5(0.9)	3.7(1.0)	3.6(0.8)	**DiffKS+KI**	3.9(0.9)	4.0(0.9)	4.0(0.9)	4.2(0.8)
Human Response	4.73(1.23)	3.21(1.24)	4.73(1.23)	4.21(1.24)	4.4(0.7)	4.3(0.8)	4.5(0.7)	4.5(0.7)	**Human Response**	4.5(0.7)	4.3(0.8)	4.4(0.8)	4.2(0.8)

Table 2: Average of human annotations results on Appropriateness (Appr.) and Informativeness (Info.). Standard deviations are shown in the brackets.

Dataset	Transformer			Transformer+KI		
	sent/s	tok/s	Time(s)	sent/s	tok/s	Time(s)
DailyDialog	215	2136	31.4	192	1980	35.1
CRD	158	5263	126.7	184	4506	108.8
WoW Seen	131	3397	25.9	133	2925	25.4
WoW Unseen	152	3943	22.4	140	3331	24.3

Table 3: Number of sentences/tokens decoded per second in testing and the total decoding time (in seconds).

Setting	# Para	wikiF1	BLEU-4	Distinc-1/2	Info.	Appr.
Transformer	42.9M	6.83	2.02	1.95/4.44	3.0(1.3)	3.2(1.2)
Bert2Rnd	147.9M	4.89	0.94	3.96/15.35	2.2(1.2)	2.4(1.2)
Ernie2Rnd	147.9M	5.15	0.91	3.95/19.73	2.2(1.1)	2.3(1.2)
Transformer+KI	43.2M	**11.25**	**2.85**	5.66/18.68	**3.7(0.9)**	**3.5(0.9)**
Bert2Rnd+KI	148.5M	5.19	1.31	**8.23/40.98**	2.6(1.2)	2.6(1.2)
Ernie2Rnd+KI	148.5M	5.02	1.19	5.01/21.27	2.3(1.1)	2.4(1.2)

Table 4: Results on LMs-based dialog generation.

2021) or prompt (Liu et al., 2021)).

7 Method Analysis

In this section, we perform an in-depth analysis to understand the effectiveness of KI.

7.1 Working Principle of KI

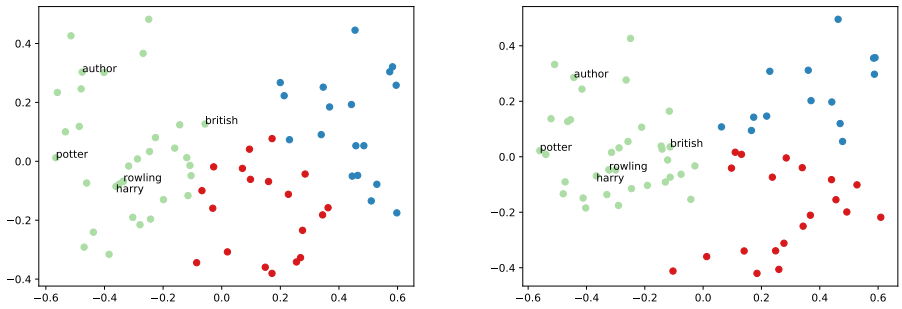
We investigate the working principle of KI by visualizing the token embeddings learned on WoW. We use principal component analysis (PCA) to map embeddings into a two-dimensional space as shown in Fig 2. Since there is no co-occurrence of British and Rowling in WoW, their embeddings learned by Transformer are distant (see Fig 2(a)). However, their embeddings learned by Transformer+KI (see Fig 2(b)) are much closer. This is because KI injects lexical knowledge (i.e., a British author) into the embedding of Rowling. Specifically, the Euclidean distances between British and Rowling are 0.37 for Transformer and 0.22 for Transformer+KI, respectively. This observation sheds light on the working principle of KI: the contrastive learning objective shortens the embedding distance between a token and tokens from its lexical knowledge. Thus when decoding, if a token is predicted (e.g. Rowling), its relevant knowledge tokens (e.g., British) are likely to receive high probabilities and be selected in the following steps (see the J.K Rowling example in Fig 1(b).

7.2 Effectiveness of Token-level Knowledge

Firstly, we experiment with a model variant (denoted as Random), which randomly assign knowledge to each utterance token. Results in Table 5 (Row 2) validate the effectiveness of the proposed token-knowledge retriever.

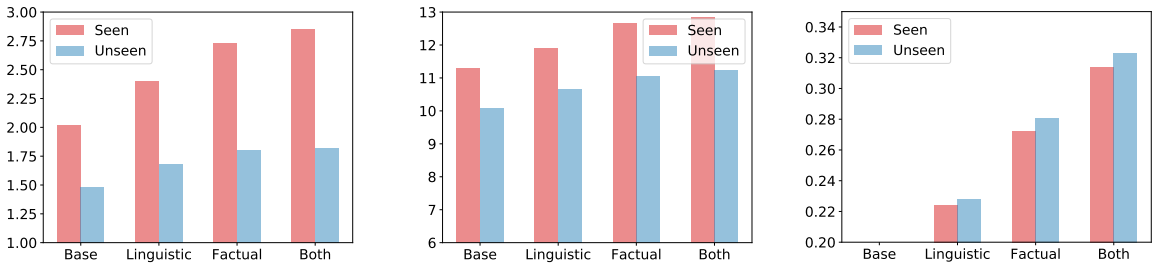
To further show the advantage of token-level knowledge, we consider a model variant in which we degenerate token-level KI to sentence-level by assigning all utterance tokens to a same lexical knowledge (we denote it as *Sentence-level knowledge*). Given the lexical knowledge retrieved for each token in an utterance, the sentence-level knowledge is chosen as the most-frequent one among all token-level knowledge. The results are summarized in Table 5 (Row 3). Note that token-level knowledge results in better performance than sentence-level knowledge. This shows that fine-grained information is useful in promoting more informative and diverse responses.

Lastly, we dive deep into the lexical knowledge retrieved to investigate which type of knowledge is most helpful in response generation. We classify a retrieved knowledge into two types: factual knowledge, which describes a real-world subject (e.g., knowledge about J.K Rowling), and is often associated with noun words in the utterance; linguistic knowledge, which explains the meaning of certain words (e.g., knowledge about donate, see Fig 1(b), and is often associated with words except nouns.We use part-of-speech (POS) tags to classify tokens and their associated knowledge. We consider two model variants that only use factual/linguistic knowledge in KI respectively, denoted as *factual* and *linguistic*. In Fig 3, we compare these two model variants to a vanilla model without KI (denoted as *base*), and a full model that uses both knowledge (denoted as *both*). We find that injecting factual knowledge brings significant improvements on BLEU-4 and ROUGE-l. We also observe similar, albeit smaller improvements when equipping with linguistic knowledge. More interestingly, these two types of knowledge can complement one another to further improve the model



(a) Transformer (b) Transformer+KI

Figure 2: Visualization of word embeddings learned by Transformer and Transformer+KI. We use words from two sources: 1) lexical knowledge retrieved for *Rowling*: "J.K. Rowling is a British author and philanthropist." 2) tokens from WoW that co-occur with "Rowling" in a sentence. Note that there is no co-occurrence of Rowling and British/author in WoW. All words are lower cased in the visualization. We use the K-means algorithm to group tokens into 3 clusters (shown in different colors).



(a) BLEU-4 (b) ROUGE-l (c) Knowledge Coverage

Figure 3: Automatic evaluation results on WoW Test Seen and Unseen. *Base* is the Transformer baseline without KI. *Both* is the Transformer+KI, with both linguistic and factual knowledge. *Linguistic*/*Factual* only considers linguistic/factual knowledge in KI, respectively.

Row	Setting	wikiF1	BLEU-4	ROUGE-l	Distinc-1/2	%safe
1	**Token-level**	11.25	2.85	12.84	5.66/18.68	35.18
2	**Random**	5.38	1.27	9.39	1.01/2.31	92.10
3	**Sentence-level**	8.41	2.31	11.72	2.98/7.77	66.32

Table 5: Comparison of model variants for Transformer+KI, using different type of knowledge. Models are evaluated on WoW Test Seen.

performance. This emphasizes the need to consider non-factual knowledge in KGD, which is usually ignored in previous study. To understand what causes the difference between using factual and linguistic knowledge, we compute *Knowledge Coverage*: the percentage of ground truth response tokens that have been recalled in the retrieved knowledge. As we can see from Fig 3(c), factual knowledge is more helpful because people tend to respond based on knowledge related to subjects (usually nouns) appearing in the dialog.

7.3 Case Study

We show an example case in Appendix E to demonstrate how KI improves dialog generation and what the limitation is.

8 Conclusion

We propose knowledge internalization (KI), which aims to incorporate the lexical knowledge into neural dialog models. Models with KI can generate informative and diverse responses without explicitly conditioning on external knowledge. To provide the fine-grained knowledge needed in KI, we also build an effective token-level lexical knowledge retriever that contextually align tokens in a sentence to their related knowledge. We show the effectiveness and general applicability of KI by evaluating KI on various datasets and diversified model structures.

9 Acknowledgement

This project is supported by the Tencent AI Lab Rhino-Bird Focused Research Program, the Shanghai Committee of Science and Technology, China (Grant No. 21DZ1100100). This research is also partly supported by the HKU-TCL Joint Research Center for Artificial Intelligence fund (Project 200009430).

References

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

Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 2000. Longman grammar of spoken and written english.

Wenhu Chen, Yu Su, Xifeng Yan, and William Yang Wang. 2020. KGPT: Knowledge-grounded pre-training for data-to-text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8635–8648, Online. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A Smith, and Yejin Choi. 2021. Scarecrow: A framework for scrutinizing machine text. *arXiv preprint arXiv:2107.01294*.

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 AAAI Conference on Artificial Intelligence*, volume 32.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020. Sequential Latent Knowledge Selection for Knowledge-Grounded Dialogue. In *International Conference on Learning Representations*.

Adhiguna Kuncoro, Lingpeng Kong, Daniel Fried, Dani Yogatama, Laura Rimell, Chris Dyer, and Phil Blunsom. 2020. Syntactic structure distillation pretraining for bidirectional encoders. *Transactions of the Association for Computational Linguistics*, 8:776–794.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. 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*, pages 7871–7880, Online. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Zekang Li, Cheng Niu, Fandong Meng, Yang Feng, Qian Li, and Jie Zhou. 2019. Incremental transformer with deliberation decoder for document grounded conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 12–21, Florence, Italy. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Shuman Liu, Hongshen Chen, Zhaochun Ren, Yang Feng, Qun Liu, and Dawei Yin. 2018. Knowledge

diffusion for neural dialogue generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1498, Melbourne, Australia. Association for Computational Linguistics.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-bert: Enabling language representation with knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2901–2908.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

Yuxian Meng, Shuhe Wang, Qinghong Han, Xiaofei Sun, Fei Wu, Rui Yan, and Jiwei Li. 2020. Openvidial: A large-scale, open-domain dialogue dataset with visual contexts. *arXiv preprint arXiv:2012.15015*.

Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M. Khapra. 2018. Towards exploiting background knowledge for building conversation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2322–2332, Brussels, Belgium. Association for Computational Linguistics.

Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios Spithourakis, and Lucy Vanderwende. 2017. Image-grounded conversations: Multimodal context for natural question and response generation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 462–472, Taipei, Taiwan. Asian Federation of Natural Language Processing.

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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Prasanna Parthasarathi and Joelle Pineau. 2018. Extending neural generative conversational model using external knowledge sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 690–695, Brussels, Belgium. Association for Computational Linguistics.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. 2021. Kilt: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473.

Lianhui Qin, Michel Galley, Chris Brockett, Xiaodong Liu, Xiang Gao, Bill Dolan, Yejin Choi, and Jianfeng Gao. 2019. Conversing by reading: Contentful neural conversation with on-demand machine reading. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5427–5436, Florence, Italy. Association for Computational Linguistics.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge.

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 (Volume 1: Long Papers)*, pages 1577–1586, Beijing, China. Association for Computational Linguistics.

Wei Shen, Jianyong Wang, and Jiawei Han. 2014. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*.

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 *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 196–205, Denver, Colorado. 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 AAAI Conference on Artificial Intelligence*, volume 31.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0:

A continual pre-training framework for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8968–8975.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 2*, pages 3104–3112.

Hao Tan and Mohit Bansal. 2020. Vokenization: Improving language understanding with contextualized, visual-grounded supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2066–2080, Online. Association for Computational Linguistics.

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 *Proceedings of the 30th International Conference on Neural Information Processing Systems*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Sixing Wu, Ying Li, Dawei Zhang, Yang Zhou, and Zhonghai Wu. 2020. Diverse and informative dialogue generation with context-specific commonsense knowledge awareness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5811–5820, Online. Association for Computational Linguistics.

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, Florence, Italy. Association for Computational Linguistics.

Yan Xu, Etsuko Ishii, Zihan Liu, Genta Indra Winata, Dan Su, Andrea Madotto, and Pascale Fung. 2021. Retrieval-free knowledge-grounded dialogue response generation with adapters. *arXiv preprint arXiv:2105.06232*.

Tom Young, Erik Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. 2018. Augmenting end-to-end dialogue systems with commonsense knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Houyu Zhang, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. 2020. Grounded conversation generation as guided traverses in commonsense knowledge graphs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2031–2043, Online. Association for Computational Linguistics.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

Xueliang Zhao, Wei Wu, Chongyang Tao, Can Xu, Dongyan Zhao, and Rui Yan. 2019. Low-resource knowledge-grounded dialogue generation. In *International Conference on Learning Representations*.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020. Knowledge-grounded dialogue generation with pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3377–3390, Online. Association for Computational Linguistics.

Chujie Zheng, Yunbo Cao, Daxin Jiang, and Minlie Huang. 2020. Difference-aware knowledge selection for knowledge-grounded conversation generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 115–125, Online. Association for Computational Linguistics.

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.

Qingfu Zhu, Lei Cui, Wei-Nan Zhang, Furu Wei, and Ting Liu. 2019. Retrieval-enhanced adversarial training for neural response generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3763–3773, Florence, Italy. Association for Computational Linguistics.

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

Dataset	Train	Valid	Test
WoW	166,787	17,715	8,715/8,782
CRD	3,384,185	20,000	10, 000
DailyDialog	54,889	6,005	5,700

Table 6: Dataset statistics. WoW includes two test sets: Test Seen/Unseen, where the latter contains topics that never appear in Train and Valid set.

A Dataset Statistics

B Implementation Details

The vocabulary size for DailyDialog/WoW/CRD is 14,696/22,168/22,512, respectively, with sentences tokenized using BERT's tokenizer provided by Transformers (Wolf et al., 2020). For Seq2Seq and Transformer, we use a shared vocabulary between the encoder and the decoder. In Seq2Seq, we adopt a 2-layer bidirectional LSTM as the encoder and an unidirectional one as the decoder. The hidden size is set to 256, with a dropout probability of 0.3. The Transformer we used has 6 encoder/decoder layers. The dimensions of the input layer, output layer, and inner feed-forward layer are set to 512, 512, and 1,024, respectively. The number of attention heads is set to 4.

We use Adam with $\beta_1 = 0.9$, $\beta_2 = 0.98$ for model optimization and start training with a warm-up phase where we linearly increase the learning rate from 10^{-7} to 0.005. After that we decay the learning rate proportional to the number of updates. Each training batch contains at most 4,096 source/target tokens. We early-stop the training if validation loss does not improve over ten epochs. We perform beam search with a beam size of 5. The λ (see Eq 3) is set to 1 in all our evaluation.

For Bert2Rnd and Ernie2Rnd, we initialize the Transformer's encoder with the pre-trained LMs using the Transformers (Wolf et al., 2020) and keep the decoder the same as above. Note that due to the exist of encoder-decoder attention, we modify the dimensions of input/output layer to 768 to be compatible with BERT (*bert-base-uncased*) and ERNIE (*nghuyongernie-2.0-en*). We share the embeddings between encoder and decoder. Models are learned with Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$. Learning rate is set to 1e4 with a linear scheduler. Each training batch contains 128 samples. The LMs are fine-tuned together with the decoder. We also experimented with LMs frozen, but this generally works worse.

Number of knowledge items	WoW	DailyDialog	CRD
per token	30	26	38
per sentence	15	12	9

Table 7: Averaged number of knowledge items associated with each token/sentence.

Context: one of our favorite books is the wonderful wizard of oz by author l . frank ba ##um and published in 1900 !
Knowledge: The Wonderful Wizard of Oz is an American children's novel written by author L. Frank Baum and illustrated by W. W. Denslow, originally published by the George M. Hill Company in May 1900.
Context: it ' s about a young wizard at hog ##wart ##s , right ?
Knowledge: The book follows Harry Potter, a young wizard, in his third year at Hogwarts School of Witchcraft and Wizardry.

Table 8: An example case from WoW. Given different contexts, the token *wizard* is aligned to different knowledge items.

C Analysis of Token-level Knowledge Retrieval

Since our retrieval component is based on the contextualized representations (see § 4.2), the same token can be aligned to different knowledge when it occurs in different contexts. As the supporting evidence, in Table 7, we report the averaged number of knowledge items associated with each token. In Table 8, we show an example of the same token being aligned to different knowledge items when giving different contexts. In addition, our approach exposes each dialog sentence to very diverse knowledge items. The rich lexical knowledge, both at the token-level and sentence-level, is the key to KI's good performance.

We further conduct an ablation study to investigate the effectiveness of two additional retrieval strategies: stopword masking and exact matching (§ 4.3). We remove each strategy and keep the other unchanged. The results are presented in Table 9. As we can see, both strategies are useful for generating appropriate (based on PPL, BLEU-4, and ROUGE-l), informative (based on WikiF1), and diversified (based on Distinc-1/2) responses.

Row	Setting	wikiF1	BLEU-4	ROUGE-l	Distinc-1/2	PPL
1	**Transformer+KI**	11.25	2.85	12.84	5.66/18.68	46.68
2	*wo stopwords masking*	5.23	2.63	12.68	5.48/21.74	57.27
3	*wo exact matching*	10.74	2.54	11.99	4.75/15.27	47.65

Table 9: Retrieval strategy ablation results on WoW Test Seen.

Model	PPL	Ent.
CCM	39.18	1.18
Transformer+KI	28.50	3.29

Table 10: Automatic evaluation on CRD. Numbers of CCM are taken from their paper.

Context: SpeakerA: I like Dylan's Bars, do you? SpeakerB: Yes Dylan's Candy Bar is my favorite boutique candy store.
Utterance: They have everything! I just love it.
Gold Response: Yes Ralph Lauren's daughter Dylan Lauren owns them.
Transformer: I'm not sure , but I do know that they have been around for a long time!
Transformer+KI: I love their chocolate chip cookies! They're actually the second largest candy company in the world! **Knowledge for Dylan's:** Lauren was inspired to create the store, which is asserted to be the "largest unique candy store in the world", by the Roald Dahl story of Willy Wonka the Chocolate Factory **Knowledge for like:** In English, the word like has a very flexible range of uses, ranging from conventional to non-standard.

Table 11: An example case from WoW Seen.

D Comparison with CCM

Similar to KI, CCM augments dialog corpora with token-level commonsense knowledge. In each encoding and decoding step, CCM explicitly uses the retrieved commonsense knowledge triples by concatenating their representations with the token representation. As existing KGD models, CCM also requires extra knowledge as input during both training and inference. Training CCM on the CRD dataset takes about a week on one Titan X GPU. The comparison of model performance is shown in Table 10. As we can see, there is a significant gap between CCM and Transformer+KI. Thus in §6.2, we consider applying KI on a more state-of-the-art and recent KGD model: DiffKS.

E Case Study

We show an example case in Table 11 to demonstrate how KI improves dialog generation and what the limitation is. From the generated results, Transformer returns a vacuous response, as it has no idea on what "Dylan's Candy Bar" is. However, Transformer+KI, which perceives the knowledge about "Dylan's Candy Bar" during training, gives a much more informative response. Meanwhile, we further observe some inaccuracy during the knowledge transfer ("largest" becomes "second largest"). We take this as an interesting future work.

F Comparison with BART

In § 6.3, we observe that KI can outperform models whose encoders are initialized with pre-trained BERT or ERNIE. Here we dive deeper to compare KI with a fully pre-trained seq2seq model: BART (Lewis et al., 2020). BART has demonstrated superior performance on conditional language generation, including translation, summarization, and dialogue response generation. We start from the BART-base checkpoint [6]. We fine-tune the model for five epochs with a learning rate of 3e-5. We do not report PPL since these two models use different tokenization methods. As we can see from Table 12, by introducing only a few extra parameters and computation, KI can significantly boost the Transformer's performance. Although a pre-trained BART model can generate slightly more diverse responses than Transformer+KI (higher *Distinc-1/2*), these generated responses are often inconsistent with the input (lower *BLEU-4/ROUGE-l*) or less informative (lower *WikiF1*).

[6] https://huggingface.co/facebook/bart-base

Row	Model	DailyDialog				CRD			
		BLEU-4	**ROUGE-l**	**Distinc-1/2**	**%safe**	**BLEU-4**	**ROUGE-l**	**Distinc-1/2**	**%safe**
1	**Transformer**	2.65	12.97	1.48/5.10	7.14	2.12	11.88	2.01/7.40	23.90
2	**Transformer+KI**	6.13	17.48	4.39/21.88	0.53	3.01	11.92	3.24/17.81	8.05
3	**BART-base**	0.65	13.40	4.95/19.47	5.51	0.48	11.60	5.03/26.89	7.40
Row	**Model**	**WoW Test Seen**				**WoW Test Unseen**			
		WikiF1	**BLEU-4/ROUGE-l**	**Distinc-1/2**	**%safe**	**WikiF1**	**BLEU-4/ROUGE-l**	**Distinc-1/2**	**%safe**
4	**Transformer**	6.83	2.02/11.29	1.95/4.44	83.69	5.43	1.48/10.08	1.43/3.27	84.67
5	**Transformer+KI**	10.69	2.85/12.84	5.66/18.68	35.18	7.13	1.82/11.23	3.82/12.98	41.62
6	**BART-base**	8.85	1.99/11.30	6.43/24.91	15.79	6.51	1.65/11.85	5.16/20.90	15.82

Table 12: Automatic evaluation results for Transformer+KI and BART-base.