Conversations Powered by Cross-Lingual Knowledge

Weiwei Sun[1*] Chuan Meng[1*] Qi Meng[2] Zhaochun Ren[1†]
Pengjie Ren[1†] Zhumin Chen[1] Maarten de Rijke[3,4]
[1]Shandong University, Qingdao, China [2]Microsoft Research Asia, Beijing, China
[3]University of Amsterdam [4]Ahold Delhaize Research, Amsterdam, The Netherlands
sunnweiwei@gmail.com,mengchuan@mail.sdu.edu.cn,meq@microsoft.com
{zhaochun.ren,chenzhumin@sdu.edu.cn}@sdu.edu.cn,jay.ren@outlook.com,m.derijke@uva.nl

ABSTRACT

Today's open-domain conversational agents increase the informativeness of generated responses by leveraging external knowledge. Most of the existing approaches work only for scenarios with a massive amount of monolingual knowledge sources. For languages with limited availability of knowledge sources, it is not effective to use knowledge in the same language to generate informative responses. To address this problem, we propose the task of *cross-lingual knowledge grounded conversation* (CKGC), where we leverage large-scale knowledge sources in another language to generate informative responses. Two main challenges come with the task of cross-lingual knowledge grounded conversation: (1) knowledge selection and response generation in a cross-lingual setting; and (2) the lack of a test dataset for evaluation.

To tackle the first challenge, we propose the *curriculum self-knowledge distillation* (CSKD) scheme, which utilizes a large-scale dialogue corpus in an auxiliary language to improve cross-lingual knowledge selection and knowledge expression in the target language via knowledge distillation. To tackle the second challenge, we collect a cross-lingual knowledge grounded conversation test dataset to facilitate relevant research in the future. Extensive experiments on the newly created dataset verify the effectiveness of our proposed curriculum self-knowledge distillation method for cross-lingual knowledge grounded conversation. In addition, we find that our proposed unsupervised method significantly outperforms the state-of-the-art baselines in cross-lingual knowledge selection.

CCS CONCEPTS

• **Computing methodologies** → **Discourse, dialogue and pragmatics**; **Natural language generation**.

KEYWORDS

Knowledge-grounded conversation; Cross-lingual information retrieval; Knowledge selection; Knowledge distillation

*These two authors contributed equally.
†Corresponding author.

SIGIR '21, July 11–15, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3462883

ACM Reference Format:
Weiwei Sun, Chuan Meng, Qi Meng, Zhaochun Ren, Pengjie Ren, Zhumin Chen, and Maarten de Rijke. 2021. Conversations Powered by Cross-Lingual Knowledge. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3404835.3462883

1 INTRODUCTION

Recent years have witnessed a rapid development of technology to support open domain human-machine conversations [1, 14, 44, 50, 55]. Although existing models are capable of generating fluent responses based on the conversational history, there is still a clear gap when people converse with such systems, compared with conversations between humans. One primary reason is that a lack of proper knowledge in generated responses makes it difficult for conversational methods to dive deeply into a specific topic [24]. To bridge this gap, the task of knowledge-grounded conversation (KGC) has been proposed so as to leverage external knowledge sources to enhance open-domain conversational models [10]. KGC has seen its first applications to the task of *conversational information retrieval* during the past few years [33, 42].

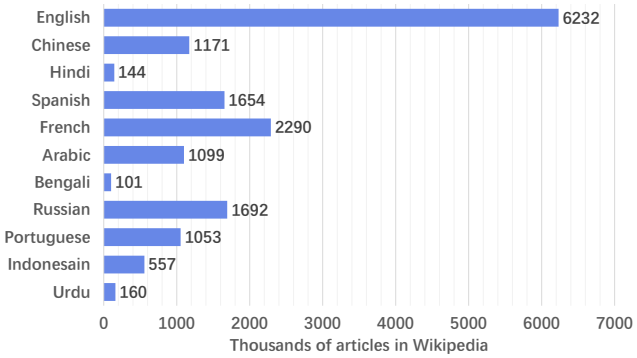

Figure 1: The number of Wikipedia articles in different languages. Note that the languages are ranked in descending order of the number of speakers.

As far as we know, existing KGC studies only focus on modeling the knowledge-grounded dialogue scenario with monolingual knowledge. However, the amount of knowledge available in different languages is extremely imbalanced. In Figure 1 we list statistics about the number of articles in Wikipedia for various languages.[1]

[1]https://meta.wikimedia.org/wiki/List_of_Wikipedias

Figure 2: (a) Open-domain dialogue grounded by monolingual knowledge (previous work); (b) For knowledge scarce languages, there is a lack of rich knowledge bases; (c) We propose cross-lingual knowledge grounded conversation (CKGC) and ground open-domain dialogues in a knowledge scarce language using cross-lingual knowledge.

We observe an imbalanced distribution over various languages, e.g., the number of documents in English is 6 times that of Chinese, whereas Urdu only has 160 thousands articles on Wikipedia. There is a theoretical possibility to establish an individual large-scale knowledge base for each language to alleviate the knowledge scarcity. But the high cost and effort required make this impractical.

To address this problem of limited knowledge sources in some languages, we propose the task of CKGCs to leverage abundant knowledge in other languages. Figure 2 shows an example of CKGC. Unlike existing KGC approaches, which only work in a monolingual domain, CKGC aims to retrieve relevant and accurate sentences from external knowledge in an auxiliary language to improve the informativeness of the response generation in the target language. Two main challenges come with the task of CKGC: (1) searching and representing cross-lingual knowledge; and (2) the lack of a dataset for evaluation. For (1), since it is difficult to construct a large-scale parallel dialogue corpora for training, we have to learn a model to retrieve and express knowledge from an auxiliary language without human annotation. For (2), we need to establish a dataset to evaluate the performance of CKGC approaches.

To tackle the first challenge, we design a two-phase framework for CKGC, with a *cross-lingual knowledge retrieval* (CKR) layer and a *multilingual response generation* (MRG) layer. On this basis, we propose a curriculum self-knowledge distillation (CSKD) scheme. With large-scale non-parallel dialogue corpora in both target and auxiliary languages, CSKD applies knowledge distillation [17] to improve cross-lingual knowledge selection (KS) and knowledge expression (KE). To this end, CSKD is composed of 3 ingredients: (1) *parallel dialogue mining*, (2) *self-knowledge distillation*, and (3) *curriculum learning*. Specifically, in *parallel dialogue mining*, we automatically extract pseudo parallel dialogues; in *self-knowledge distillation*, we distillate the knowledge selection and knowledge expression ability from an auxiliary language to the target language; and in the *curriculum learning* part, we incorporate curriculum learning [3] into the distillation process to handle the noise caused by automatic parallel dialogue mining.

To tackle the second challenge, we collect a CKGC test dataset including about 3,000 conversations in three languages. As the first benchmark on the CKGC task, our work helps to facilitate relevant research in the future.

Using our newly created dataset, we conduct extensive experiments to assess the effectiveness of our proposed CKGC method. Evaluation results in terms of both automatic metrics and human evaluation indicate that CSKD can significantly outperform all baselines in cross-lingual knowledge selection without using any parallel corpus or human annotation. Moreover, we find that CSKD increases the topic richness of responses in the target language by leveraging knowledge in an auxiliary language.

In summary, this paper makes the following contributions:

- We are the first to propose the cross-lingual knowledge grounded conversation (CKGC) task.
- We propose a CSKD scheme, which learns to search and represent cross-lingual knowledge without annotations.
- We propose the first cross-lingual knowledge grounded conversational dataset to facilitate research on CKGC.
- We conduct extensive experiments on three languages to empirically validate the effectiveness of the proposed methods.

2 RELATED WORK

We discuss related work on knowledge-grounded conversations, cross-lingual information retrieval, and knowledge distillation.

2.1 Knowledge-grounded conversation

Unlike existing task-oriented dialogue generation approaches [19, 22, 23], open-domain dialogue systems focus on providing natural-sounding replies automatically to interact with humans on various domains [4]. In recent years, a variety of knowledge-grounded conversation (KGC) approaches have been proposed to improve the informativeness of open-domain dialogues [32]. Existing KGC methods can be categorized into two groups: *structured*-KGC and *unstructured*-KGC. The former conditions response generation on knowledge triples [29, 52, 58], whereas the latter conditions on free text [39]. Recently, many methods in the second group focus on leveraging *document-based unstructured knowledge* (e.g., Wikipedia articles) to enhance KGC [31, 33–35].

Importantly, existing KGC studies focus on grounding conversations with monolingual knowledge, which is difficult for languages with limited knowledge resources. Unlike previous work, the task we propose aims to leverage auxiliary knowledge sources to alleviate the problem of limited knowledge sources. We release a new dataset for KGC in a cross-lingual scenario.

Because annotations in KGC are expensive, recent studies explore unsupervised learning, without human knowledge annotations. Specifically, Lian et al. [26] devise a posterior knowledge

selection (PostKS) model to reparameterize the non-differentiable knowledge sampling process. Zhao et al. [56] propose KnowledGPT, which integrates a pre-trained language model with reinforcement learning. ZRKGC applies generalized EM to optimize two latent variables (for knowledge selection and knowledge expression) in KGC [25]. Unlike previous methods, our proposed CSKD is the first to tackle the KGC challenge in a cross-lingual scenario.

2.2 Cross-lingual information retrieval

The task of cross-lingual information retrieval (CLIR) aims to retrieve documents based on queries written in different languages [36]. Traditional CLIR systems transform the cross-lingual problem into a monolingual problem by translating queries or documents [37, 38, 57]. To address the translation ambiguity problem, embedding-based alignment approaches have successfully been applied to CLIR [28, 51]. Recently, multi-lingual pre-trained language models have been proposed to extract language-agnostic representations [8, 9, 12, 30]. Jiang et al. [18] apply multi-lingual language models to learn the relevance for English queries of foreign-language documents for CLIR.

Unlike previous work, in CKGC the query is a conversational context and the retrieved content is subsequently automatically rewritten in the form of a conversational response.

2.3 Knowledge distillation

Knowledge distillation (KD) [17] aims to transfer knowledge defined as soft output distributions from a teacher model to a student model. It has successfully been applied in numerous NLP tasks [5, 45, 54] Recent research applies KD in a cross-lingual scenario to bridge the language gap. Xu and Yang [53] use soft labels to supervise the learning of a low-resource language classifier with a parallel corpus. Duan et al. [11] introduce a hybrid distillation strategy in the summarization task. Inspired by knowledge distillation methods in model compression, self knowledge distillation (SKD) distills self-knowledge from a current model in the training process [16]. Sun et al. [46] propose a SKD objective during back-translation on an unsupervised neural machine translation task.

Our work differs from previous work in the following important ways: (1) no previous study focuses on the CKGC task; and (2) we propose curriculum self-knowledge distillation by incorporating curriculum learning to remove noisy signals.

3 PROBLEM FORMULATION

Suppose we have a dialogue corpus in target language $\mathcal{D}_T = \{(C_i^T, R_i^T)\}_{i=1}^{|\mathcal{D}_T|}$ with $|\mathcal{D}_T|$ context-response pairs, where for all i, C_i^T refers to a dialogue context with response R_i^T. Also, we have a knowledge corpus $\mathcal{K} = \{K_s\}_{s=1}^{|\mathcal{K}|}$ with $|\mathcal{K}|$ pieces of knowledge (i.e., sentences in the corpus) in an auxiliary language A. In addition to $\mathcal{D}_T$, we further assume that there is a large-scale dialogue corpus in the auxiliary language, i.e., $\mathcal{D}_A = \{(C_j^A, R_j^A)\}$.

As shown in Figure 3, we devise a sequential two-phase framework for CKGC. It includes a *cross-lingual knowledge retrieval* (CKR) layer and a *multilingual response generation* (MRG) layer. In CKR, we estimate $P(K|C, \mathcal{K})$ to retrieve a piece of knowledge from $\mathcal{K}$ given dialogue context C. In MRG, we learn $P(R|C, K)$ to generate a response R conditioned on both context C and knowledge K drawn from $P(K|C, \mathcal{K})$. Next, we introduce the details of these two layers.

Cross-lingual knowledge retrieval. Given a conversation (C^T, R^T) in language T, we retrieve a piece of knowledge from $\mathcal{K}$ in a two-step paradigm by following [10]. In the first step, we retrieve a knowledge pool $KP(C^T) = \{\dot{K}_s\}_{s=1}^{|KP(C^T)|}$, which contains many pieces of knowledge relevant to the conversation. In the second step, given context C, we select a piece of knowledge K from the knowledge pool by optimizing $P(K|C^T, KP(C^T))$. We utilize the response R to construct a more relevant knowledge pool $KP(R^T)$ during training.

Multilingual response generation. Given a conversational context C^T as well as a piece of knowledge K in an auxiliary language retrieved by the CKR layer, we aim to estimate $P(R^T|C^T, K)$ for generating response R^T. The MRG layer generates the response token by token. Thus, we define $P(R^T|C^T, K)$ as follows:

$$P(R^T|C^T, K) = \prod_{t=1}^{|R^T|} P(R_t^T|R_{1:t-1}^T, C^T, K). \quad (1)$$

4 METHOD

We first introduce the neural parameterization method of two layers in Section 4.1. Then in Section 4.2, we introduce the pre-training method on an auxiliary language. After that, we detail CSKD in Section 4.3, which learns $P(K|C^T)$ and $P(R^T|C^T, K)$ in an unsupervised manner.

4.1 Neural parameterization

Cross-lingual knowledge retrieval layer. Given a conversation $(C^{T/A}, R^{T/A})$ in language T or A, and a large knowledge corpus $\mathcal{K}$, we first use a transformer encoder with parameters θ to encode the context sentence C, so we have:

$$\mathbf{h}^{C^{T/A}} = \text{L2Norm}(\text{AvgPooling}(\text{Encoder}(C^{T/A}, \theta))) \in \mathbb{R}^{1\times d}, \quad (2)$$

where $\mathbf{h}^{C^{T/A}} \in \mathbb{R}^{1\times d}$ refers to the sentence representation with hidden size d, L2Norm indicates a L2-normalization operation, and AvgPooling means average-pooling. Likewise, we encode all the knowledge sentences in $\mathcal{K}$ to $\mathbf{h}^{\mathcal{K}} \in \mathbb{R}^{|\mathcal{K}|\times d}$ using the same method. Then, we construct the knowledge pool $KP(C^{T/A})$ using the nearest neighbors of $C^{T/A}$:

$$KP(C^{T/A}) = \text{KNN}(\mathbf{h}^{C^{T/A}}, \mathbf{h}^{\mathcal{K}}), \quad (3)$$

where KNN denotes a K Nearest Neighbors function using cosine similarity. We use FAISS [20] to simultaneously search neighborhoods for all dialogues in an efficient manner.

Given the knowledge pool $KP(C^{T/A})$, we use an attention mechanism to perform a fine-grained selection of which knowledge sentences are used to generate the response:

$$P(K|C^{T/A}, KP(C^{T/A})) = \frac{\exp(\mathbf{h}^{C^{T/A}} \cdot \mathbf{h}^{K^\top})}{\sum_{\dot{K}\in KP(C^{T/A})} \exp(\mathbf{h}^{C^{T/A}} \cdot \mathbf{h}^{\dot{K}^\top})}. \quad (4)$$

When training the model, we build the knowledge pool $KP(R^{T/A})$ by replacing $C^{T/A}$ with response $R^{T/A}$.

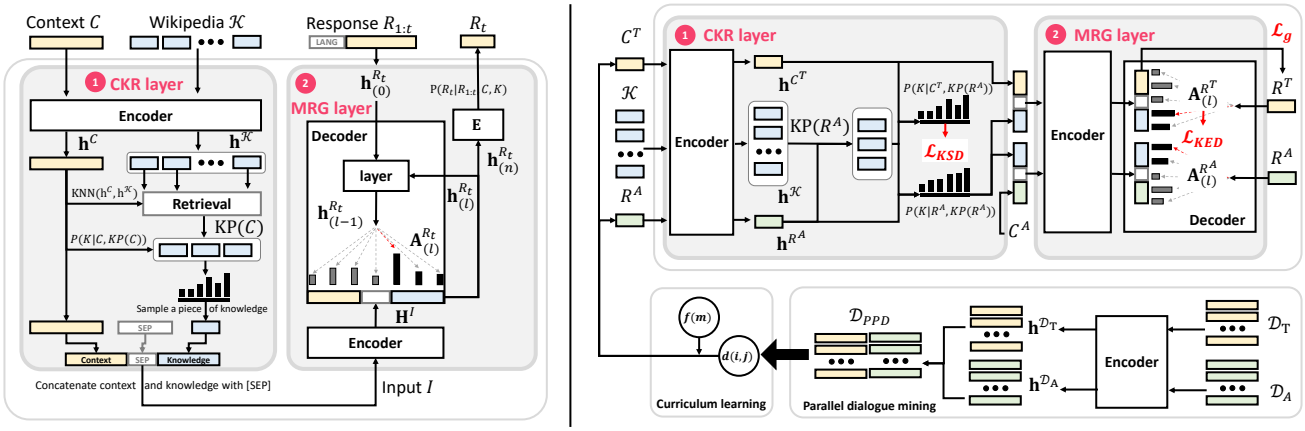

Figure 3: An overview of CKGC framework and CSKD scheme.

Multilingual response generation layer. Given multilingual inputs, we employ a transformer encoder-decoder network with parameters ϕ in the MRG layer to generate the response. Specifically, we first concatenate $C^{T/A}$ and the selected knowledge K obtained from the CKR layer with a $[SEP]$ token to get the input, i.e., $I = \{K; [SEP]; C^{T/A}\}$. Then we encode I into a latent representation $\mathbf{H}^I \in \mathbb{R}^{|I| \times d}$, where $|I|$ denotes the number of tokens in I and d denotes the hidden size. In the decoding stage, we decode to generate the response token by token with a start language identification $[LANG]$. Concretely, the probability of generating token R_t from a predefined vocabulary V at the timestamp t is defined as:

$$P(R_t^{T/A}|R_{1:t}^{T/A}, C, K) = \text{Softmax}(\mathbf{h}_{(n)}^{R_t^{T/A}} \cdot \mathbf{E}^\top) \in \mathbb{R}^{1\times|V|}, \quad (5)$$

where $\mathbf{E} \in \mathbb{R}^{|V|\times d}$ denotes an embedding matrix of vocabulary V; $\mathbf{h}_{(n)}^{R_t^{T/A}} \in \mathbb{R}^d$ denotes d-dimension word representations in R_t for a decoder with n layers. Thus, for the l-th layer in the decoder, we have:

$$\begin{aligned}
\mathbf{h}_{(l)}^{R_t^{T/A}} &= \text{FFN}(\text{LN}(\mathbf{A}_{(l)}^{R_t^{T/A}} \cdot \mathbf{H}^{I\top} + \mathbf{s}_{(l)}^{R_t^{T/A}})) \in \mathbb{R}^{1\times d}, \\
\mathbf{A}_{(l)}^{R_t^{T/A}} &= \text{Softmax}(\mathbf{s}_{(l)}^{R_t^{T/A}} \cdot \mathbf{H}^{I\top}) \in \mathbb{R}^{1\times|I|}, \\
\mathbf{s}_{(l)}^{R_t^{T/A}} &= \text{LN}(\text{SA}(\mathbf{h}_{(l-1)}^{R_t^{T/A}}, \mathbf{H}_{(l-1)}^{R_{1:t}^{T/A}\top}) + \mathbf{h}_{(l-1)}^{R_t^{T/A}}) \in \mathbb{R}^{1\times d},
\end{aligned} \quad (6)$$

where $\mathbf{A}^{R_t^{T/A},(l)}$ denotes the attention distribution at the l-th layer, which represents the encoder inputs during the generation process; LN denotes the layer normalization operation, whereas FFN is a position-wise fully connected feed-forward network with GeLU non-linear activation; $\mathbf{H}_{(l-1)}^{R_{1:t}^{T/A}} = \{\mathbf{h}_{(l-1)}^{R_\tau^{T/A}}\}_{\tau=0}^{t} \in \mathbb{R}^{t\times d}$ denotes a representation matrix of $R_{1:t}$ at the l-th layer. Following [49], we write SA for the self attention block.

4.2 Pretraining on an auxiliary language

In this stage, we aim to pre-train the network in auxiliary language given dialogues $\mathcal{D}_A$ and a knowledge corpus $\mathcal{K}$. Inspired by Zhao et al. [56], we pre-train the network auxiliary language use a two-stage paradigm: warm-up stage and reinforcement learning stage.

Algorithm 1 CSKD training

Input: dialogue corpus $\mathcal{D}^A$ in auxiliary language, dialogue corpus $\mathcal{D}^T$ in target language, knowledge corpus $\mathcal{K}$, pre-trained mBART, maximum step M
Initialize parameters θ of CKR and ϕ of MRG with mBART
Pre-train θ and ϕ on auxiliary language ▷ See §4.2
Construct pseudo parallel dialogue $\mathcal{D}_{PPD}$. ▷ See §4.3.1
Sort $\mathcal{D}_{PPD}$ according to Eq. 16
for training step m = 1, ..., M **do**
    Sample a batch B_m in $\mathcal{D}_{PPD}$ based Eq. 17;
    Sample $\{(C^T, R^T), (C^A, R^A)\}$ from B_m;
    Update θ based on Eq. 13;
    Update ϕ based on Eq. 14 and Eq. 15;
end for

4.2.1 Warm-up stage. We first warm up the network by collecting pseudo ground truth [56]. Specifically, we construct the pseudo knowledge pool $KP(\hat{R^A})$ and pseudo ground truth knowledge $\hat{K}$ by BM25 using responses as queries. With the $KP(\hat{R^A})$ and $\hat{K}$, CKR are optimized via the maximum likelihood estimation:

$$\mathcal{L}_{CKR} = \sum_{(C^A,R^A)\in\mathcal{D}_A} -\log(P(\hat{K}|R^A, KP(\hat{R^A}))), \quad (7)$$

and MLG layers are optimized via NLL loss:

$$\mathcal{L}_{MLG} = \sum_{(C,R)\in\mathcal{D}_A} -\log(P(R^A|C^A, \hat{K})), \quad (8)$$

4.2.2 Reinforcement learning stage. After the warm-up stage, the CKR layer is further improved via the policy gradient [47] with a reward function from the MRG layer. We draw knowledge K from $P(K|C^A, KP(\hat{R^A}))$ and define the reinforcement learning objective as follows:

$$\mathcal{L}_{RL} = -\sum_{(C^A,R^A)\in\mathcal{D}_A} \mathbb{E}_{K\sim P(K|\cdot)}\, \tilde{\mu}_{(R^A,C^A,K)} \cdot \log(P(K|\cdot)), \quad (9)$$

where

$$\begin{aligned}\tilde{\mu}_{(R^A,C^A,K)} &= \mu_{(R^A,C^A,K)} - b \\ \mu_{(R^A,C^A,K)} &= P(R^A|C^A,K)^{\varepsilon/|R^A|}.\end{aligned} \tag{10}$$

In Eq. 9 and 10, we write $P(K|\cdot)$ as a shorthand for $P(K|C^A, K\hat{P}(R))$, ε is the temperature, whereas $|R^A|$ is the length of the response. Following [7], we write b for the baseline to reduce the variance of gradient estimation, so we have:

$$b = \frac{1}{|K\hat{P}(R^A)|} \sum_{\dot{K} \in K\hat{P}(R^A)} \mu_{(R^A,C^A,\dot{K})}, \tag{11}$$

where $\mu_{(R^A,C^A,K)}$ is the reward function given context C^A, knowledge K, and the target response R^A.

4.3 Curriculum self-knowledge distillation

After pretraining, our method is able to select and express the knowledge in an auxiliary language. However, the cross-lingual knowledge selection and expression is still limited. By assuming that a parallel dialogue has similar knowledge expression, we propose a CSKD scheme in order to distillate knowledge from an auxiliary language to a target language. The learning algorithm of CSKD is summarized in Algorithm 1. CSKD is composed of three ingredients: (1) *parallel dialogue mining*: we extract pseudo parallel dialogues automatically; (2) *distillation on knowledge selection and expression*: we distillate the knowledge selection and expression abilities from auxiliary language to target language; and (3) *curriculum learning*: we incorporate the curriculum learning into the distillation to handle the noise in the parallel dialogues mining stage.

4.3.1 Parallel dialogue mining. In this part, we extract pseudo parallel dialogues and construct $\mathcal{D}_{PPD} = \{(C_i^T, R_i^T), (C_j^A, R_j^A)\}_{i=1}^{|\mathcal{D}_T|}$, where $\forall i$, (C_j^A, R_j^A) is the pseudo parallel dialogue of (C_i^T, R_i^T).

To mine parallel dialogues without supervised signals, we first encode each response R_i^T in $\mathcal{D}^T$ into a representation $\mathbf{h}^{R_i^T}$ via Eq. 2, and each R_j^A in the auxiliary language is encoded into $\mathbf{h}^{R_j^A}$ in the same way. Let $\mathbf{h}^{\mathcal{D}_T} = \{\mathbf{h}^{R_i^T}\}_{i=1}^{|\mathcal{D}_T|} \in \mathbb{R}^{|\mathcal{D}_T| \times d}$ and $\mathbf{h}^{\mathcal{D}_A} = \{\mathbf{h}^{R_i^A}\}_{i=1}^{|\mathcal{D}_A|} \in \mathbb{R}^{|\mathcal{D}_A| \times d}$ be the encoded dialogue corpus in target and auxiliary languages, respectively.

Then, we use the same KNN function in Eq. 3 to find N nearest neighbors of each dialogue in the target language. Therefore, for the i-th dialogue in the target language (C_i^T, R_i^T), we have $\mathrm{NN}(R_i^T) = \mathrm{KNN}(\mathbf{h}^{R_i^T}, \mathbf{h}^{\mathcal{D}_A})$. Given a pair of dialogues in the target and auxiliary languages, i.e., (C_i^T, R_i^T) and (C_j^A, R_j^A), we calculate their semantic similarity by a ratio margin function [2]:

$$\mathrm{S}(i,j) = \cos(\mathbf{h}^{R_i^T}, \mathbf{h}^{R_j^A}) / \left[\sum_{\mathbf{z} \in \mathrm{NN}(R_i^T)} \cos(\mathbf{h}^{R_i^T}, \mathbf{z}) + \sum_{\mathbf{z} \in \mathrm{NN}(R_j^A)} \cos(\mathbf{z}, \mathbf{h}^{R_j^A}) \right], \tag{12}$$

where $\mathbf{z} \in \mathbb{R}^{1\times d}$, $\mathbf{z} \in \mathrm{NN}(R_{i/j}^{T/A})$ denotes the representation of a dialogue in $\mathrm{NN}(R_{i/j}^{T/A})$. Next, we construct the pseudo parallel dialogue for each $(C_i^T, R_i^T) \in \mathcal{D}^T$ using the dialogue (C_j^A, R_j^A) with the highest $\mathrm{S}(i,j)$.

4.3.2 Distillation on knowledge selection and expression. We propose two distillation objectives, knowledge selection distillation (KSD) and knowledge expression distillation (KED), in order to distill the knowledge selection and expression abilities from the auxiliary language to target language. Specifically, given parallel dialogues $\{(C^T, R^T), (C^A, R^A)\}$, we define the objective function of KSD as:

$$\mathcal{L}_{KSD} = \sum_{K \in KP} P(K|R^A, KP) \log \left(\frac{P(K|R^A, KP)}{P(K|C^T, KP)} \right), \tag{13}$$

where the knowledge pool KP denotes $KP(R^A)$ constructed by the response in the auxiliary language. The objective function of KED is defined as:

$$\begin{aligned}\mathcal{L}_{KED} &= \sum_{l=1}^{n} 1/n \ \mathrm{KL}(\mathrm{AP}(\mathbf{A}_{(l)}^{R^A}), \mathrm{AP}(\mathbf{A}_{(l)}^{R^T})), \\ \mathbf{A}_{(l)}^{R} &= \{\mathbf{A}_{(l)}^{R_\tau}\}_{\tau=1}^{|R|} \in \mathbb{R}^{|R|\times|I|},\end{aligned} \tag{14}$$

where n is the number of layers in decoder; I is the concatenation of context and knowledge sequences; KL denotes the Kullback-Leibler divergence; $A_{(l)}^R \in \mathbb{R}^{|R|\times|I|}$ denotes the attention matrix of response R attending to encoder input I in the l-th decoder layer; we apply an average pooling operation AP on the attention matrix and obtain $\mathrm{AP}(\mathbf{A}_{(l)}^R) \in \mathbb{R}^{1\times|I|}$, where we mask the context part in $\mathrm{AP}(\mathbf{A}_{(l)}^R)$ to ensure that the two input variables of $\mathrm{KL}(\cdot,\cdot)$ have the same size. The masked $\mathrm{AP}(\mathbf{A}_{(l)}^R)$ indicates the average intensity of attention of each knowledge token during the response generation process. Besides, we optimize the mixed-language-aware response generation (MLG) via an additional NLL loss:

$$\mathcal{L}_g = -\log(P(R^T|C^T, K)), \tag{15}$$

where $K \sim P(K|R^A, KP(R^A))$ is a piece of knowledge chosen by R^A.

4.3.3 Curriculum learning. In order to reduce the impact of noise caused in parallel dialogues mining, we introduce curriculum learning [3] into the distillation process. We design a curriculum training scheduler to provide the model with easy samples first, then gradually increase the difficulty of samples. The curriculum training scheduler is arranged by sorting each pseudo parallel dialogue according to the difficulty defined as follows:

$$d(i,j) = \mathrm{S}(i,j) P(K_{best}|R_j^A, KP(R_j^A)), \tag{16}$$

where the lower $d(i,j)$ is, the more difficult the pseudo parallel dialogue is; $\{(C_i^T, R_i^T), (C_j^A, R_j^A)\}$ is a pseudo parallel dialogue from $\mathcal{D}_{PPD}$; the function $\mathrm{S}(\cdot,\cdot)$ is defined in Eq. 12; the function $P(\cdot|\cdot,\cdot)$ is defined in Eq. 4; and K_{best} is the knowledge chosen from $KP(R_j^A)$ by R_j^A.

During training, the model will first train on data with lower difficulty according to the training scheduler, and gradually increase the proportion of difficult data until all the data is used. At training step m, a batch of training samples is obtained from the top $\mathrm{f}(m)$

part of the entire sorted training samples. Following [40], we define the function $f(m)$ as:

$$f(m) \triangleq \min\left(1, \sqrt{\frac{m(1-\alpha_0^2)}{M} + \alpha_0^2}\right). \quad (17)$$

where α_0^2 denotes the percentage of data used in the initial training stage, M is the maximum step.

5 DATASET

As far as we know, existing KGC datasets only focus on grounding conversations in a monolingual knowledge corpus. Thus, we annotate a cross-lingual knowledge grounded conversation (CKGC) test dataset. Following [10], we consider a one-to-one conversation scenario in CKGC, and only one participant (i.e., the *wizard*) has access to an information retrieval system that shows the worker paragraphs from Wikipedia possibly relevant to the conversation, while the other is a curious learner (the *apprentice*).

Before the start of the conversation, two participants engage in chitchat and will be randomly assigned the role of *wizard* or *apprentice*; the *apprentice* chooses the topic of conversation. Then, the two participants chat one by one, while the *wizard* can access knowledge that is unobservable to the *apprentice*. The conversation repeats until one of the conversation partners ends the chat.

Topic selection. Dinan et al. [10] crowd-sourced a set of natural, open-domain dialogue topics (each linked to a Wikipedia article). Thus, we use topics that have appeared in an unseen test set in Wizard of Wikipedia [10] as topic sets in our dataset. Before the conversation, we show several topics randomly selected from the topic sets to the *apprentice*. The *apprentice* then chooses a topic of interest as the start topic of the conversation. During the conversation, the topic is allowed to naturally change.

Knowledge retrieval. During the conversation, the *wizard* has access to a set of passages of knowledge that may be relevant to the given dialogue context. There are two types of knowledge shown to the *wizard*, one is about the original topic, and the other is the knowledge updated in real time as the conversation progresses. For the first type, since each topic is linked to a Wikipedia article, we use the first 10 sentences in this article, which is usually the summary of the article. For the second type, we retrieve the top 7 articles (first paragraph only) for the last two turns of dialogue to adapt to the topic transition in conversation. Specifically, we first translate each utterance in the conversation context using Google translation. Then, we retrieve the articles via Apache Solr, an open source enterprise search platform built on Apache Lucene (basically a BM25 model [43]). We sort the paragraphs based on the unigram F1 [10] correlation of the dialogue context and each paragraph.

Quality assurance. We hired 6 experienced experts to score each conversation collected in the previous step. We ask experts to evaluate three aspects of the data, including knowledge relevance (whether the selected knowledge is relevant to the context), correctness of knowledge representation (whether the wizard understands and uses the knowledge correctly), and dialogue coherence (whether the two parties are engaged in the dialogue), and assign a score in $\{0, 1, 2\}$ (representing "bad," "fair," and "good"). All data is evaluated by two experts repeatedly to eliminate bias. We deleted all data scored as "bad" in any of the three aspects.

Table 1: Statistics of the cross-lingual knowledge grounded conversation (CKGC) dataset.

CKGC	Chinese	French	Spanish
#Utterances	8,271	7,858	7,410
#Dialogues	1,000	950	940
#Topics	399	360	354
Average turns per dialogue	4.1	4.0	4.0
Knowledge database	6.2M articles, 106M sentences		

Table 2: Statistics of the training data.

Language	English	Chinese	French	Spanish
Crawled from	Reddit	Tieba	Reddit	Reddit
#Dialogues	20,308,634	3,059,173	1,026,462	2,012,992
Avg. #words	14.1	15.7	49.3	37.7
Wikipedia	#Sentences: 106,824,651, Avg. #words: 21.1			

The final dialogue dataset we collect consists of 1,000 dialogues in Chinese, 950 dialogues in French and 940 dialogues in Spanish. Overall data statistics can be found in Table 1. Here, we define a continuous pair of (Apprentice-utterance, Wizard-utterance) as a "turn".

6 EXPERIMENTS

6.1 Research questions

We aim to answer the following research questions with our experiments: (RQ1) How does our proposed method, CSKD, perform on CKGC? Can CSKD help to select and express knowledge in an auxiliary language? (See §7.1 and §7.2) (RQ2) What is the effect of each ingredient in CSKD? (See §7.3) (RQ3) Can knowledge in an auxiliary language enrich the conversation topics? If so, what is the effect of these topics? (See §7.4) In addition, we show a number of cases to demonstrate the performance of CSKD; see §7.5.

6.2 Training data

We establish the knowledge corpus using a Wikipedia dump,[2] where text is extracted[3] and split into sentences using NLTK.[4] In total, there are 6,167,445 articles and 106,824,651 sentences. The English dialogue corpus is constructed from the Reddit Conversation Corpus, which contains 20,308,634 conversations. The French dialogue corpus[5] is collected from Reddit, including 1,026,462 conversations. The Chinese dialogue corpus is constructed from an online communication platform, Tieba,[6] with 3,059,173 conversations. To establish a Spanish dialogue corpus, we collect comments from Reddit during 2019. We first download all the comments submitted in 2019 from the pushshift;[7] then we identify the language of each comment using a classifier trained by fastText [15, 21]; finally, we match comments based on their parent_id to construct the context. Our Spanish Reddit dialogue corpus contains 2,012,992 conversations. Summary statics of our training datasets are given in Table 2.

[2] https://dumps.wikimedia.org/enwiki/latest/enwiki-latest-pages-articles.xml.bz2
[3] https://github.com/attardi/wikiextractor/wiki
[4] https://www.nltk.org/
[5] https://www.kaggle.com/breandan/french-reddit-discussion
[6] https://github.com/codemayq/chinese_chatbot_corpus
[7] https://files.pushshift.io/reddit/

6.3 Baselines

To verify the effectiveness of CSKD, we compare it with the following models:[8]

- **DialoGPT** [41] is a dialogue generation model that attains human-close performance in evaluation. The model follows the architecture of OpenAI GPT-2. In our work, we use checkpoints trained in Chinese,[9] French,[10] and Spanish.[11]
- **mBART** [30] is a multilingual model pretrained by denoising auto-encoder objective on the common-crawl-25 corpus; the mbart.cc25 checkpoint was used and we finetune the model on the dialogue corpus shown in Table 2.
- **Pipeline** [6] translates input context sentences in a target language into English. Then it performs an English KGC model to produce the response, and finally translates it back to the target language. We use the KGC model trained in §4.2 and a transformer based model trained on OPUS data [48] for translation.
- **RLKS** [56] supervises KS according to the pseudo ground-truth labels they construct and the signals gained in a reinforcement learning way; we use the response translated by the model trained on OPUS data [48] as the query, and employ BM25 to retrieve the pseudo knowledge pool.
- **XNLG** [6] fine tunes a pre-trained language model on English dialogue data first, and directly performs inference on non-English test data in a zero-shot setting. We adopted the (Fine-Tuning for Any-to-Others NLG) suggestion in [6], that keeps the decoder and the word embeddings frozen and only updates the encoder parameters during fine-tuning, to avoid catastrophic forgetting of target language controllability.

6.4 Evaluation metrics

We use Recall@1 as automatic metric to evaluate the cross-lingual knowledge retrieval task. To evaluate response generation, we choose unigram F1[12] and ROUGE-1/2 [27] as automatic metrics. We also assess the performance of all methods in terms of human annotations by following [25]. We randomly sample 300 dialogues and their corresponding generations from our model as well as the baselines. We recruit 6 experienced translation experts as annotators. Specifically, given the conversation context, the knowledge pool used at the current turn, the selected knowledge, as well as the generated responses, each expert needs to give a preference (i.e., "bad", "fair", and "good") in terms of three aspects: *fluency*, *context coherence*, and *knowledge relevance*. Fluency measures if the generated response is smooth; context coherence measures if the generated response and dialogue context are coherent; and knowledge relevance measures if the selected knowledge and dialogue context are relevant. Each response receives 3 scores per aspect, and agreement among the annotators is measured via Fleiss' kappa [13].

6.5 Implementation details

In the inference phase, our model uses the knowledge pool in the dataset (see §5 for the construction method) just like the baseline. We use FAISS [20] to accelerate the dense vectors index process. We use mbart.cc25 checkpoint [30] (680M parameters)[13] as the backbone network of our model and all the baselines. The model is trained on 4 NVIDIA TITAN RTX GPUs, with a *batch size* = 64. We set the maximum input length as 128, and a maximum output length as 64. We set the knowledge pool size as 10. All models are optimized using the AdamW optimizer with $lr = 2e-5$, $\beta 1 = 0.9$, and $\beta 2 = 0.999$. During decoding, we use beam search algorithm and set beam size = 3.

7 EXPERIMENTAL RESULTS

7.1 Automatic evaluation (RQ1)

Table 3 shows the evaluation results on the automatic metrics. Generally, CSKD achieves the best performance in terms of all metrics for all 3 languages. Based on the results, we have three main observations.

First, CSKD significantly outperforms the open-domain dialogue baseline that neglects external knowledge during the response generation. This shows that CSKD can effectively increase the informativeness of generated responses by introducing foreign knowledge.

Second, we compare CSKD with translation-based methods. And we find that even though no parallel corpus is used, CSKD significantly surpasses the baseline method in cross-lingual retrieval tasks. This shows that our end-to-end CKGC solution can solve the ambiguity problem which frequently occurs during machine translation. Moreover, we find that CSKD outperforms the pipeline method in terms of the quality of response generation.

Third, CSKD significantly outperforms RLKS, which only uses the target language dialogue, and XNLG, which only uses an auxiliary dialogue. In contrast with these monolingual methods, CSKD makes full use of the dialogue data of the two languages. This shows the effectiveness of the usage of knowledge distillation for extending the ability to select and express cross-lingual knowledge.

7.2 Human evaluation (RQ1)

We conduct human evaluations to confirm the improvements of CSKD. Table 4 shows the human evaluation results. Overall, CSKD achieves the best performance in terms of all metrics on three languages. CSKD outperforms baseline methods in terms of the Fluency and Coherence metrics, with significant advantages according to the KG Relevance metric. The Kappa value confirms that the increase in performance is unanimously agreed on by experts.

7.3 Ablation studies (RQ2)

To analyze the effect of each component in CSKD, we conduct an ablation study. Table 5 shows the results on the three languages, where we consider four settings: (1) No KED in §4.3.2 (-KED in Table 5), i.e., we remove the knowledge expression distillation objective; (2) No KSD in §4.3.2 (-KSD in Table 5), i.e., we remove the knowledge selection distillation objective; (3) No curriculum learning (CL) in §4.3.3 (-CL in Table 5), i.e., we remove the curriculum learning, train model directly on the whole data; and (4) No KED,

[8] For a fair comparison, we replaced all baseline backbone networks with mBART except the DialoGPT.
[9] https://github.com/yangjianxin1/GPT2-chitchat
[10] https://huggingface.co/antoiloui/belgpt2
[11] https://huggingface.co/datificate/gpt2-small-spanish
[12] https://github.com/facebookresearch/ParlAI/blob/master/parlai/core/metrics.py
[13] https://github.com/pytorch/fairseq/tree/master/examples/mbart

Table 3: Automatic evaluation results on the CLKGC. Bold face indicates the best result in terms of the corresponding metric, significant improvements over the best baseline are marked with * (t-test, $p < 0.05$).

Methods	Chinese				French				Spanish			
	R@1	F1	ROUGE-1	ROUGE-2	R@1	F1	ROUGE-1	ROUGE-2	R@1	F1	ROUGE-1	ROUGE-2
DialoGPT[41]	-	8.30	10.22	1.22	-	7.63	7.50	0.76	-	11.13	11.58	1.32
mBART[30]	-	11.18	14.92	4.33	-	11.37	9.98	1.45	-	11.11	11.56	1.61
Pipeline	6.92	15.12	18.78	4.93	6.90	13.16	11.13	1.98	6.82	14.23	13.24	2.39
RLKS[56]	8.76	14.49	18.25	4.41	8.82	13.05	10.89	1.92	8.67	13.95	13.31	2.20
XNLG[6]	7.07	5.33	12.26	3.06	6.59	5.27	4.64	0.63	6.42	5.46	5.50	0.81
CSKD (ours)	**9.29***	**15.34***	**19.21***	**5.27***	**9.39***	**13.58***	**11.56***	**2.11***	**9.27***	**14.62***	**13.76***	**2.58***

Table 4: Human evaluation results.

Methods	Chinese				French				Spanish			
	Fluency	Coherence	Relevance	Kappa	Fluency	Coherence	Relevance	Kappa	Fluency	Coherence	Relevance	Kappa
mBART	1.65	0.91	0.76	0.73	1.66	1.13	0.89	0.71	1.52	1.24	0.84	0.62
Pipeline	1.78	1.34	1.17	0.73	1.71	1.36	1.15	0.64	1.60	1.45	1.10	0.57
CSKD	1.79	**1.57**	**1.39**	0.61	1.74	**1.53**	**1.38**	0.65	1.65	**1.55**	**1.39**	0.60
Human	1.98	1.89	1.87	0.56	1.97	1.87	1.91	0.61	1.96	1.80	1.79	0.63

KSD and CL (-KED-KSD-CL in Table 5), i,e., we remove all ingredients in CSKD and only use the knowledge selected in auxiliary language as pseudo ground truth to train the MRG layer.

The results show that all components are helpful for CSKD because removing any of them will decrease the results. Without KED, the model only optimizes the MRG layer through NLL loss defined in Eq. 15 during training. It can be seen from the experimental results that although the model is still highly accurate in knowledge selection, the model is significantly weaker than CSKD in knowledge expression. Concretely, this shows that KED can effectively supervise model learning how to express a piece of auxiliary language knowledge in target language.

Without KSD, we find that the model faces a huge performance degradation in knowledge selection. Concretely, it drops 2.22%, 2.80% and 2.85% in terms of R@1 on Chinese, French and Spanish, respectively. As a result, although the model still uses the CSKD method in the generation part, the decline in knowledge selection ability directly leads to the low knowledge relevance of the generated responses, and poor generation performance.

Without CL, we find that although the distillation can still bring a certain improvement, the ability of the model is severely damaged due to the influence of noise during the parallel dialogue mining stage. Specifically, it drops 1.40%, 1.73% and 1.76% in terms of R@1 on three languages separately, indicating that the curriculum learning scheduler can effectively estimate the credibility of the data, and make full use of the data with noisy.

When removing all the components from CSKD, the model degenerates to only use NLL loss to train MRG, while not doing any finetuning on the CKR layer. The experimental results show that although the model still has a certain ability of knowledge expression, it can no longer generate satisfactory responses compared with CSKD.

7.4 Comparison with monolingual KGC (RQ3)

In order to verify whether the use of rich foreign language knowledge can increase the richness of dialogue knowledge, we further compare the performance of models using different language knowledge on our dataset. We use the French dialogue dataset and French Wikipedia to train the monolingual model using the method introduced in §4.2. Next, we retrieve knowledge using a standard IR system based on the dialogue context. We replace the *wizard* with the monolingual KGC model on French, and the performance on the test set is shown in Table 7.

The results of automatic evaluation show that the generation performance of the single-language model is significantly weaker than that of the CKGC method. Since there is no label of French knowledge, we cannot report the automatic metric of knowledge retrieval. In order to further analyze the weaknesses of the monolingual model, we hired three experts to manually evaluate the knowledge select of the model. We ask experts to score the relevance of the selected knowledge, divided into three levels, $\{0, 1, 2\}$. We analyze the scoring results of experts and find that about 11.45% of the data was scored as 0, which indicates that monolingual KGC has difficulties in knowledge retrieval.

7.5 Case studies

In Table 6 we show three examples in different languages from the test set to assess the performance of CSKD, Pipeline and human. We see that CSKD chooses more appropriate knowledge and hence generates accurate and engaging responses with the aid of cross-lingual knowledge. For instance, given the current user utterance in the Chinese example, both CSKD and Pipeline choose the right knowledge. But the Pipeline model incorrectly expresses knowledge during translation, i.e., it mistakes the knowledge “Colombia, is a country in the north of south America” for “Colombia is a country in North America”. In contrast, CSKD shows its advantages

Table 5: Ablation study on the CKGC. -KED denotes removing the knowledge expression distillation. -KSD denotes removing the knowledge selection distillation. -CL denotes removing the curriculum learning scheduler.

Methods	Chinese				French				Spanish			
	R@1	F1	ROUGE-1	ROUGE-2	R@1	F1	ROUGE-1	ROUGE-2	R@1	F1	ROUGE-1	ROUGE-2
CSKD	**9.29**	**15.34**	**19.21**	**5.27**	**9.39**	**13.58**	**11.56**	**2.11**	**9.27**	**14.62**	**13.76**	**2.58**
-KED	9.29	14.86	18.94	4.77	9.39	13.38	11.37	1.95	9.27	14.57	13.56	2.38
-KSD	7.07	13.96	17.62	4.48	6.59	12.73	10.68	1.67	6.42	13.14	12.70	2.01
-CL	7.89	14.53	18.36	4.51	7.66	12.94	10.65	1.80	7.51	13.84	13.13	2.29
-CL-KED-KSD	7.07	12.95	16.37	4.30	6.59	11.67	10.25	1.57	6.42	12.79	12.06	1.83

Table 6: Case study.

	Chinese (Topic: Colombia)	French (Topic: Canadian Football)	Spanish (Topic: Electric guitar)
Context	你知道哥伦比亚吗? (Do you know Colombia?)	Le sport est-il populaire? (Is sports popular?)	¿Podrías introducirme algo sobre este instrumento musical? como su origen y los métodos de tocar? (Can you introduce this instrument to me? As its origin and playing method?)
Knowledge	K_1: Colombia, officially the republic of colombia , is a country in the north of south America. K_2: colombia is one of the world's 17 megadiverse countries and has the second-highest level of biodiversity in the world.	K_1: Senior leagues such as the Alberta Football League have grown in popularity in recent years. K_2: The CFL is the most popular and only major professional Canadian football league.	K_1: Invented in 1932, the electric guitar was adopted by jazz guitar players, who wanted to play single-note guitar solos in large big band ensembles. K_2: The sound of an electric guitar can be modified by new playing techniques such as string bending, tapping, and hammering-on, using audio feedback, or slide guitar playing.
Pipeline	哥伦比亚是北美洲的一个国家。(Colombia is a country in North America.).	Le sport est populaire, tant en nombre de participants qu'en nombre de spectateurs. (The sport is very popular, both for the participants and the audience.).	En el año 2000, la banda lanzó su primer álbum de estudio, titulado "Spirit", el cual fue lanzado en el año 2001. El álbum fue grabado en los estudios. (In 2000, the band released their first studio album, spirit, which was released in 2001. The album was recorded in the studio.).
CSKD	哥伦比亚是南美的国家。(Colombia is a country in South America.).	Oui, la Ligue Senior a gagné en popularité ces dernières années. (Yes, in the past few years, the Premier League has become more and more popular.).	Un instrumento de cuerda es una guitarra que tradicionalmente está hecha de madera y es conocida por su tonalidad por sus cuerdas intestinales, de acero o de nylon. (Stringed instruments are guitars that are traditionally made of wood and are known for their liner, steel or nylon string tones.).
Human	哥伦比亚是位于南美洲北部的国家。(Colombia is a country in the north of South America.).	Oui, souvent chaque lycée ou collège a leur propre équipe. (Yes, usually every high school or college has its own team.).	La guerra electrónica fue inventada en 1932, primeramente fue adoptada or los jugadores de jazz ,que querían una sola nota. (Electric guitar was invented in 1932 and was initially adopted by jazz players.).

Table 7: Comparison with monolingual KGC.

Methods	French			
	R@1	F1	ROUGE-1	ROUGE-2
cross-lingual KGC	9.39	13.58	11.56	2.11
monolingual KGC	–	12.87	10.04	1.73

in knowledge expression because it avoids the accumulation of translation errors.

In French, the Pipeline model chooses the wrong knowledge, while CSKD chooses knowledge accurately. CSKD outperforms Pipeline for knowledge selection.

In Spanish, although neither CSKD nor Pipeline selects the right knowledge, the knowledge selected by CSKD appears to be more relevant to the dialogue context.

8 CONCLUSION

In this paper, we have focused on the knowledge-grounded conversation problem for languages with limited availability of knowledge sources. We have proposed the task of cross-lingual knowledge grounded conversations (CKGC) to alleviate the knowledge deficits. We have devised a 2-phase framework for CKGC. On this basis, we have proposed a curriculum self-knowledge distillation (CSKD) scheme to tackle challenges in CKGC. CSKD is composed of 3 ingredients: parallel dialogue mining, self-knowledge distillation, and curriculum learning. As there is no benchmark test dataset on our topic, we have collected a real-world CKGC dataset.

Using the CKGC dataset, we have run extensive experiments to verify the effectiveness of CSKD. For both automatic evaluation and human evaluation, CSKD outperforms all the baselines.

A limitation of our work is the limited improvement over translation-based baselines in terms of the quality of generated responses. As to future work, we would like to extend our method, CSKD, to combine with knowledge in multiple languages to enhance the performance. Semi-supervised learning methods also provide insights for cross-lingual knowledge grounded conversations.

REPRODUCIBILITY

To facilitate reproducibility of our results, we are sharing our code and dataset at https://github.com/sunnweiwei/ckgc.

ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China with grant No. 2020YFB1406704, the Natural Science Foundation of China (61902219, 61972234, 62072279), the Key Scientific and Technological Innovation Program of Shandong Province (2019JZZY010129), the Tencent WeChat Rhino-Bird Focused Research Program (JR-WXG-2021411), the Fundamental Research Funds of Shandong University, the Hybrid Intelligence Center, a 10-year program funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl.

All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

REFERENCES

[1] Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a Human-like Open-Domain Chatbot. *arXiv preprint arXiv:2001.09977* (2020).

[2] Mikel Artetxe and Holger Schwenk. 2019. Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings. In *ACL*.

[3] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum Learning. In *ICML*.

[4] Hongshen Chen, Zhaochun Ren, Jiliang Tang, Yihong Eric Zhao, and Dawei Yin. 2018. Hierarchical Variational Memory Network for Dialogue Generation. In *WWW*. 1653–1662.

[5] Yen-Chun Chen, Zhe Gan, Yu Cheng, Jingzhou Liu, and Jing jing Liu. 2020. Distilling Knowledge Learned in BERT for Text Generation. In *ACL*.

[6] Zewen Chi, Li Dong, Furu Wei, Wenhui Wang, Xian-Ling Mao, and Heyan Huang. 2020. Cross-lingual Natural Language Generation via Pre-training. In *AAAI*.

[7] Kevin Clark and Christopher D. Manning. 2016. Deep Reinforcement Learning for Mention-Ranking Coreference Models. In *EMNLP*.

[8] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *ACL*.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

[10] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-Powered Conversational Agents. *ICLR*.

[11] Xiangyu Duan, Mingming Yin, Min Zhang, Boxing Chen, and Weihua Luo. 2019. Zero-Shot Cross-Lingual Abstractive Sentence Summarization through Teaching Generation and Attention. In *ACL*.

[12] Fangxiaoyu Feng, Yin-Fei Yang, Daniel Matthew Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT Sentence Embedding. *arXiv preprint arXiv:2007.01852* (2020).

[13] Joseph L. Fleiss. 1971. Measuring Nominal Scale Agreement among Many Raters. *Psychological Bulletin* 76 (1971), 378–382.

[14] Chongming Gao, Wenqiang Lei, Xiangnan He, Maarten de Rijke, and Tat-Seng Chua. 2021. Advances and Challenges in Conversational Recommender Systems: A Survey. *arXiv preprint arXiv:2101.09459* (January 2021).

[15] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning Word Vectors for 157 Languages. *arXiv preprint arXiv:1802.06893* (2018).

[16] Sangchul Hahn and Heeyoul Choi. 2019. Self-Knowledge Distillation in Natural Language Processing. In *RANLP*.

[17] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531* (2015).

[18] Zhuolin Jiang, Amro El-Jaroudi, William Hartmann, Damianos Karakos, and Lingjun Zhao. 2020. Cross-lingual Information Retrieval with BERT. *arXiv preprint arXiv:2004.13005* (2020).

[19] Xisen Jin, Wenqiang Lei, Zhaochun Ren, Hongshen Chen, Shangsong Liang, Yihong Zhao, and Dawei Yin. 2018. Explicit State Tracking with Semi-Supervision for Neural Dialogue Generation. In *CIKM*. 1403–1412.

[20] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale Similarity Search with GPUs. *arXiv preprint arXiv:1702.08734* (2017).

[21] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *EACL*.

[22] Wenqiang Lei, Xiangnan He, Yisong Miao, Qingyun Wu, Richang Hong, Min-Yen Kan, and Tat-Seng Chua. 2020. Estimation-action-reflection: Towards Deep Interaction between Conversational and Recommender Systems. In *WSDM*. 304–312.

[23] Wenqiang Lei, Xisen Jin, Min-Yen Kan, Zhaochun Ren, Xiangnan He, and Dawei Yin. 2018. Sequicity: Simplifying Task-oriented Dialogue Systems with Single Sequence-to-sequence Architectures. In *ACL*. 1437–1447.

[24] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A Diversity-Promoting Objective Function for Neural Conversation Models. *NAACL*.

[25] Lin-Xiao Li, Can Xu, W. Wu, Yufan Zhao, Xueliang Zhao, and Chongyang Tao. 2020. Zero-Resource Knowledge-Grounded Dialogue Generation. *NeurIPS*.

[26] Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. Learning to Select Knowledge for Response Generation in Dialog Systems. *IJCAI*.

[27] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *ACL*.

[28] Robert Litschko, Goran Glavas, Simone Paolo Ponzetto, and Ivan Vulic. 2018. Unsupervised Cross-Lingual Information Retrieval Using Monolingual Data Only. *SIGIR* (2018).

[29] Shuman Liu, Hongshen Chen, Zhaochun Ren, Yang Feng, Qun Liu, and Dawei Yin. 2018. Knowledge Diffusion for Neural Dialogue Generation. In *ACL*.

[30] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. *TACL* (2020).

[31] Longxuan Ma, Weinan Zhang, Runxin Sun, and Ting Liu. 2020. A Compare Aggregate Transformer for Understanding Document-grounded Dialogue. In *Findings of EMNLP*.

[32] Longxuan Ma, Wei-Nan Zhang, Mingda Li, and Ting Liu. 2020. A Survey of Document Grounded Dialogue Systems (DGDS). *arXiv preprint arXiv:2004.13818* (2020).

[33] Chuan Meng, Pengjie Ren, Zhumin Chen, Christof Monz, Jun Ma, and Maarten de Rijke. 2020. RefNet: A Reference-aware Network for Background Based Conversation. In *AAAI*.

[34] Chuan Meng, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tengxiao Xi, and Maarten de Rijke. 2021. Initiative-Aware Self-Supervised Learning for Knowledge-Grounded Conversations. In *SIGIR*. ACM.

[35] Chuan Meng, Pengjie Ren, Zhumin Chen, Weiwei Sun, Zhaochun Ren, Zhaopeng Tu, and Maarten de Rijke. 2020. DukeNet: A Dual Knowledge Interaction Network for Knowledge-Grounded Conversation. In *SIGIR*. ACM.

[36] Jian-Yun Nie. 2010. *Cross-Language Information Retrieval*. Morgan & Claypool.

[37] Vassilina Nikoulina, Bogomil Kovachev, Nikolaos Lagos, and Christof Monz. 2012. Adaptation of Statistical Machine Translation Model for Cross-Lingual Information Retrieval in a Service Context. In *EACL*.

[38] Douglas W. Oard. 1998. A Comparative Study of Query and Document Translation for Cross-Language Information Retrieval. In *AMTA*.

[39] Prasanna Parthasarathi and Joelle Pineau. 2018. Extending Neural Generative Conversational Model using External Knowledge Sources. In *EMNLP*.

[40] Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Póczos, and Tom Michael Mitchell. 2019. Competence-based Curriculum Learning for Neural Machine Translation. In *NAACL*.

[41] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

[42] Pengjie Ren, Zhumin Chen, Zhaochun Ren, Evangelos Kanoulas, Christof Monz, and Maarten de Rijke. 2020. Conversations with Search Engines. *arXiv preprint arXiv:2004.14162* (2020).

[43] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *TREC*.

[44] Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models. In *AAAI*.

[45] Siamak Shakeri, Abhinav Sethy, and Cheng Cheng. 2019. Knowledge Distillation in Document Retrieval. *AMLC* (2019).

[46] Haipeng Sun, Rui Wang, Kehai Chen, M. Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2020. Knowledge Distillation for Multilingual Unsupervised Neural Machine Translation. *arXiv preprint arXiv:2004.10171* (2020).

[47] Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. 1999. Policy Gradient Methods for Reinforcement Learning with Function Approximation. In *NIPS*.

[48] Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – Building open translation services for the World. In *EAMT*.

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. *NIPS*.

[50] Oriol Vinyals and Quoc V. Le. 2015. A Neural Conversational Model. *arXiv preprint arXiv:1506.05869* (2015).

[51] Ivan Vulic and Marie-Francine Moens. 2015. Monolingual and Cross-Lingual Information Retrieval Models Based on (Bilingual) Word Embeddings. *SIGIR* (2015).

[52] Jun Xu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Conversational Graph Grounded Policy Learning for Open-domain Conversation Generation. In *ACL*. 1835–1845.

[53] Ruochen Xu and Yiming Yang. 2017. Cross-lingual Distillation for Text Classification. *ACL* (2017).

[54] Ze Yang, Linjun Shou, Ming Gong, Wutao Lin, and Daxin Jiang. 2020. Model Compression with Two-stage Multi-teacher Knowledge Distillation for Web Question Answering System. *WSDM* (2020).

[55] 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 *ACL*.

[56] Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020. Knowledge-Grounded Dialogue Generation with Pre-trained Language Models. In *EMNLP*.

[57] Dong Zhou, Mark Truran, Tim Brailsford, Vincent Wade, and Helen Ashman. 2012. Translation Techniques in Cross-language Information Retrieval. *Comput. Surveys* 45, 1 (2012), Article 1.

[58] Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense Knowledge Aware Conversation Generation with Graph Attention. In *IJCAI*.