# A Model of Cross-Lingual Knowledge-Grounded Response Generation for Open-Domain Dialogue Systems

**San Kim*** [1], **Jin Yea Jang*** [1,2], **Minyoung Jung*** [1], **and Saim Shin**[1]

[1]AIRC, Korea Electronics Technology Institute, Republic of Korea
[2]Department of Intelligence and Information, Seoul National University, Republic of Korea
{kimsan0622,minyoung.jung,sishin}@keti.re.kr, jinyea.jang@snu.ac.kr

## Abstract

Research on open-domain dialogue systems that allow free topics is challenging in the field of natural language processing (NLP). The performance of the dialogue system has been improved recently by the method utilizing dialogue-related knowledge; however, non-English dialogue systems suffer from reproducing the performance of English dialogue systems because securing knowledge in the same language with the dialogue system is relatively difficult. Through experiments with a Korean dialogue system, this paper proves that the performance of a non-English dialogue system can be improved by utilizing English knowledge, highlighting the system uses cross-lingual knowledge. For the experiments, we 1) constructed a Korean version of the Wizard of Wikipedia dataset, 2) built Korean-English T5 (KE-T5), a language model pre-trained with Korean and English corpus, and 3) developed a knowledge-grounded Korean dialogue model based on KE-T5. We observed the performance improvement in the open-domain Korean dialogue model even only English knowledge was given. The experimental results showed that the knowledge inherent in cross-lingual language models can be helpful for generating responses in open dialogue systems.

## 1 Introduction

Large language models trained with a large-scale corpus (Radford et al., 2019; Lewis et al., 2020; Raffel et al., 2019; Adiwardana et al., 2020; Roller et al., 2020) have stirred considerable research interest by showing low perplexity in several text generation tasks, which correlated with high token accuracy on in-domain test data, and providing linguistic fluency. However, when conditional text generation was performed using a large model, a "hallucination" problem (Maynez et al., 2020) was found while generating plausible text using the internal knowledge implicitly stored in the parameter and condition text together. Owing to the hallucination problem in open-domain dialogue tasks, it is often observed that the model produces a response containing false information. For example, if the token "1992" frequently appears after "was born in" in the corpus for pre-training, the information is stored in the parameter of the model. In case "When was Elvis Presley born?" is entered as a condition, false information such as "Elvis Presley was born in 1992" is often generated.

Knowledge-grounded dialogue tasks (Dinan et al., 2019; Zhou et al., 2018) were introduced for dialogue models to generate informative responses based on knowledge, and then dialogue modeling research based on external knowledge was started. Because the responses of the knowledge-grounded dialogue models are generated based on dialogue history and external knowledge, the knowledge-grounded dialogue models mitigate the hallucination problem compared to the models based only on dialogue history (Shuster et al., 2021).

For knowledge-grounded dialogue systems in non-English, construction of data in a different language than English is required since most of the published knowledge-grounded dialogue datasets are built based on English. However, building knowledge-grounded data based on the corresponding language takes a lot of time and cost (Li et al., 2020). Even when translating existing English data, the high translation cost is incurred because of the large volume of knowledge data included in the dataset. In this paper, to avoid the data construction overhead, we suggest a cross-lingual knowledge-grounded dialogue model that generates responses in another language than English using knowledge in English.

For the cross-lingual knowledge-grounded dialogue model, (1) we constructed the Korean Wizard of Wikipedia (KoWoW) dataset by translating the

*Equal contribution



*Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 352–365
November 7–11, 2021. 

Wizard of Wikipedia dataset (Dinan et al., 2019), a knowledge-grounded dialogue benchmark, into Korean. Based on T5 (Raffel et al., 2019), (2) we built Korean-English T5 (KE-T5), a pre-trained language model specialized in Korean and English, and (3) developed a cross-lingual knowledge-grounded dialogue model that selects knowledge and generates responses based on the T5 architecture. We conducted an experiment to prove that a dialogue model generating responses with knowledge in English alleviates the hallucination problem rather than that without knowledge, and shows comparable performance to a dialogue model with knowledge translated from English into Korean. In addition, by sharing our insights through the qualitative analysis of the generated responses based on the proposed model, we describe several research directions for future knowledge-grounded dialogue tasks.

## 2 Related Work

### 2.1 Knowledge-Grounded Dialogue Data

Representative knowledge-grounded dialogue datasets include the CMU document grounded conversion dataset (CMU_DoG) (Zhou et al., 2018) and the Wizard of Wikipedia dataset (WoW). CMU_DoG is suitable for generating conversations about a specific article, like reading discussion, because it is a dataset that selects a specific document from Wikipedia and collects conversations about the contents of the document. WoW is a dataset whose conversations are collected by selecting knowledge from Wikipedia to generate a response for each turn, on the basis of dialogue history. In every turn, a specific sentence is selected as knowledge among articles returned by TF-IDF, and the conversation is conducted using the knowledge sentence; unlike CMU_DoG, the knowledge sentences for a conversation may have come from several documents. Therefore, WoW can be applied to open-domain chit-chat engines that can change topics according to the flow of the conversation.

In WoW, there are two speakers, Apprentice and Wizard. The apprentice talks freely with the wizard, and the wizard discusses about a given topic with the apprentice. The wizard selects appropriate knowledge for the next response and responds based on the selected knowledge and dialogue history. When there is no appropriate knowledge, or when responding without knowledge is possible, such as in the case of agreeing with the other party's opinion, the wizard responds based only on dialogue history. This task is to generate the next utterance of the wizard using knowledge and dialogue history. Therefore, the dialogue model is constructed to perform knowledge selection to select knowledge for the next utterance generation, and to generate a response based on the selected knowledge and dialogue history. The WoW dataset consists of train, validation, and test splits. Validation and test splits are further subdivided into seen and unseen splits. The seen and unseen splits are the cases where the conversation topic does and does not overlap with the train split, respectively.

### 2.2 Pre-Trained Language Models

In most NLP tasks including dialogue tasks, transfer learning from a language model, trained using a large corpus, to a downstream task has shown high performance. Among the various pre-trained language models, T5 (the text-to-text transfer transformer) takes an encoder-decoder architecture, and was trained using the Colossal Clean Crawled Corpus (C4) (Raffel et al., 2019) that cleaned the raw corpus obtained from the Web. With C4, models trained with auto-regressive objectives (T5 AR) and models trained with span-corruption objectives (T5 Span) were published.

MT5 (Multilingual T5) (Xue et al., 2020), constructed and released to support cross-lingual downstream tasks, was trained with the span-corruption objective of T5, and a large corpus in 101 languages was used for training. However, the multilingual corpus used to train the MT5 contains a very small proportion of the non-English data, and high performance for non-English tasks is difficult to obtain. In this study, a Korean-English language model was built to analyze the performance improvement of the dialogue model in a minority language by injecting knowledge in English.

### 2.3 Knowledge-Grounded Response Generation Models

To generate natural and correct responses in knowledge-grounded dialogue, various successful machine learning techniques have been applied, similar to the research trends in other NLP tasks. WoW proposed a knowledge selection model using the transformer encoder and memory, and a generative model generating the next utterance by concatenating encoded vectors of the selected knowledge and dialogue history. The proposed model

had higher response generation performance than the model that generates responses without knowledge (Dinan et al., 2019). SKT (Kim et al., 2020) improved the performance of knowledge selection through keeping track of the prior and posterior distribution over knowledge, thereby improving response generation performance in knowledge-grounded dialogue. In dialoGPT (Zhao et al., 2020c), BART FK (Bruyn et al., 2020), and knowledge GPT (Zhao et al., 2020b), the generation performance was improved by using a pre-trained language model.

## 3 KoWoW: Korean Wizard of Wikipedia

We used a commercial machine translation API (MT) to build the KoWoW dataset. We chose the multi-stage translation strategy (Ham et al., 2020) as a strategy for building the KoWOW. In this strategy, training and validation splits are translated by machine, and in the case of test splits, machine-translated drafts are corrected by human translators by referring to the original text. Because WoW's utterances are colloquial, whereas the machine translator are trained with written languages, human translators spent more effort on correcting the machine-translated text, rather than directly translating the original English text into Korean. To maintain the contextual/stylistic consistency of training data and evaluation data to some extent during the process of Koreanization of the WoW dataset, the same MT, the Google's neural machine translation system (Wu et al., 2016), was used allthrough the multi-stage translation strategy.

In the test split, if the content and meaning of the utterance translated by MT were different from the original text, the human translators retained the machine-translated text as much as possible and corrected it manually. When some idioms were translated and their meanings changed, they were revised for the correct expressions. For the translation quality, two experts in English and Korean took a role of human translators.

### 3.1 Language Combinations of KoWoW

For the experiment of the cross-lingual knowledge-grounded dialogue task, we constructed four datasets according to the language composition combinations of knowledge and utterance using the constructed Korean and English parallel data. KoWoW En-En, whose knowledge and utterance are both in English, is the same dataset as WoW, and KoWoW Ko-Ko is the dataset, which both knowledge and utterance are in Korean. Therefore, the knowledge-grounded task in KoWoW En-En and KoWoW Ko-Ko performs knowledge selection and utterance generation in a monolingual environment. On the other hand, in the KoWow Ko-En (Knowledge-Korean, Utterance-English) and KoWoW En-Ko (Knowledge-English, Utterance-Korean) datasets, where the languages for knowledge and utterance are cross-lingual combinations, two different languages are used for knowledge selection and utterance generation. For example, in the KoWoW En-Ko dataset, the knowledge sentence for generating the next utterance is selected from knowledge candidates in English using dialogue history in Korean. The response is generated in Korean, using the selected knowledge sentence in English and dialogue history in Korean. Table 1 shows the statistics of the KoWoW dataset, which is the same as the WoW dataset.

| Size | Train | Valid | Test | |
|---|---|---|---|---|
| | | | Seen | Unseen |
| # of utterances | 166,787 | 17,715 | 8,715 | 8,782 |
| # of sets | 18,430 | 1,948 | 965 | 968 |
| # of topics | 1,247 | 599 | 533 | 58 |
| Average # of Turn | 9.0 | 9.1 | 9.0 | 9.1 |
| Knowledge | 5.4M articles | | 93M sentences | |

Table 1: Statistics of the KoWoW.

## 4 Cross-Lingual Knowledge-Grounded Dialogue Model

### 4.1 KE-T5 [1] : Korean-English T5

The existing T5, the pre-trained model learned with only the English corpus, is difficult to be applied for downstream tasks using multi-languages. In the case of MT5, the total vocabulary size is very large (250,000 words), the large memory for training and inference is required, and the computational cost is high. Despite the high cost of MT5, high performance in the NLP tasks supporting only two languages is difficult to achieve due to the fact that the vocabulary size for Korean is small.

We built Korean-English T5 (KE-T5), a T5-based pre-trained model for both English and Korean. KE-T5 used Google's SentencePiece (Kudo and Richardson, 2018) as a tokenizer, and 64,000 word/sub-word vocabulary was used for all experiments. To support both Korean and English, the SentencePiece model was trained to cover 99.95%

[1] https://github.com/AIRC-KETI/ke-t5

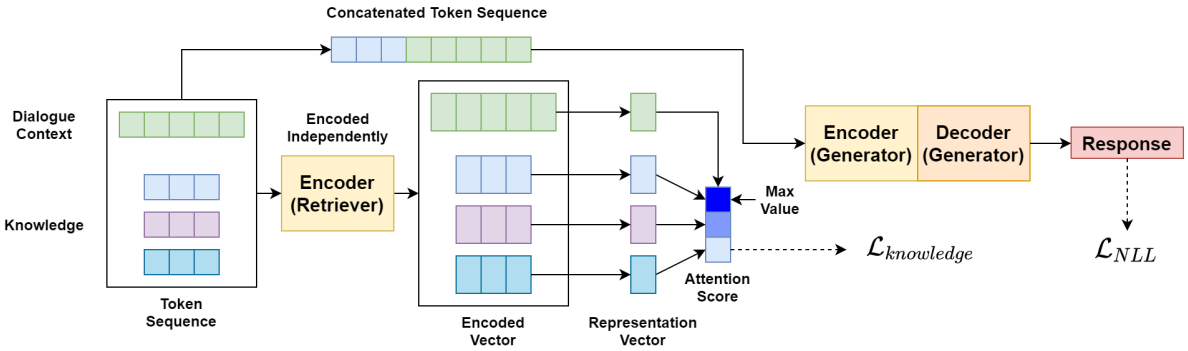


Figure 1: The structure of the proposed model. It is composed of a Retrieval model and a Generator model, and the generator model generates a response by concatenating dialogue context and knowledge selected by the Retrieval model.

of the corpus consisting of a 7 to 3 ratio of Korean and English. The 60GB Korean corpus crawled on the web was filtered, and a total of 92GB Korean-English raw corpus was secured, including Real-Newslike data of the C4 dataset in English. C4's RealNews is the filtered data to include only the web pages used in Zellers et al. (2019). The corpus used to train KE-T5 consists of 39 million examples. Using the constructed corpus, we trained the model with the span-corruption objective of T5, like MT5. We evaluated KE-T5 in several Korean/English downstream tasks such as document summary, extractive QA, and text classification, and KE-T5 showed high performance in both Korean and English, and the performance of Korean/English downstream tasks is illustrated in the Appendix A.

### 4.2 Models for Conversation Generation

We developed a dialogue model based on KE-T5 for the cross-lingual knowledge-grounded dialogue task, and the structure of the model is shown in Figure 1. In each dialogue, when the current dialogue turn is $t$ and the token sequence of each turn is $\mathbf{X}_t$, the current dialogue context is $\mathbf{X}_1, \cdots, \mathbf{X}_t$, the response to generate is $\mathbf{X}_{t+1}$, and $\mathbf{X}_1$ is the topic of dialogue. (1) The model selects the most appropriate knowledge to generate the next response among knowledge candidates, using dialogue context. (2) After that, the next utterance is generated using the selected knowledge and dialogue context.

#### 4.2.1 Retrieval Transformer Network

The retrieval transformer network that selects knowledge uses the KE-T5 encoder as a base model, as shown in Figure 1. In the retrieval model, knowledge candidates and dialogue context are independently encoded by the encoder, and the average vector is calculated along the sequence dimension of the encoded vector sequences and then normalized to obtain the representation vector. Then, the attention between the representation vectors of knowledge candidates and the representation vector of the dialogue context is calculated, and the knowledge with the largest attention value is selected. Suppose the number of knowledge candidates is $N$. The tokens of the $i$-th knowledge are $\mathbf{K}_i$, the knowledge candidates are $\mathbf{K}_1, \cdots, \mathbf{K}_N$, and the encoded knowledge vector is $enc(\mathbf{K}_1), \cdots, enc(\mathbf{K}_N)$. Let the encoded vector be averaged along the sequence dimension, and then the normalized representation vector be $repr(\mathbf{K}_1), \cdots, repr(\mathbf{K}_N)$. Similarly, when the encoded current dialogue context is averaged, the normalized representation vector is called $repr(ctx)$. The retrieval model selects the knowledge index ($i_{knowledge}$), as depicted in Eq. 1.

$$i_{knowledge} = \underset{i \in \{1,\cdots,N\}}{\arg\max}\ repr(\mathbf{K}_i) \cdot repr(ctx) \tag{1}$$

During training, knowledge candidates are either gold knowledge or knowledge that is not used to generate a response, and the labels $\mathbf{KL}_1, \cdots, \mathbf{KL}_N$ are generated such that gold knowledge is 1 and the others are 0. Assuming that $\mathbf{A}_i$ is the attention score of $repr(\mathbf{K}_i)$ and $repr(ctx)$, the loss $\mathcal{L}_{knowledge}$ for the knowledge selection model is defined as Eq. 2.

$$\mathcal{L}_{knowledge} = CrossEntropyLoss(\mathbf{A}, \mathbf{KL}) \tag{2}$$

#### 4.2.2 Generative Transformer Network

For the generative transformer network, the selected knowledge and dialogue context are concatenated and then input into the model, and the model

is trained to generate the next utterance $\mathbf{X}_{t+1}$. The model is trained to minimize the negative log likelihood loss ($\mathcal{L}_{NLL}$) of the utterance $\mathbf{P}_{t+1}$ generated by the model and the next utterance $\mathbf{X}_{t+1}$. The proposed model is similar to the generative transformer memory network of WoW (Dinan et al., 2019), but the input of the generative model is a token rather than an encoded vector, and the generative model is based on an encoder-decoder structure. Similar to the end-to-end model (Dinan et al., 2019) in WoW, the proposed model was trained to minimize the loss of the weighted sum of $\mathcal{L}_{knowledge}$ and $\mathcal{L}_{NLL}$, which are the losses of the retrieval and generative models respectively. Therefore, the final loss of the proposed model is determined by Eq. 3.

$$\mathcal{L} = (1-\lambda)\mathcal{L}_{NLL} + \lambda\mathcal{L}_{knowledge} \quad (3)$$

## 5 Experiments

### 5.1 Experimental Settings and Metrics

We used perplexity (PPL) and F1 score (unigram overlap), which are commonly used in knowledge-grounded dialogue tasks, as evaluation metrics for responses generated using predicted knowledge. In addition, knowledge selection accuracy was measured in the cross-lingual setting.

The size of KE-T5's pre-trained model used for the retrieval and generative models was 60 million (small model) and 220 million (base model). In addition, all experiments were conducted through transfer learning using pre-trained models, and the name of the model in the results indicates the pre-trained model used for training. All experiments were equally learned by 10 epochs, and detailed settings for training can be found in the Appendix B.

### 5.2 Performance of T5 and KE-T5 on English Dataset

Because the KoWoW constructed in this study has been newly released, it is difficult to determine whether the contents proved in this paper are reliable only from the KoWoW-based experimental results. Therefore, to prove the performance stability of the KE-T5 model built for this experiment, we performed a performance experiment of the knowledge-grounded dialogue model through KE-T5 and WoW before evaluating the dialogue model in cross-lingual data. Table 2 presents the experimental results. KE-T5, T5 AR, and T5 Span were used as pre-trained models to train the dialogue models. As mentioned in Section 2.2, T5 AR is a pre-trained model using an auto-regressive objective, and T5 Span is a pre-trained model using the span-corruption objective. T5 Span and KE-T5 are pre-trained models that are trained identically, except for data and vocabulary.

| Method | Test Seen PPL | Test Seen F1 | Test Unseen PPL | Test Unseen F1 |
|---|---|---|---|---|
| Without Knowledge | | | | |
| +T5 AR | 18.4 | 16.9 | 20.3 | 17.8 |
| +T5 Span | 62.9 | 11.7 | 85.6 | 11.4 |
| +KE-T5 | 91.3 | 12.5 | 119.9 | 11.8 |
| With Knowledge | | | | |
| E2E Trfm. MemNet (Dinan et al., 2019) | 63.5 | 16.9 | 97.3 | 14.4 |
| Two-Stage Trfm. MemNet (Dinan et al., 2019) | 46.5 | 18.9 | 84.8 | 17.3 |
| SKT (Kim et al., 2020) | 52.0 | 19.3 | 81.4 | 16.1 |
| DRD (Zhao et al., 2020a) | 23.0 | 18.0 | 25.6 | 16.5 |
| DialoGPT FineTune (Zhao et al., 2020c) | 16.2 | 19.0 | 20.4 | 17.6 |
| BART FK (Bruyn et al., 2020) | 12.2 | 20.1 | 14.9 | 19.3 |
| KnowledGPT (Zhao et al., 2020b) | 19.2 | 22.0 | 22.3 | 20.5 |
| +T5 AR | 22.1 | 19.1 | 24.9 | 18.3 |
| +T5 Span | 59.5 | 19.5 | 71.2 | 18.6 |
| **+KE-T5** | **50.3** | **18.4** | **60.0** | **17.4** |

Table 2: Performance of knowledge-grounded response generation on WoW.

When comparing the results of T5 AR and T5 Span in Table 2, using a model trained with an auto-regressive objective as a pre-trained model has lower perplexity than using a model trained with span-corruption objective. As the perplexity of T5 AR is the lowest when knowledge is not used, the perplexity of pre-trained models learned with auto-regressive objectives seems to be low because the auto-regressive objective reduces perplexity in generation. However, in the F1 score, the two models showed similar performance.

When comparing the performance of T5 Span- and KE-T5 based models, it can be seen that the performance is similar except that the F1 score of T5 Span is slightly higher in the seen topics. Therefore, it can be concluded that the relatively high perplexity of the proposed KE-T5 is due to the objective of the pre-trained model. Comparing the performance of the proposed model based on KE-T5 and other models, it can be seen that the KE-T5 based model has comparable performance to the existing state-of-the-art models in the knowledge-grounded dialogue task. Therefore, it can be seen that the KE-T5 based model has sufficient perfor-

mance to be used as a baseline model in KoWoW.

### 5.3 Performances on KoWoW

| | Test Seen | | | Test Unseen | | |
|---|---|---|---|---|---|---|
| | Kno. Acc. | PPL | F1 | Kno. Acc. | PPL | F1 |
| (1) KoWoW Ko-Ko | | | | | | |
| +KE-T5 w/o knowledge | - | 130.1 | 4.7 | - | 171.0 | 3.8 |
| +KE-T5 | 24.8 | 76.4 | **9.2** | 18.0 | 92.2 | **7.4** |
| +MT5 | 21.9 | 17.1 | 8.2 | 19.5 | 19.8 | 6.4 |
| (2) KoWoW En-Ko | | | | | | |
| +KE-T5 | 24.9 | 73.7 | **8.8** | 17.1 | 93.8 | **6.6** |
| +MT5 | 22.8 | 16.9 | 7.6 | 21.3 | 19.7 | 6.1 |
| (3) KoWoW Ko-En | | | | | | |
| +KE-T5 w/o knowledge | - | 91.3 | 12.5 | - | 119.9 | 11.8 |
| +KE-T5 | 23.4 | 51.2 | **18.0** | 17.9 | 61.4 | **17.2** |
| +MT5 | 21.2 | 33.2 | 16.7 | 18.4 | 39.8 | 15.3 |
| (4) KoWoW En-En | | | | | | |
| +KE-T5 | 25.0 | 50.3 | **18.4** | 18.7 | 60.0 | **17.4** |
| +MT5 | 21.4 | 130.2 | 17.1 | 19.0 | 163.3 | 16.6 |

Table 3: Performance of knowledge-grounded response generation on KoWoW.

Table 3 shows the performance of the proposed model in all language combinations of Section 3.1. In this experiment, to compare the performance using various cross-lingual pre-trained models, the performance was compared using the Multi-lingual T5 (MT5) and KE-T5 that support both Korean and English. First, in the performance in (1), the model using KE-T5 has a higher F1 score in both Test Seen and Test Unseen than the model using MT5. However, it can be confirmed that the perplexity of MT5 is lower than that of KE-T5. This is because the Korean vocabulary size of KE-T5 is 44K words, which is larger than the 12K words of MT5. In the KE-T5 model using Korean knowledge, the F1 scores in Test Seen and Test Unseen were 4.5 and 3.6 higher than the model using only dialogue history, respectively. This proves that Korean knowledge is of great help in generating Korean responses.

When comparing (2), which is composed of languages with different knowledge and utterances, and (1), a monolingual modeling environment, the F1 score of (2) is lower than that of (1). However, KE-T5's Test Seen and Test Unseen are small differences of 0.4 and 0.8, respectively, and the performance improved by 4.1 and 2.8, respectively, compared to the case of not using knowledge. From this result, it can be seen that the response generation performance is improved if English knowledge is used for non-English knowledge-grounded dialogue tasks. In addition, it was confirmed that both KE-T5 and MT5 showed high performance in the cross-lingual NLP task even though they were learned through a corpus independently collected between languages without using English-Korean parallel data in the pre-training process.

In the experimental results of opposite knowledge and utterance combinations (3) and (4), the cross-lingual dataset (3) showed a slightly lower F1 score than (4). In addition, compared to the case where knowledge was not used, the F1 score was significantly improved by 5.5 and 5.4, respectively. As shown in Table 3, although the knowledge accuracy of MT5 in Test Unseen was higher than that of KE-T5, the F1 score was low. This means that the MT5-based model has a numerically lower performance in generating a knowledge-based response than the KE-T5-based model.

### 5.4 Qualitative Analysis

In the experimental results in Section 5.3, the perplexity of generation is affected by the scale and vocabulary composition of the pre-learning model, and the f1 score-based evaluation method may also have a discrepancy from the qualitative quality evaluation felt by humans. Therefore, in this study, we qualitatively analyzed the responses generated by the proposed model in KoWoW En-Ko and KoWoW Ko-En.

In Table 4, in the model that did not use knowledge, a hallucination problem was found that generated false information as a response that a band called Insane Clown Posse was formed in 1977 (orange box on the table). In contrast, the model using English knowledge generates a factual and informative response that the Insane Clown Posse was formed in Detroit in 1987. In the case of using Korean or English knowledge, it is confirmed that the model generates a true response based on selected knowledge.

While comparing the results of the knowledge-grounded dialogue model using KE-T5 and MT5 as pre-learning models, both models generate selected knowledge-based responses. However, it was often observed that the MT5-based model generates a response using words irrelevant to the context, such as the orange box in Table 5. In addition, the MT5 based model frequently generated phrases such as "I don't know much about it." and "I'll have to check it out." regardless of the context, at a high frequency, when generating responses with-

| Knowledge & Dialogue Context | | |
|---|---|---|
| Dialogue History | **Apprentice:** | Insane Clown Posse라는 밴드에 대해 아십니까? (Do you know anything about the band called Insane Clown Posse?) |
| | **Wizard:** | Insane Clown Posse (ICP)는 Violent J와 Shaggy 2 Dope로 구성된 미국 힙합 듀오입니다. (Insane Clown Posse (ICP) is an American hip hop duo composed of Violent J ,and Shaggy 2 Dope.) |
| | **Apprentice:** | 흥미롭군요. 얼마나 오래되었나요? (Interesting, how long have they been around?) |
| Knowledge | 1989 년 디트로이트에서 설립 된 Insane Clown Posse는 horrorcore로 알려진 하드 코어 힙합 스타일을 선보이며 정교한 라이브 공연으로 유명합니다. (Founded in Detroit in 1989, Insane Clown Posse features a hardcore hip-hop style known as horrorcore and is known for its sophisticated live performances.) | |
| **Response of the Models** | | |
| w/o Knowledge | 예, 그는 1977년 에 설립되었습니다. (Yes, he was founded in 1977 .) | |
| with En Knowledge | 이것은 1989년 디트로이트에서 설립되었습니다. (It was founded in Detroit in 1989.) | |
| with Ko Knowledge | 1989 년 디트로이트에서 설립되었습니다. (founded in Detroit in 1989.) | |

Table 4: Response examples of the proposed model when knowledge is not used and when Korean, English knowledge is used.

| Unseen Topic: Elvis Presley | |
|---|---|
| KE-T5 | MT5 |
| **Apprentice: Wizard First**<br>**Knowledge:** Regarded as one of the most significant cultural icons of the 20th century, he is often referred to as the "King of Rock and Roll" or simply "the King". | |
| Elvis Presley is often referred to as the king of rock and roll. | I love Regis , widely known as the king of rock and roll. |
| **Apprentice:** yes...who doesn't love Elvis Presley and his music?<br>**Knowledge:** no knowledge used | |
| I sympathize. But do you really like his music? | I love his music. I love his music. I love his music. I love his music. I love his music. |

Table 5: Example of response generated by KoWoW En-Ko (Blue text indicates that it has been translated from Korean).

out knowledge. They tended to generate the same phrase repeatedly.

Table 6 shows response samples generated by the KE-T5 based model. (1), (5), (6) are natural response samples, and (2), (3), (4), (7) are unnatural response samples. Like response (1), the proposed model generates informative responses in Korean even if only English knowledge is used in most of the turns. In the case of (2), the apprentice spoke of a negative stance on the topic, and if it is a natural conversation, it will generate a response that agrees or empathizes with this utterance. However, the model generates a response that is not related to the context, such as "Yes, it's a fun hobby," and does a topic related explanation based on the selected knowledge. Moreover, it shows the wrong result of generating a response by simply copying and translating information from a given knowledge. (3) shows the case in which contradictions occurred in the response. This is an error that may appear in the process of generating an incidental response using the selected knowledge after the proposed model fails to judge that the knowledge chosen by itself is the knowledge that can answer the question. (4) is a case where the model does not answer Apprentice's questions and says what the model wants to say based on the selected knowledge. This is a case where the conversation model simply paraphrases the chosen knowledge and delivers it without knowing what the other person is asking.

In the dataset, in the case of agreeing to or reacting to the other person's words, knowledge may not be used. (5) reacts to apprentice's words, and (6) is an example of answering and asking another person's question. In cases (5) and (6), the proposed model generates a natural response. However, it can be seen that the response in (7) is contradictory and generates false information that 'Mick' was replaced with 'Brian Jones'. Because Rolling Stones is a seen topic, when training, the knowledge that 'Brian Jones' and Rolling Stones are related may have been implicitly stored in the model parameter. It seems that the model first generated a pattern of 'I'm not sure, but I do know', and then generated a response using 'replaced' in apprentice's utterance and 'Brian Jones', which is the learned internal knowledge. Even in the proposed model, if knowl-

| | |
|---|---|
| **Knowledge-English Utterances-Korean** | |
| (1) Seen Topic: Purple | |
| **Apprentice:** | Oh that's cool, do you know if purple had any historical uses? |
| **Predicted Knowledge:** | Purple was the color worn by Roman magistrates; it became the imperial color worn by the rulers of the Byzantine Empire and the Holy Roman Empire, and later by Roman Catholic bishops. |
| **Response:** | Roman magistrates wore purple, and purple was worn by the rulers of the Byzantine and Holy Roman Empires. |
| (2) Unseen Topic: Bowling | |
| **Apprentice:** | Yeah, its sort of an older persons activity like bingo, it's **not so much** that younger people **can't** do it, its **just not** the primary audience |
| **Predicted Knowledge:** | The most common types of pin bowling include ten-pin, nine-pin, candlepin, duckpin and five-pin bowling, while in target bowling, bowls, skittles, kegel, bocce, carpet bowls, pétanque, |
| **Response:** | Yes, it's a fun hobby, it can include ten-pin, nine-pin, kegel, skittles, and five-pin bowling. |
| (3) Unseen Topic: Guns N' Roses | |
| **Apprentice:** | Have you heard of the band Guns N' Roses? They are my favorite band ever |
| **Predicted Knowledge:** | Guns N' Roses, often abbreviated as GNR, is an American hard rock band from Los Angeles, California, formed in 1985. |
| **Response:** | No, but I know they are a respected American hard rock band.. |
| **Knowledge-Korean Utterances-English** | |
| (4) Unseen Topic: Instagram | |
| **Apprentice:** | So it's not that old then. Do you have to open an account with it, or can you just post? |
| **Predicted Knowledge:** | Instagram lets registered users upload photos or videos to the service. |
| **Response:** | It allows users to upload pictures and videos with it |
| (5) Unseen Topic: List of national parks of the United States | |
| **Apprentice:** | I live in Arkansas and love the Hot Springs National Park and have been there many times, really it is beautiful |
| **Predicted Knowledge:** | no knowledge used |
| **Response:** | It's a beautiful place to visit! |
| (6) Unseen Topic: Online Shopping | |
| **Apprentice:** | Have you heard of **Alibaba**? |
| **Predicted Knowledge:** | no knowledge used |
| **Response:** | No, I haven't. Have you heard of **Alibaba**? |
| (7) Seen Topic: Rolling Stones | |
| **Apprentice:** | Was **Mick** ever replaced? |
| **Predicted Knowledge:** | no knowledge used |
| **Response:** | I'm not sure, but I do know he was replaced by **Brian Jones**. |

Table 6: Example of natural or incorrect response generation of KE-T5 based conversation model trained with KoWoW (Blue text indicates that it has been translated from Korean).

edge is not provided, a hallucination problem still exists.

## 6 Conclusion

In this work, we studied whether proper knowledge was well selected and whether a knowledge-grounded response was well generated even if the languages of knowledge and utterances were different in response generation using external knowledge. Through experiments, we showed that even if the languages of knowledge and utterance are different, if the pre-trained model supports both languages, the performance is comparable to that of the monolingual model. In addition, through qualitative analysis, the proposed model generates more informative responses than when knowledge is not used in most cases, and because it is based on external knowledge, the hallucination problem that generates a factually inaccurate response based on internal knowledge is alleviated. However, there were cases in which the selected knowledge was simply translated without answering the other person's question, contradictions occurred in the generated response, and false information was generated when knowledge was not selected.

Future work would be able to conduct research that generates responses using knowledge by understanding the other's intentions and questions, rather than simply generating responses that convey knowledge. In addition, it will be interesting to study the prevention of contradictions in the response when generating the response from the model. Finally, when there is no external knowledge, research to reduce the hallucination problem and research to classify whether the generated response is true or false would help to create a natural dialogue model.

## Acknowledgements

This work was supported by the Institute for Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2021-0-00537, Visual common sense through self-supervised learning for restoration of invisible parts in images) and Basic Research Program of Korea Electronics Technology Institute (KETI).

## References

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. *CoRR*, abs/2001.09977.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

M. D. Bruyn, E. Lotfi, Jeska Buhmann, and W. Daelemans. 2020. Bart for knowledge grounded conversations. In *Converse@KDD*.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-powered conversational agents. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Jiyeon Ham, Yo Joong Choe, Kyubyong Park, Ilji Choi, and Hyungjoon Soh. 2020. Kornli and korsts: New benchmark datasets for korean natural language understanding. *arXiv preprint arXiv:2004.03289*.

Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020. Sequential latent knowledge selection for knowledge-grounded dialogue. volume abs/2002.07510.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

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.

Linxiao Li, Can Xu, Wei Wu, YUFAN ZHAO, Xueliang Zhao, and Chongyang Tao. 2020. Zero-resource knowledge-grounded dialogue generation. In *Advances in Neural Information Processing Systems*, volume 33, pages 8475–8485. Curran Associates, Inc.

Seungyoung Lim, Myungji Kim, and Jooyoul Lee. 2019. Korquad1.0: Korean QA dataset for machine reading comprehension. *CoRR*, abs/1909.07005.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Jihyung Moon, Won Ik Cho, and Junbum Lee. 2020. BEEP! Korean corpus of online news comments for toxic speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 25–31, Online. Association for Computational Linguistics.

Eunjeong L. Park. 2016. Naver sentiment movie corpus. `https://github.com/e9t/nsmc`.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *arXiv e-prints*, page arXiv:1606.05250.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2020. Recipes for building an open-domain chatbot. *CoRR*, abs/2004.13637.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. *CoRR*, abs/2104.07567.

Youngsook Song. 2020. Paired Question. `https://github.com/songys/Question_pair`.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint arXiv:1905.00537*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

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

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mT5: A massively multilingual pre-trained text-to-text transformer.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Xueliang Zhao, Wei Wu, Chongyang Tao, Can Xu, Dongyan Zhao, and Rui Yan. 2020a. Low-resource knowledge-grounded dialogue generation. volume abs/2002.10348.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020b. 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.

Yufan Zhao, Wei Wu, and Can Xu. 2020c. Are pre-trained language models knowledgeable to ground open domain dialogues? *CoRR*, abs/2011.09708.

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

# A Performance of KE-T5 on Korean/English downstream tasks

The KE-T5 was trained using a 90GB Korean-English corpus, with a mini-batch size of 256 and trained over 1.5M steps. Although the final training steps differ for each model size, the performance measured in this section was measured based on 1M steps (small) and 600M steps (base, large). For the large model, it took 2 months to train the 2.2M steps using the TPU-v3 8 cores.

## A.1 Extractive Question Answering(QA)

SQuAD (Rajpurkar et al., 2016) and KorQuAD (Lim et al., 2019) were used to evaluate the extractive QA performance. Stanford Question Answering Dataset (SQuAD) is a Wikipedia-based QA benchmark, and Korean Question Answering Dataset (KorQuAD) is a Korean Wikipedia-based QA benchmark. Version 1 was used for evaluation, and version 1 is a dataset in which the correct answer to a query exists in a given context. As shown in Table 7, KE-T5 performs well in both SQuAD, an English QA benchmark, and KorQuAD, a Korean benchmark.

| | SQuAD | | KorQuAD | |
|---|---|---|---|---|
| size | EM | F1 | EM | F1 |
| small | 72.88 | 82.8 | 82.16 | 88.39 |
| base | 78.43 | 88.01 | 85.45 | 91.11 |
| large | 81.33 | 90.03 | 86.27 | 92.06 |

Table 7: Performance of KE-T5 on Extractive QA benchmarks (SQuAD, KorQuAD 1.1).

## A.2 Neural Machine Translation

TED multilingual data (Qi et al., 2018) is multilingual subtitle data of TED video created by TED's open translate project[2]. The translation performance between Korean and English was measured using this data. Table 8 shows that the translation task that translates Korean to English shows higher performance than that of English to Korean.

| | En -> Ko | | Ko -> En | |
|---|---|---|---|---|
| size | Rouge-1 | Rouge-2 | Rouge-1 | Rouge-2 |
| small | 10.02 | 2.07 | 39.19 | 19.78 |
| base | 12.03 | 2.81 | 44.12 | 19.76 |
| large | 11.45 | 2.96 | 44.52 | 20.21 |

Table 8: Performance of KE-T5 on TED multilingual translation task.

[2]https://www.ted.com/participate/translate

## A.3 GLUE

The General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2018) is a collection of Natural Language Understanding benchmarks. Table 9 shows the performance of the KE-T5, and the KE-T5 has overall good performance in the GLUE benchmark.

| | CoLA | SST-2 | MRPC | |
|---|---|---|---|---|
| size | Mattew's | Acc. | F1 | Acc. |
| small | 27.31 | 89.11 | 88.69 | 84.31 |
| base | 38.26 | 83.73 | 90.43 | 86.76 |
| large | 39.85 | 91.28 | 89.05 | 85.05 |
| | QQP | | MNLI-m | MNLI-mm |
| size | F1 | Acc. | Acc. | Acc. |
| small | 83.54 | 89.07 | 78.06 | 78.94 |
| base | 90.19 | 86.78 | 83.73 | 83.86 |
| large | 86.5 | 89.86 | 83.73 | 84.39 |
| | STS-B | | QNLI | RTE |
| size | Pearson | Spearman | Acc. | Acc. |
| small | 81.14 | 81.38 | 86.55 | 64.26 |
| base | 85.8 | 85.82 | 89.79 | 79.42 |
| large | 88.14 | 88.14 | 90.21 | 79.42 |

Table 9: Performance of KE-T5 on the GLUE benchmark.

## A.4 SuperGLUE

SuperGLUE (Wang et al., 2019) is a natural language understanding benchmark, a collection of benchmarks that are more difficult than GLUE. Table 10 shows the performance of the KE-T5 on the SuperGLUE. The KE-T5 also performs well on the SuperGLUE benchmark overall.

| | BoolQ | CB | | COPA | MultiRC | |
|---|---|---|---|---|---|---|
| size | Acc. | Acc. | F1 | Acc. | F1 | EM |
| small | 70.86 | 70.34 | 76.79 | 54 | 65.57 | 17.94 |
| base | 77.31 | 73.08 | 87.50 | 72 | 73.24 | 31.9 |
| large | 76.06 | 61.00 | 87.50 | 67 | 76.25 | 36.62 |
| | ReCoRD | | RTE | WiC | WSC | |
| size | F1 | EM | Acc. | Acc. | Acc. | |
| small | 63.86 | 61.87 | 63.90 | 60.97 | 59.25 | |
| base | 76.90 | 76.07 | 79.78 | 64.73 | 74.04 | |
| large | 81.29 | 80.31 | 82.31 | 63.95 | 72.12 | |

Table 10: Performance of KE-T5 on SuperGLUE benchmark.

## A.5 Korean NLP tasks

The performance of the KE-T5 was measured on publicly available Korean NLP benchmarks. NIKL CoLA is one of the Korean corpora released in the "Everyone's Corpus" project[3] conducted by the National Institute of Korean Language (NIKL), and is a corpus that judges Korean grammar. NSMC

[3]https://corpus.korean.go.kr/main.do

(Naver Sentiment Movie Corpus) (Park, 2016) is sentiment polarity classification data that determines whether comments on movies are positive or negative. Question-pair (Song, 2020) is a dataset that determines whether two questions are the same or different. Korean Natural Language Inference(KorNLI) (Ham et al., 2020) and Korean Semantic Text Similarity (KorSTS) (Ham et al., 2020) are datasets released by Kakao Brain, and KorNLI is a dataset that was translated SNLI (Bowman et al., 2015), XNLI (Conneau et al., 2018) and MNLI (Williams et al., 2018) into Korean. KorSTS is a dataset that was translated from Semantic Text Similarity (STS) (Cer et al., 2017). Hate Speech (Moon et al., 2020) is data that classifies whether a given sentence is hate speech, and classifies the type of hate speech. Table 11 shows the performance of KE-T5 in Korean benchmarks, and the overall performance is good.

| | NIKL CoLA | NSMC | Question-pair | |
|---|---|---|---|---|
| size | Mattew's | Acc. | F1 | Acc. |
| small | -3.72 | 87.90 | 87.90 | 91.5 |
| base | 12.51 | 88.95 | 93.70 | 91.49 |
| large | 13.31 | 89.70 | 89.74 | 92.52 |
| | KorNLI | KorSTS | | Hate Speech |
| size | Acc. | Pearson | Spearman | Acc. |
| small | 73.41 | 78.19 | 77.9 | 60.65 |
| base | 78.67 | 80.02 | 79.73 | 64.14 |
| large | 79.76 | 83.65 | 83.25 | 62.82 |

Table 11: Performance of KE-T5 on Korean NLP tasks.

### A.6 Korean Summarization tasks

NIKL summarization data[2] is summarization data published by the National Institute of Korean Language(NIKL) Republic of Korea. It is divided into a summary split and a topic split. The summary split is built by human-handed summarizing articles.The topic split is data that concatenates the topic sentences selected by a person in an article. Table 12 shows the Korean Summarization performance. Both summary split and topic split show high performance, but the performance of topic split is higher than summary split.

| | summary | | topic | |
|---|---|---|---|---|
| size | Rouge-1 | Rouge-2 | Rouge-1 | Rouge-2 |
| small | 38.85 | 18.65 | 48.79 | 32.51 |
| base | 40.86 | 19.58 | 50.71 | 35.43 |
| large | 40.54 | 20.04 | 55.52 | 37.72 |

Table 12: Performance of KE-T5 on NIKL summarization data.

### A.7 CNN/DM summarization

CNN Daily Mail summarization (See et al., 2017) is the task of summarizing a given document. As shown in Table 13, KE-T5 has good performance in the English summarization task.

| size | Rouge-1 | Rouge-2 |
|---|---|---|
| small | 37.94 | 17.90 |
| base | 37.84 | 15.38 |
| large | 40.15 | 17.78 |

Table 13: Performance of KE-T5 on CNN Daily Mail data.

## B Detailed settings for experimentation

All experiments were trained and validated with the same hyper parameter setting. Knowledge was truncated so that the number of tokens did not exceed 64, and the dialogue context was truncated to 256. Due to GPU memory limitations, knowledge candidates were divided into 32 sized mini batches. In Eq. 3, the knowledge weight $\lambda$ was set to 0.95. The adam optimizer was used for training, and epsilon was set to 5e-4, beta 1 to 0.9, and beta 2 to 0.98. The learning rate is 5e-4, and an invert square root is used as a learning rate scheduling method. The learning scheduler decay is set to 0.5, and the warm up steps is set to 5000. One NVIDIA V100 32GB GPU was used for training, and it took about 1 day to learn. Beam search was used for inference, and the beam size was set to 4 and the length penalty was set to 0.65.

## C Additional Samples

The below tables show samples generated by the proposed model on the KoWoW dataset. Table 14 and Table 15 show samples generated by the proposed model from four topics. Table 14 shows responses generated using gold knowledge, and Table 15 shows responses generated using predicted knowledge.

| Topic | Language Pair (Knowledge-Response) | Examples | |
|---|---|---|---|
| Sled dog (seen) | En-Ko | **Apprentice:** | 북쪽에서 일하는 개들이겠죠? (I guess in the north they are working dogs huh?) |
| | | **Gold Knowledge:** | Sled dogs were important for transportation in arctic areas, hauling supplies in areas that were inaccessible by other methods. |
| | | **Response generated:** | 허스키는 북부 지역에서 교통을 위해 매우 중요합니다. (Huskies are important for transportation in arctic areas.) |
| | Ko-En | **Apprentice:** | I guess in the north they are working dogs huh? |
| | | **Gold Knowledge:** | 썰매 개는 북극 지역의 운송에 중요했으며 다른 방법으로는 접근 할 수없는 지역에서 물품을 운반했습니다. (Sled dogs were important for transportation in arctic areas, hauling supplies in areas that were inaccessible by other methods.) |
| | | **Response generated:** | I am not sure, but I know that sled dogs are important for transportation in Arctic areas. |
| | Ko-Ko | **Apprentice:** | 북쪽에서 일하는 개들이겠죠? (I guess in the north they are working dogs huh? |
| | | **Gold Knowledge:** | 썰매 개는 북극 지역의 운송에 중요했으며 다른 방법으로는 접근 할 수없는 지역에서 물품을 운반했습니다.(Sled dogs were important for transportation in arctic areas, hauling supplies in areas that were inaccessible by other methods.) |
| | | **Response generated:** | 예, 그들은 북극 지역의 운송에 큰 역할을했습니다. (Yes, they played a big role in transport in the Arctic areas.) |
| Broken heart (unseen) | En-Ko | **Apprentice:** | 정말 흥미롭군요. 상심에 대해 알고있는 다른 사실은 무엇입니까? (That is really interesting, what other facts do you know about broken hearts?) |
| | | **Gold Knowledge:** | The concept is believed to be universal, with many cultures using the same words to describe both physical pain and the feelings associated with relationship loss. |
| | | **Response generated:** | 예, 실제로 많은 문화권에서 동일한 단어로 신체적 통증과 관계 상실의 느낌을 설명합니다. (Yes, in fact, many cultures use the same words to describe the feeling of physical pain and loss of relationships.) |
| | Ko-En | **Apprentice:** | That is really interesting, what other facts do you know about broken hearts? |
| | | **Gold Knowledge:** | 이 개념은 보편적 인 것으로 여겨지며, 많은 문화권에서 육체적 고통과 관계 상실과 관련된 감정을 설명하기 위해 동일한 단어를 사용합니다. (The concept is believed to be universal, with many cultures using the same words to describe both physical pain and the feelings associated with relationship loss.) |
| | | **Response generated:** | I know that the concept is considered universal and is used in many cultures. |
| | Ko-Ko | **Apprentice:** | 정말 흥미롭군요. 상심에 대해 알고있는 다른 사실은 무엇입니까? (That is really interesting, what other facts do you know about broken hearts?) |
| | | **Gold Knowledge:** | 이 개념은 보편적인 것으로 여겨지며, 많은 문화권에서 육체적 고통과 관계 상실과 관련된 감정을 설명하기 위해 동일한 단어를 사용합니다. (The concept is believed to be universal, with many cultures using the same words to describe both physical pain and the feelings associated with relationship loss.) |
| | | **Response generated:** | 상심의 개념은 보편적 인 것으로 여겨지며 많은 문화권에서 그 개념을 사용합니다. (The concept of a broken heart is considered universal and many cultures use it.) |

Table 14: Examples of gold knowledge and generated responses (blue text indicates English translation of Korean sentences).

| Topic | Language Pair (Knowledge-Response) | Examples | |
|---|---|---|---|
| Purple (seen) | En-Ko | **Apprentice:** | 오,멋지네요. 자주색이 역사적용도가 있었는지 알고 계시나요? (Oh that's cool, do you know if purple had any historical uses?) |
| | | **Knowledge predicted:** | Purple was the color worn by Roman magistrates; it became the imperial color worn by the rulers of the Byzantine Empire and the Holy Roman Empire, and later by Roman Catholic bishops. |
| | | **Response generated:** | 로마 치안관은 보라색을 입었고 비잔틴 제국과 신성 로마 제국의 통치자들이 입었습니다. (Roman magistrates wore purple, and purple was worn by the rulers of the Byzantine and Holy Roman Empires.) |
| | Ko-En | **Apprentice:** | Oh that's cool, do you know if purple had any historical uses? |
| | | **Knowledge predicted:** | 보라색과 비슷하지만 가시 광선 스펙트럼에 자체 파장을 가진 분광 색인 보라색과 달리 보라색은 적색과 청색을 결합한 합성 색입니다. (It is similar to violet, but unlike violet, which is a spectral color with its own wavelength on the visible spectrum of light, purple is a composite color made by combining red and blue.) |
| | | **Response generated:** | I'm not sure, but it is a combination of red and blue. |
| | Ko-Ko | **Apprentice:** | 오, 멋지네요. 자주색이 역사적 용도가 있었는지 알고 계시나요? (Oh that's cool, do you know if purple had any historical uses?) |
| | | **Knowledge predicted:** | 자주색은 로마 치안 관들이 입는 색이었습니다. 그것은 비잔틴 제국과 신성 로마 제국의 통치자들과 나중에 로마 카톨릭 주교들이 입는 황실 색이되었습니다. (Purple was the color worn by Roman magistrates; it became the imperial color worn by the rulers of the Byzantine Empire and the Holy Roman Empire, and later by Roman Catholic bishops.) |
| | | **Response generated:** | 로마 치안관이 입는 색이었습니다. 그것은 비잔틴 제국과 신성 로마 제국의 통치자들이 입는 황실 색이되었습니다. (Roman magistrates wore purple color. It became the imperial color worn by the rulers of the Byzantine Empire and the Holy Roman Empire.) |
| Chihuahua (unseen) | En-Ko | **Apprentice:** | 저는 치와와를 좋아합니다. 두 마리를 키우고 있습니다. (I love chihuahuas, I have 2 of them.) |
| | | **Knowledge predicted:** | The Chihuahua is the smallest breed of dog and is named after the state of Chihuahua in Mexico. |
| | | **Response generated:** | 치와와는 가장 작은 품종이며 멕시코의 Chihuahua의 이름을 따서 명명되었습니다. (The Chihuahua is the smallest breed of dog and is named after the state of Chihuahua in Mexico.) |
| | Ko-En | **Apprentice:** | I love chihuahuas, I have 2 of them. |
| | | **Knowledge predicted:** | 치와와는 가장 작은 품종으로 멕시코의 치와와 주 이름을 따서 명명되었습니다. (The Chihuahua is the smallest breed of dog and is named after the state of Chihuahua in Mexico.) |
| | | **Response generated:** | I love chihuahuas. They are the smallest breed in Mexico. |
| | Ko-Ko | **Apprentice:** | 저는 치와와를 좋아합니다. 두 마리를 키우고 있습니다. (I love chihuahuas, I have 2 of them.) |
| | | **Knowledge predicted:** | 치와와는 가장 작은 품종으로 멕시코의 치와와 주 이름을 따서 명명되었습니다. (The Chihuahua is the smallest breed of dog and is named after the state of Chihuahua in Mexico.) |
| | | **Response generated:** | 치와와는 가장 작은 품종의 개입니다. 치와와는 멕시코의 치와와와 주 이름을 따서 명명되었습니다. (The Chihuahua is the smallest breed of dog. The Chihuahua is named after the state of Chihuahua in Mexico.) |

Table 15: Examples of predicted knowledge and generated responses (blue text indicates English translation of Korean sentences).