CECADA: Cause-Effect Conjunctive Adverb-based Data Augmentation Method in Low-Resource Knowledge-Grounded Dialogue

Taesuk Hong[1,2], **Dongkyu Lee**[2,3], **Janghoon Han**[2], **Stanley Jungkyu Choi**[2], **Jungyun Seo**[1,2]

[1]Sogang University, [2]LG AI Research,
[3]Hong Kong University of Science and Technology
lino.taesuk@gmail.com, dleear@cse.ust.hk,
{janghoon.han, stanleyjk.choi}@lgresearch.ai, seojy@sogang.ac.kr

Abstract

A large body of research has investigated in drawing an interesting and engaging conversation with a user, and one of the effort is incorporating a knowledge in generation. Accordingly, a growing need for knowledge-incorporated dialogue dataset has gained attention. However, coupling a response and a knowledge in a context-specific manner is laborious and challenging, and hence the amount of data collected is often insufficient. In this light, this study proposes a simple but effective data augmentation method by leveraging the linguistic features of cause-effect conjunctive adverbs in a natural language; we reformulate a plain document with a cause-effect conjunctive adverb as a knowledge-grounded dialogue data instance. With the proposed data augmentation technique, we observe a marked gain in generalization of a model in both knowledge selection and knowledge-grounded dialogue generation. In particular, the proposed method demonstrates its effectiveness in a low-resource setting in which dialogue systems generally suffer from.

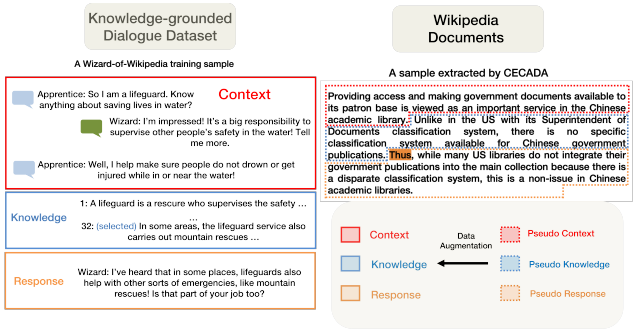


Figure 1: Description of our proposed data augmentation method. Here, cause-effect conjunctive word, '`thus`', captures a knowledge-grounded dialogue data-like structure in a document. This instance is added to the augmentation dataset.

1 Introduction

Neural language models have recently demonstrated remarkable performances across a variety of tasks, one of which is dialogue response generation. Dialogue systems are generally divided into two types; the first is task-oriented that seeks to achieve a specific purpose through dialogue, and the other is *general-purpose chit-chat dialogue model*. One common challenge faced by chit-chat systems is that models often generate a dull and uninformative responses, and such responses fail to draw engaging and interesting interactions with a user (Zhao et al. 2020b; Li et al. 2016b). Therefore, a large body of research has been conducted to mitigate the issue (Rashkin et al. 2019; Smith et al. 2020; Zhao et al. 2020b), and one of the strategies is incorporating knowledge as a conditioning variable to a language model (Ghazvininejad et al. 2018; Dinan et al. 2019; Li et al. 2022; Zhou et al. 2022).

A knowledge-grounded dialogue model conditions on an external knowledge, such as documents, tables, and pictures, when generating a response to a user. Therefore, a training dataset is required to be equipped with not just dialogue turns, but with external knowledge to be grounded, such as Dinan et al. (2019) and Feng et al. (2020). However, annotating such data is time-consuming and requires a sophisticated tool to label; each response is required to be coupled with a specific knowledge that caters to a given context (dialogue history). Thus, the amount of knowledge-grounded dialogue data tends to be limited (Dinan et al. 2019; Li et al. 2020; Zhao et al. 2020a).

In this study, we propose a simple automatic data augmentation method to alleviate the data shortage problem of a knowledge-grounded dialogue system. The proposed method automatically collects data from a large corpus using *linguistic features*. The intuition is rooted from an observation; each example of knowledge-grounded dialogue data consists of the following three elements: 1) dialogue context, 2) grounding knowledge, and 3) a response that reflects the knowledge and fits the context. We observe that such combination is found in a plain text **when a Cause-Effect Conjunctive Adverb (CECA) is present**. Cause-effect conjunctive adverbs, such as `therefore`, `thus`, and `hence`, signify that the connection between two sentences has a cause-effect relationship. This structure resembles to the knowledge-grounded dialogue data which has

Accepted to Workshop on Knowledge Augmented Methods for Natural Language Processing, in conjunction with AAAI 2023.

been considered laborious. Therefore, the proposed data augmentation method exploits this structural similarity and constructs pseudo knowledge-grounded dialogue data from a large unlabeled corpus, which the process is illustrated in Figure 1.

The contribution of this study is summarized as follows:

- We propose a simple data augmentation method, called CECADA, for knowledge-grounded dialog system that removes the need of human annotation.
- Empirical results demonstrate that the proposed data augmentation improves performance of a model in both 1) the knowledge selection task and 2) the knowledge-grounded response generation task.
- The proposed augmentation technique enables a model to be robust even when trained in a low-resource training environment.

2 Approach

2.1 Task Description

Let $D = \{(c_i, k_i, r_i)\}_i^n$ be a dialogue corpus with size n, where c, k and r denote a context, a knowledge, and a response respectively. In neural knowledge-grounded dialogue system, we are interested in finding two functions: 1) a knowledge selection function f with parameter θ that selects a context-specific knowledge within a knowledge pool, and 2) a knowledge-grounded response generation function g with parameter ϕ that maps a context and a knowledge to a response that is fluent, context-appropriate and reflects the given knowledge, $f_\theta : \mathcal{C} \times \mathcal{K} \rightarrow [0, 1]$, $g_\phi : \mathcal{C} \times \mathcal{K} \rightarrow \mathcal{R}$. In this paper, the goal is to *improve the generalization of the two functions with a data augmentation technique.*

2.2 Cause-Effect Conjunctive Adverb-based Data Augmentation (CECADA)

Definition 1 *(CECA) We define CECA as any conjunctive adverb that connects two sentences with a cause-effect relation with the condition that the conjunctive adverb comes after the cause and before the effect.*

When a CECA appears in a paragraph, the sentence following the CECA is grounded on the sentence immediately preceding it. Except for these two cause-effect sentences, the rest of the preceding sentences act as a context in which the cause-effect sentences fit in. Inspired by the fact that it is similar to the structure of knowledge-grounded response generation data, we propose a data augmentation method as follows:

- **Step 1. Select a Domain of Interest**: Selecting the right domain is essential, as the out-of-domain augmented data can impede the training of the targeting domain. We select Wikipedia web pages as a collection target for data augmentation. Specifically, we use WikiText (Merity et al. 2016), and English Wikipedia-dump of the '20200501.en'.

https://huggingface.co/datasets/wikipedia. The original datatset can be found in Wikimedia Foundation: https://dumps.wikimedia.org

Cause-Effect Conjunctive Adverb	Number of Examples
`therefore`	91,407
`thus`	134,957
`hence`	21,785
`consequently`	25,429
`accordingly`	12,534
`henceforth`	1,255
Total	287,367

Table 1: The statistics on the collected data using our proposed method, CECADA.

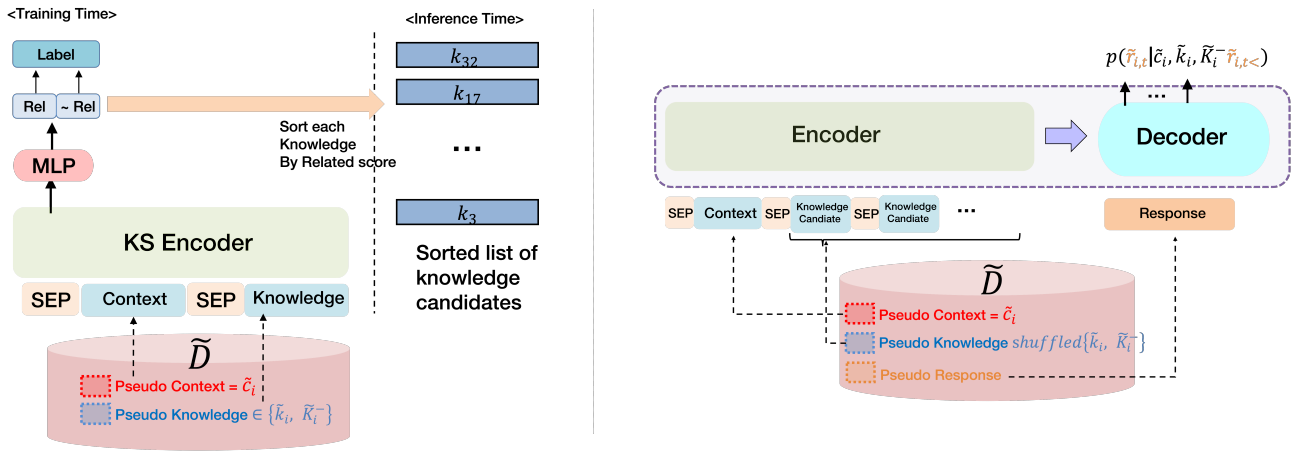
- **Step 2. Define List of Cause-Effect Conjunctive Adverb (CECA)**: A CECA could be any word that represents a cause and effect, yet we confine such vocab to the ones listed in Table 1.
- **Step 3. Collection Process**: We augment the data so that the augmented data correspond to the structure of an example of knowledge-grounded dialogue data. When one of the words in the CECA list exactly matches in a corpus, we augment an example; the sentence after the CECA is regarded as a response (pseudo response), the sentence immediately before the CECA is considered as knowledge (pseudo knowledge), and the five sentences before the pseudo knowledge are considered as dialogue context (pseudo context). The pseudo context, pseudo knowledge and pseudo response make up an augmented training instance. Lastly, the CECA and commas around the CECA are removed, so that a model does not overfit to a pattern.

Following the above steps, we have collected 287,367 examples to be augmented. Table 1 shows the count of occurrences of each CECA in the corpus, thus forming a CECADA example. We denote the augmented collection of data $\tilde{D} = \{(\tilde{c}_i, \tilde{k}_i, \tilde{r}_i)\}_i^M$ where $\tilde{c}_i, \tilde{k}_i, \tilde{r}_i$ refers to a pseudo context, a pseudo knowledge, and a pseudo response, and M is the total number of the augmented data.

2.3 Training a Model with CECADA

As we construct *pseudo* training examples, our strategy for both the knowledge selection and the knowledge-grounded generation task is to adopt *curriculum learning-like* fine-tuning. Before fine-tuning on a downstream task data, we train the model with the CECADA dataset. Once the training is done on CECADA dataset, we further fine-tune the model on a downstream task dataset. In this section, we describe in detail with training objectives for each model in knowledge selection task and the knowledge-grounded response generation task.

Utilizing CECADA in Knowledge Selection One common strategy in training knowledge selection model is to fine-tune a large pretrained model, such as RoBERTa-large (Zhuang et al. 2021) for binary classification model. Given the dialogue context and one of the knowledge in a set of candidates, the model is trained to predict the relatedness between them. At inference time we sort knowledge candidates by ranking the model's output relatedness scores. The model is depicted in Figure 2a.



(a) Knowledge Selection Model

(b) Knowledge-Grounded Generation Model

Figure 2: our knowledge selection model in (a) takes the pseudo context and one of the pseudo knowledge from the augmented set of knowledge candidates as input, then classifies if the knowledge is related – or not related – to the given context by the binary scores. The generation model in (b) takes the form of sequence-to-sequence. The pseudo context and knowledge candidates are encoded in a pre-trained encoder, then the model is trained to generate a pseudo response given encoded vector.

Here, in training the model with the CECADA dataset, a set of negative examples are required for the training and is constructed via retrieval and random sampling.

Retrieving The concatenated sentences of $\tilde{c}_i$ and $\tilde{k}_i$ are encoded with a sentence-transformers encoder (Reimers and Gurevych 2019), which the mapping process is denoted as h. The output vector, $h([\tilde{c}_i; \tilde{k}_i])$, is then compared to every pre-encoded Wikipedia corpus, and the objective is to find the most similar negative examples for training.

$$S_i^- = \underset{\substack{|S_i^-|=o \\ s_j \in S}}{\arg\max} \cos(h([\tilde{c}_i; \tilde{k}_i]), h(s_j)) \tag{1}$$

where $\cos$, S and S_i^- refer to the cosine similarity, the set of all Wikipedia paragraphs, and the shortlisted negative paragraphs respectively. We confine the number of negative samples for each instance to o.

Random Sampling Each element in the retrieved output is a paragraph. To align the format of the negative samples to that of knowledge, we randomly sample p sentences as negative knowledge, denoted as $\tilde{k}_i^-$, from each retreived paragraph $s_i \in S_i^-$; the set of the sampled negative knowledge has the size of $o \times p$ and is termed $\tilde{K}_i^-$. We add the negative knowledge for each pseudo instance, hence CECADA dataset being $\tilde{D} = \{(\tilde{c}_i, \tilde{k}_i, \tilde{r}_i, \tilde{K}_i^-)\}_i^M$.

To be specific on the training stage on the CECADA dataset for the knowledge selection model, we align the pseudo data format to Wizard-of-Wikipedia (Dinan et al. 2019) sample. A training instance takes one of the pseudo knowledge candidates given the pseudo context ($\tilde{c}_i$), and forms an input for the model by concatenating them with the separator token in between. The encoded hidden state of the first token at the last layer is projected through the binary classification layer. The training objective in finding the f is minimizing the Cross-Entropy loss as shown in the Equation 2. The trained knowledge selection model with the CECADA dataset is consecutively fine-tuned with the downstream task dataset.

$$\begin{aligned}\mathcal{L}_{ks} = \mathbb{E}_{\{(\tilde{c}_i, \tilde{k}_i, \tilde{K}_i^-)\}_i^B \sim \tilde{D}}[-\log g_\phi(\tilde{k}_i, \tilde{c}_i) \\ + \log 1 - g_\phi(\tilde{k}_i^-, \tilde{c}_i)], \text{ where } \tilde{k}_i^- \in \tilde{K}_i^-\end{aligned} \tag{2}$$

Utilizing CECADA in Knowledge-Grounded Response Generation The knowledge-grounded response generation shares the same training scheme as the knowledge selection task. We chose BART-large (Lewis et al. 2020) for the large pre-trained model.

In training the generation model with the CECADA dataset, we now include the pseudo response when aligning the pseudo data format to the downstream task sample. Figure 2b depicts the training process. We concatenate $\tilde{c}_i$ and all the knowledge candidate, the pseudo knowledge $\tilde{k}_i$ and the set of negative knowledge sampled $\tilde{K}_i^-$. The order of knowledge sentences is randomly shuffled, and each sentence is divided with a special token. A model is trained to generate the pseudo response $\tilde{r}_i$ by minimizing the negative log-likelihood.

$$\begin{aligned}\mathcal{L}_{NLL} = \mathbb{E}_{\{(\tilde{c}_i, \tilde{k}_i, \tilde{r}_i, \tilde{K}_i^-)\}_i^B \sim \tilde{D}} \\ [-\frac{1}{T}\sum_t^T \log P(\tilde{r}_{i,t}|\tilde{c}_i, \tilde{k}_i, \tilde{K}_i^-, \tilde{r}_{i,<t}; \theta)]\end{aligned} \tag{3}$$

where $\tilde{r}_{i,t}$ refers to the token to be generated at t-th time step and $\tilde{r}_{i,<t}$ is the preceding tokens at time step t. After

WoW	Train	Valid	test-seen	test-unseen
Utterances	166,787	17,715	8,715	8,782
Dialogues	18,430	1.948	965	968
Topics	1,247	599	533	58

Table 2: The statistics on the Wizard-of-Wikipedia (WoW) dataset. We denote train dataset as D.

training the model with the CECADA dataset, we fine-tuned the model with a downstream task dataset to adapt to the targeting domain with the same loss as in Equation 3.

3 Experiments

3.1 Dataset

Wizard-of-Wikipedia (WoW) (Dinan et al. 2019) is a knowledge-grounded dialogue benchmark consisting of a one-to-one conversation between an apprentice who wants to learn about a specific topic in an open-domain environment and a wizard who responds by referring to a sentence from a Wikipedia document. For each wizard's turn of the dialogue, the wizard has to select one of the preferring knowledge sentence among the knowledge candidates then to respond which reflects the selected knowledge.

The data statistics are shown in Table 2. We hereafter denote D – defined in Section 2.1 Task Description – as the training dataset of Wizard-of-Wikipedia. Note that $D = (c_i, k_i, r_i, K_i^-)_i^n$ contains the set of negative knowledge candidates, K_i^-. The models we present in the experiments are all trained with this downstream task dataset, D.

3.2 Experimental Setup

We demonstrate our presented models for clarity. Two models below are used in the knowledge selection task:

- **KS** has a structure depicted in Figure 2a as a binary classification model that predicts the relatedness score of given knowledge and the dialogue context. However, for the purpose of comparison, it *is not trained* with the CECADA dataset. The model is only fine-tuned with the WoW dataset, D.
- **KS**$^+$ takes the same structure as the KS. However, it *is trained with* the CECADA dataset and then fine-tuned with the WoW dataset.

The below four models are used in knowledge-grounded generation task:

- **BART** is only fine-tuned with the WoW dataset to generate a response given dialogue context and knowledge candidates with negative log-likelihood (NLL) objective. The model's structure can be seen in Figure 2b. However, this model *is not trained* with the CECADA dataset. In the inference phase, the model takes concatenated sequence of dialogue context (c_i) and a set of knowledge candidates ($k_i \cup K_i^-$) as the input. Here, the set of knowledge candidates are shuffled.
- **BART**$^+$ differs from the above BART in that it is trained with the CECADA dataset and then fine-tuned with the WoW dataset. The input to the model in the inference phase is the same as one with BART.
- **KS-BART** first receives the scored list of knowledge candidates by the KS model. Then, starting from the top-most similar knowledge in the list to the descending order, the input for BART is formed by concatenating knowledge sentence to the dialogue context until it reaches or exceeds the max length of the model. BART is only fine-tuned with the WoW dataset.
- **KS**$^+$**-BART** has a difference between the KS-BART in that the KS$^+$ model provides the scored list of knowledge candidates to the fine-tuned BART. Here, the generation model is only fine-tuned with the WoW dataset.

3.3 Training Details

In training the knowledge selection model, KS$^+$, with the CECADA dataset, we trained with a batch size, B, of 16. The learning rate was set to $2e-5$, and the model was trained for three epochs with AdamW optimizer (Loshchilov and Hutter 2019) whose adam epsilon was $1e-8$. The concatenated input for the knowledge selection model was truncated from the foremost part of the context, with the model's max length of 256. The o and p are 2 and 3, respectively. Fine-tuning steps that follow the training with the CECADA dataset, the hyper-parameter setups are the same as above.

In training the generation model, BART$^+$, with the CECADA dataset, the model was trained with a batch size of 8. The model's max length was set to 256, and each knowledge that exceeds the length of 64 was truncated. Given the context and the sequence of knowledge candidates, if the whole sequences are longer than the max length of the model, then we truncated the knowledge sequence to fit in. The model was trained with AdamW, a learning rate of $5e-5$, adam epsilon, $\epsilon = 1e-8$, training epoch of 3. Fine-tuning with the downstream task dataset follows the same setting as the one with training with the CECADA dataset. In the inference time, generation was performed with greedy decoding with sampling. It took six hours and five hours to train for one epoch with two RTX3090 GPUs in training the knowledge selection model and the generation model, respectively. We used the Huggingface Transformers library for both to download the model parameter and train the model.

3.4 Evaluation Metrics

For the knowledge selection task, we measure Recall@{1, 2, 5, 10} scores where Recall@k is the ratio of correct predictions that the label knowledge – the one that the wizard selected in the dataset – was present within the top-k list of the sorted ranks. In the response generation task, we evaluate BLEU (Papineni et al. 2002) and ROUGE (Lin 2004) scores, both of which the lexical overlaps are used between the predicted response and the label response. DIST1 (Li et al. 2016a) calculates the model's diversity of vocabulary use in generating the response. Finally, we report the perplexity score.

https://github.com/huggingface/transformers

Model	R@1	R@2	R@5	R@10
TF-IDF	11.00	18.86	37.44	56.43
Transformer MemNet	22.5	-	-	-
KS	20.29	28.44	44.17	59.53
KS^+	**25.47**	**36.32**	**55.47**	**71.78**

Table 3: knowledge selection task performance on Wizard-of-Wikipedia (WoW) test-seen dataset. R@k refers to Recall@k. For some of the scores that the baseline work does not report are marked as '-'. The units are in %p.

3.5 Baselines

Here we briefly explain each baseline model in knowledge selection and knowledge-grounded generation task.

- **Transformer MemNet** (Dinan et al. 2019): Along with releasing the Wizard-of-Wikipedia dataset, the authors proposed a baseline model which encodes each knowledge sentence and the dialogue context using Transformer encoder (Vaswani et al. 2017). Each encoded knowledge vector and the context vector are compared using dot-product attention. The model outputs the sorted list of knowledge candidates based on the similarity scores.
- **SKT** (Kim, Ahn, and Kim 2020): proposes a sequential knowledge transformer (SKT) that keeps track of the prior and posterior distribution over knowledge by using a latent variable. This technique helps maintain the temporal features of the dialogue in selecting knowledge, and the knowledge selection results are fed into the generation model.
- **LCK** (Zhao et al. 2020a): Inspired by the nature of chit-chat dialogue, its generation model splits separate probability models with respect to the language model, context, and knowledge. The encoders and the decoder used GRUs (Cho et al. 2014) in generating the response.

3.6 Result of Knowledge Selection

Table 3 shows the knowledge selection performance of various models. The TF-IDF model at the first row of the table measures tf-idf based similarity scores between the dialogue history and the list of knowledge candidates to sort relevant knowledge in descending order. KS^+ scored the highest in every measure R@1, R@2, R@5, and R@10, and improved significantly upon the KS by 5.18%p, 7.88%p, 11.3%p, 12.25%p respectively. This fact proves that our data augmentation method contributes to improving the knowledge selection task of the knowledge-grounded dialogue system. KS^+ also outperformed the baseline Transformer MemNet by 2.97%p in R@1 score.

3.7 Result of Knowledge-Grounded Response Generation Task

Table 4 presents the evaluation result on test-seen dataset of WoW in knowledge-grounded response generation task. Compared to the baseline model BART, our proposed model, $BART^+$, shows a slight performance improvement in every metrics except for the perplexity score. For KS^+-BART, the BLEU1 score was 25.09%p, which showed an improvement of 3.61%p compared to the baseline (BART). This infers that the result of the knowledge selection can considerably help boosting the knowledge-grounded response generation task. The KS^+-BART outperforms even the high-performing model SKT and LCK by 1.95%p in ROUGE1 and 3.29%p in the BLEU1 metric. To measure the effectiveness of the CECADA, we compared the KS^+-BART to the KS-BART model, whose knowledge selection model was not trained by the CECADA dataset. The KS^+-BART improved upon the KS-BART in every metric except the DIST1 by 0.59%p, 0.55%p, 0.96%p, 0.77%p in BLEU1, BLEU4, ROUGE1, and perplexity score, respectively. Therefore, the CECADA contributes to helping increase the knowledge-grounded generation capability.

3.8 Effect of CECADA in Low-resource Environment

Here, we set the low-resource environment as the knowledge-grounded dialogue datasets generally encounter the low-resource problem. Table 5 shows the decreased performance rate of BLEU1 score from where the entire data was used in fine-tuning to the ones where the fine-tuning dataset becomes proportionally reduced. Considering the Wizard-of-Wikipedia dataset belongs to the domain of 'dialogue,' we additionally performed the CECADA in the 2019 Reddit corpus from Pushshift (Baumgartner et al. 2020). The model, which is trained with the Reddit-based CECADA dataset, refers to the $BART^+_{Redd}$. $BART^+_{Wiki}$ is the same model as the $BART^+$. Note that every model in Table 5 does not take any knowledge selection result of a separate knowledge selection model.

When the training set is reduced, every model starts to degrade, except for the only case of $BART^+_{Wiki}$ where the training dataset was half the entire dataset. When the training data is reduced to 25% of the entire data, the decrease rate plummeted in the BART model to -27.19%. However, compared to the baseline model, the models that took advantage of the CECADA maintained the performance more robustly. In particular, The decrease rate of $BART^+_{Redd}$ is remarkably low of -2.16%, which is lower than the baseline model with a 25.03%p difference. When the size of the training model highly drops to 6.25%, the difference between the baseline model and $BART^+_{Redd}$ is 25.28%p. The average performance decrease rate in the four size decrease settings are -21.52%, -12.71%, -4.67% in BART, $BART^+_{Wiki}$, and $BART^+_{Redd}$, respectively. It is noteworthy that training the CECADA dataset was not significantly helpful in the knowledge-grounded response generation task, as in Table 4, when the entire training dataset was used for fine-tuning. However, our proposed method becomes essential in making the model robust to the low-resource setting. As the CECADA dataset can fill the gap that attributes to the reduced part of the WoW dataset, this result supports our hypothesis that the CECADA dataset that is automatically augmented resembles the knowledge-grounded dataset, which is manually annotated.

Model	CECADA	B1	B4	R1	DIST1	PPL
SKT (Kim, Ahn, and Kim 2020)	X	-	-	19.3	-	52.0
LCK (Zhao et al. 2020a)	X	21.8	5.5	-	-	**23.0**
BART	X	21.48	4.21	18.07	6.87	49.24
BART^{+}	O	21.70	4.22	18.1	7.07	49.39
KS-BART	X	24.50	5.52	20.29	**9.34**	30.47
KS^{+}-BART	O	**25.09**	**6.07**	**21.25**	9.33	29.70

Table 4: Knowledge-grounded response generation task performance on Wizard-of-Wikipedia (WoW) test-seen dataset. The units are in %p. B1, B4, R1, DIST1, PPL refer to the BLEU1, BLEU4, ROUGE1, Diversity-1, and perplexity score, respectively. In the CECADA column, the model with the 'O' indicates that it gained help from the CECADA in training, and 'X' otherwise.

Model	CECADA	FULL	50%	25%	12.5%	6.25%	Average
BART	X	21.48	20.89 (-2.75%)	15.64 (-27.19%)	16.89 (-21.37%)	14.01 (-34.78%)	-21.52%
BART$^{+}_{Wiki}$	O	21.70	21.72 (+0.09%)	18.47 (-14.88%)	17.75 (-18.20%)	17.83 (-17.83%)	-12.71%
BART$^{+}_{Redd}$	O	**22.20**	**21.41 (-3.56%)**	**21.72 (-2.16%)**	**21.41 (-3.56%)**	**20.09 (-9.50%)**	**-4.67%**

Table 5: The evaluation result on WoW test-seen dataset in a low-resource training environment. 50%, 25%, 12.5%, 6.25% in the first row of the table are the degree to which the dataset is reduced. CECADA column indicates whether the model was trained with the CECADA dataset before fine-tuning. The units of performance are in %p. The number in parenthesis is performance decrease rates from when the model was fine-tuned with the entire dataset to when the model is fine-tuned with the reduced dataset size.

4 Related Works

4.1 Knowledge-Grounded Dialogue Systems

Research on dialogue systems recently has been actively conducted to deliver information desired by users by grounding on the external knowledge source in the form of dialogue (Li et al. 2022; Zhou et al. 2022; Tuan et al. 2022). This external knowledge refers to the information or data that the dialogue model did not train in advance. Task-oriented dialogue systems that utilize external knowledge vary in their form of knowledge as DB (Eric et al. 2019), a description sentence for specific venues (Kim et al. 2021), images (Kottur et al. 2021; Hori et al. 2022), and documents (Feng et al. 2020; Reddy, Chen, and Manning 2019). In the case of Chit-chat dialogue systems, external knowledge can also be provided in documents that focus on specific domains (Zhou, Prabhumoye, and Black 2018; Moghe et al. 2018) or those that cover multi-domains (Dinan et al. 2019; Komeili, Shuster, and Weston 2022). Recently, a knowledge-grounded dialogue dataset that provides a knowledge graph as external knowledge with each dialogue turn is linked to its referring node is released (Moon et al. 2019). In addition, Zhang et al. (2018) is a dataset that provides a person's characteristics as a sentence and aims to chit-chat based on the characteristics.

4.2 Data Augmentation

Data augmentation is a prevalent method of collecting data without explicit collection to solve a data-insufficient problem for training data (Feng et al. 2021). The data augmentation method can be divided into rule-based, example interpolation, and model-based methods (Feng et al. 2021). Regardless of the fields, data augmentation is actively used, such as machine translation (Wang et al. 2018), question answering (Longpre et al. 2019), summarization (Fabbri et al. 2021), etc. Study of data augmentation that is involved in dialogue systems, various ways to improve the performance using data augmentation have been proposed. A method to increase the number of responses to train by splitting the dialogue is widespread (Kummerfeld et al. 2019). Han et al. (2021) pointed out that the dialogue models should train augmented fine-grained examples to capture the relationship inside the dialogue as the turn level. Whang et al. (2021) devised novel tasks that can be derived within the dialogue data to improve the model by multi-tasking. Our study takes advantage of the linguistic features of conjunctive adverbs to augment data that shares the parallel structure of the knowledge-grounded dialogue system.

5 Conclusion

The manual annotation process of knowledge-grounded dialogue data suffers from insufficient data problems because annotating each grounding knowledge to the response is time-consuming. We proposed a simple and automatic data augmentation method using the linguistic feature of cause-effect conjunctive adverbs appearing in documents. Through this method, the performance of the knowledge selection task where the model finds the most appropriate knowledge to be referenced given a dialogue context is highly improved. Furthermore, our proposed data augmentation method achieved a slight performance improvement in the knowledge-grounded response generation. In particular, when training the knowledge-grounded dialogue system with the augmented data by the proposed method, we showed that the performance was maintained robustly in a low-resource situation.

6 Acknowledgments

This work was supported by Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2020-0-00368, A Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques)

References

Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; and Blackburn, J. 2020. The Pushshift Reddit Dataset. *CoRR*, abs/2001.08435.

Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724–1734. Doha, Qatar: Association for Computational Linguistics.

Dinan, E.; Roller, S.; Shuster, K.; Fan, A.; Auli, M.; and Weston, J. 2019. Wizard of Wikipedia: Knowledge-powered Conversational Agents. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Eric, M.; Goel, R.; Paul, S.; Sethi, A.; Agarwal, S.; Gao, S.; and Hakkani-Tur, D. 2019. MultiWOZ 2.1: Multi-Domain Dialogue State Corrections and State Tracking Baselines. *arXiv preprint arXiv:1907.01669*.

Fabbri, A.; Han, S.; Li, H.; Li, H.; Ghazvininejad, M.; Joty, S.; Radev, D.; and Mehdad, Y. 2021. Improving Zero and Few-Shot Abstractive Summarization with Intermediate Fine-tuning and Data Augmentation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 704–717. Online: Association for Computational Linguistics.

Feng, S.; Wan, H.; Gunasekara, C.; Patel, S.; Joshi, S.; and Lastras, L. 2020. doc2dial: A Goal-Oriented Document-Grounded Dialogue Dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Feng, S. Y.; Gangal, V.; Wei, J.; Chandar, S.; Vosoughi, S.; Mitamura, T.; and Hovy, E. 2021. A Survey of Data Augmentation Approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 968–988. Online: Association for Computational Linguistics.

Ghazvininejad, M.; Brockett, C.; Chang, M.-W.; Dolan, B.; Gao, J.; Yih, W.-t.; and Galley, M. 2018. A Knowledge-Grounded Neural Conversation Model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Han, J.; Hong, T.; Kim, B.; Ko, Y.; and Seo, J. 2021. Fine-grained Post-training for Improving Retrieval-based Dialogue Systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1549–1558. Online: Association for Computational Linguistics.

Hori, C.; Shah, A. P.; Geng, S.; Gao, P.; Cherian, A.; Hori, T.; Le Roux, J.; and Marks, T. K. 2022. Overview of Audio Visual Scene-Aware Dialog with Reasoning Track for Natural Language Generation in DSTC10.

Kim, B.; Ahn, J.; and Kim, G. 2020. Sequential Latent Knowledge Selection for Knowledge-Grounded Dialogue. In *ICLR*.

Kim, S.; Liu, Y.; Jin, D.; Papangelis, A.; Gopalakrishnan, K.; Hedayatnia, B.; and Hakkani-Tur, D. 2021. "How robust r u?": Evaluating Task-Oriented Dialogue Systems on Spoken Conversations. arXiv:2109.13489.

Komeili, M.; Shuster, K.; and Weston, J. 2022. Internet-Augmented Dialogue Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8460–8478. Dublin, Ireland: Association for Computational Linguistics.

Kottur, S.; Moon, S.; Geramifard, A.; and Damavandi, B. 2021. SIMMC 2.0: A Task-oriented Dialog Dataset for Immersive Multimodal Conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 4903–4912. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Kummerfeld, J. K.; Gouravajhala, S. R.; Peper, J.; Athreya, V.; Gunasekara, C.; Ganhotra, J.; Patel, S. S.; Polymenakos, L.; and Lasecki, W. S. 2019. A Large-Scale Corpus for Conversation Disentanglement. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 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*, 7871–7880. Online: Association for Computational Linguistics.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2016a. A Diversity-Promoting Objective Function for Neural Conversation Models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 110–119. San Diego, California: Association for Computational Linguistics.

Li, J.; Monroe, W.; Ritter, A.; Jurafsky, D.; Galley, M.; and Gao, J. 2016b. Deep Reinforcement Learning for Dialogue Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1192–1202. Austin, Texas: Association for Computational Linguistics.

Li, L.; Xu, C.; Wu, W.; ZHAO, Y.; Zhao, X.; and Tao, C. 2020. Zero-Resource Knowledge-Grounded Dialogue Generation. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 8475–8485. Curran Associates, Inc.

Li, Y.; Peng, B.; Shen, Y.; Mao, Y.; Liden, L.; Yu, Z.; and Gao, J. 2022. Knowledge-Grounded Dialogue Generation

with a Unified Knowledge Representation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 206–218. Seattle, United States: Association for Computational Linguistics.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Longpre, S.; Lu, Y.; Tu, Z.; and DuBois, C. 2019. An Exploration of Data Augmentation and Sampling Techniques for Domain-Agnostic Question Answering. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, 220–227. Hong Kong, China: Association for Computational Linguistics.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Merity, S.; Xiong, C.; Bradbury, J.; and Socher, R. 2016. Pointer Sentinel Mixture Models. arXiv:1609.07843.

Moghe, N.; Arora, S.; Banerjee, S.; and Khapra, M. M. 2018. Towards Exploiting Background Knowledge for Building Conversation Systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2322–2332. Brussels, Belgium: Association for Computational Linguistics.

Moon, S.; Shah, P.; Kumar, A.; and Subba, R. 2019. OpenDialKG: Explainable Conversational Reasoning with Attention-based Walks over Knowledge Graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Rashkin, H.; Smith, E. M.; Li, M.; and Boureau, Y.-L. 2019. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5370–5381. Florence, Italy: Association for Computational Linguistics.

Reddy, S.; Chen, D.; and Manning, C. D. 2019. CoQA: A Conversational Question Answering Challenge. *Transactions of the Association for Computational Linguistics*, 7: 249–266.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992. Hong Kong, China: Association for Computational Linguistics.

Smith, E. M.; Williamson, M.; Shuster, K.; Weston, J.; and Boureau, Y.-L. 2020. Can You Put it All Together: Evaluating Conversational Agents' Ability to Blend Skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2021–2030. Online: Association for Computational Linguistics.

Tuan, Y.-L.; Beygi, S.; Fazel-Zarandi, M.; Gao, Q.; Cervone, A.; and Wang, W. Y. 2022. Towards Large-Scale Interpretable Knowledge Graph Reasoning for Dialogue Systems. In *Findings of the Association for Computational Linguistics: ACL 2022*, 383–395. Dublin, Ireland: Association for Computational Linguistics.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wang, X.; Pham, H.; Dai, Z.; and Neubig, G. 2018. SwitchOut: an Efficient Data Augmentation Algorithm for Neural Machine Translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 856–861. Brussels, Belgium: Association for Computational Linguistics.

Whang, T.; Lee, D.; Oh, D.; Lee, C.; Han, K.; Lee, D.-h.; and Lee, S. 2021. Do Response Selection Models Really Know What's Next? Utterance Manipulation Strategies for Multi-turn Response Selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zhang, S.; Dinan, E.; Urbanek, J.; Szlam, A.; Kiela, D.; and Weston, J. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2204–2213. Melbourne, Australia: Association for Computational Linguistics.

Zhao, X.; Wu, W.; Tao, C.; Xu, C.; Zhao, D.; and Yan, R. 2020a. Low-Resource Knowledge-Grounded Dialogue Generation. In *International Conference on Learning Representations*.

Zhao, X.; Wu, W.; Xu, C.; Tao, C.; Zhao, D.; and Yan, R. 2020b. Knowledge-Grounded Dialogue Generation with Pre-trained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3377–3390. Online: Association for Computational Linguistics.

Zhou, K.; Prabhumoye, S.; and Black, A. W. 2018. A Dataset for Document Grounded Conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Zhou, P.; Gopalakrishnan, K.; Hedayatnia, B.; Kim, S.; Pujara, J.; Ren, X.; Liu, Y.; and Hakkani-Tur, D. 2022. Think Before You Speak: Explicitly Generating Implicit Commonsense Knowledge for Response Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1237–1252. Dublin, Ireland: Association for Computational Linguistics.

Zhuang, L.; Wayne, L.; Ya, S.; and Jun, Z. 2021. A Robustly Optimized BERT Pre-training Approach with Post-training. In *Proceedings of the 20th Chinese National Con-*

ference on Computational Linguistics, 1218–1227. Huhhot, China: Chinese Information Processing Society of China.