Semantic data augmentation for meaning maintenance on Task-Oriented Conversation with Large-size Language Model

Jaehwan Lee[*], **Kwanyoung Son**[*], **EunKyung Yoo**[†], **Eugene Kim**[†], **Hoseok Do, Daeyeon Yi**
AI Lab, CTO Division, LG Electronics, South Korea
{jaehwan314.lee, kwanyoung.son, eunkyung.ryu, white.kim, hoseok.do, daeyeon.yi}@lge.com

Abstract

This paper presents our approach to building a generalized model for Track 5 in DSTC11: "Task-oriented Conversational Modeling with Subjective Knowledge" which addresses the challenge of generating responses to users' utterances based on a variety of factual and subjective knowledge. To tackle this challenge, we first augmented the training data by leveraging contextual word embedding and back translation, thereby increasing the quantity of available data. Then, we utilized a large-size language model to enhance the acceptability of the augmented data and fine-tuned the model using augmented data. Specifically, we applied the DeBERTa-v3-large model for knowledge detection and selection, and the BART-large model for response generation. Our best model achieved the seventh rank in the objective evaluation and the second rank in the final official human evaluation. These outcomes serve as solid evidence that data augmentation and using a large-size model were highly effective for developing a conversational model system that incorporates objective and subjective knowledge.

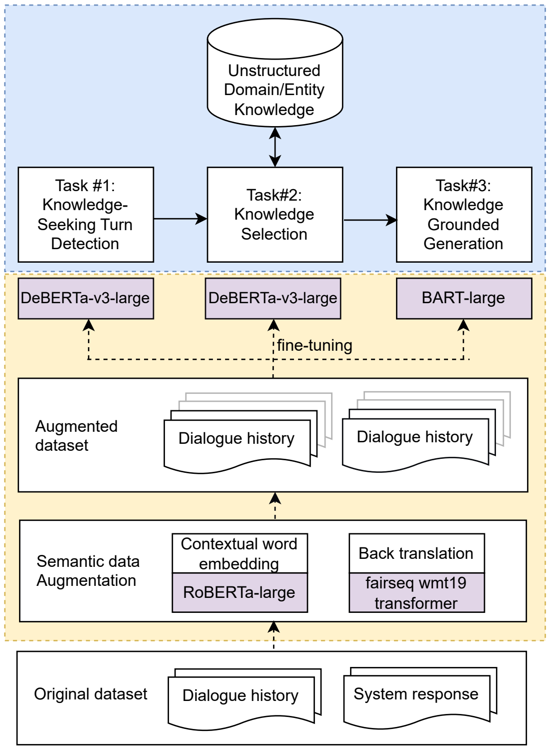

Figure 1: Overall structure of our method: The blue box is a block diagram of DSTC11 Track 5. The yellow box is an illustration of our best model. Dialog history and system response of the original dataset were augmented using contextual word embedding and back translation, and then fine-tuned to large-size models.

1 Introduction

Over the past decade, AI technology has brought about many changes and has been used in various applications and devices that permeate our lives, such as chatbots, personal assistants, and smart kiosks. Traditional studies on task-oriented dialogue systems have focused primarily on providing information and performing actions limited to specific databases or application programming interfaces (API). However, recent studies have focused on tasks beyond available APIs and databases, incorporating relevant domain knowledge. Knowledge-grounded task-oriented conversational modeling tasks are being introduced to address the demands in this research field (Kim et al., 2020; Mi et al., 2021; Kim et al., 2021).

Subjective information, such as customer reviews, rather than objective knowledge from the web, can be meaningful if used well. For these reasons, DSTC11 Track 5 proposes the challenge "Task-Oriented Conversation Modeling through Subjective Knowledge" (Zhao et al., 2023). This challenge comprises three sub-tasks: Knowledge-seeking Turn Detection, Knowledge Selection, and Knowledge-grounded Response Generation. The first sub-task determines whether knowledge access is required for a given utterance and dialog history.

[*]These authors contributed equally to this work.
[†]These authors contributed equally to this work.

Proceedings of The Eleventh Dialog System Technology Challenge, pages 166–176
September 11, 2023

The second sub-task selects the relevant subjective knowledge. Finally, the third sub-task generates an appropriate sentence as a system response based on the selected knowledge.

Our best model adopted two major improvements: data augmentation and large-size language models. As shown in Figure 1, we applied back translation and contextual word embedding for semantic-preserving data augmentation. For the knowledge-seeking turn detection and knowledge selection, we adopted the DeBERTa-v3-large language model. Additionally, for response generation, we utilized the BART-large language model. Fine-tuning our model with augmented data, we achieved outstanding performance. Our model secured the second rank in the official human evaluation.

2 Related work

2.1 Augmentation

Data augmentation is used to increase the size of the training dataset, prevent overfitting, and increase the performance of models in low-resource tasks. Recent studies have demonstrated that data augmentation plays an important role in improving the performance of text classification and text generation tasks (Bayer et al., 2022b; Xu et al., 2021). Some studies augment the data by replacing words or phrases with synonyms (Zhang et al., 2015) and replacing similar word embedding using the k-nearest neighbor algorithm and cosine similarity (Wang and Yang, 2015). By developing this methodology, a method of replacing contextualized word embedding using the language model was also devised. Several attempts have been made, such as replacing words with top-k probability using a bi-directional language model and CNN (Kobayashi, 2018), or using BERT's masked language modeling, which masks several words, predicts them at once, and replaces them from original data (Wu et al., 2019).

In the neural machine translation task, a back translation methodology is devised which translates target language sentences into source language sentences and adds them to training data (Sennrich et al., 2016). A method for generating entirely new sentences was also studied. Anaby-Tavor et al. (2020) generated new sentences by inputting part of the beginning of the original sentence to fine-tune GPT-2.

Alternatively, some studies augmented data by randomly reproducing, inserting, swapping, and deleting words in the original text like Easy Data Augmentation (EDA) without maintaining the meaning of the original text, and demonstrating that it can help prevent overfitting and increase the robustness of the model (Wei and Zou, 2019). However, other studies have shown that EDA lowers the classification scores (Bayer et al., 2022a; Luu et al., 2020). Considering the importance of entities and the meaning of the sentence in selection and generation tasks, augmentation methods that effectively preserve these elements are prioritized over techniques like EDA.

2.2 Transformer-based Language Models

Pre-trained language models (PLMs) were created based on the "Transformer" architecture (Vaswani et al., 2017). They have been applied to Natural Language Processing (NLP) tasks while making improvements in model structure, size, and training strategy. We reviewed and applied the following models in our study.

RoBERTa (Liu et al., 2019) is an improved model of BERT (Devlin et al., 2018) trained using a larger training corpus and longer steps. RoBERTa also dynamically changes the masking pattern applied to the training data. DeBERTa (He et al., 2020) is an improved version of BERT and RoBERTa models that uses a disentangled attention mechanism and an enhanced mask decoder. DeBERTa-v3 (He et al., 2021) improves the original DeBERTa model by replacing mask language modeling with replaced token detection, a more sample-efficient pre-training task. BART (Lewis et al., 2020) is a seq2seq model with a bidirectional encoder and an autoregressive decoder. It is pre-trained by corrupting the text with an arbitrary noising function and learning a model to reconstruct the original text. T5 (Raffel et al., 2020) is also a seq2seq model. It was pre-trained on a multi-task mixture of unsupervised and supervised tasks. Multiple NLP tasks were converted into a unified "text-to-text" format and used in the pre-training.

2.3 Open-ended text generation

The aim of open-ended text generation is to produce coherent and contextually relevant sentences based on a given input or context. Many studies have concentrated on decoding methods (Bengio et al., 2000; Fan et al., 2018; Holtzman et al., 2020; Su et al., 2022) and these utilize open-ended text generation tasks such as story generation (Fan et al., 2018), response generation in dialogue systems(Zhao et al., 2023), and contextual text completion (Radford et al., 2019). Greedy and beam search, a widely used deterministic method, selects a token with the highest probability of appearing next among candidates based on previous conversation history. However, these approaches are ineffective for tasks with subjective content requiring free-form answers (Murray and Chiang, 2018; Yang et al., 2018) and often suffer from dullness, repetition, and degeneration (Holtzman et al., 2020; Welleck et al., 2020; Shao et al., 2017; Li et al., 2016). In the stochastic methods, sampling-based decoding methods including randomness according to the conditional probability distribution have been introduced (Fan et al., 2018). Recently, contrastive search (Su et al., 2022) has been proposed to solve the lack of coherence by model degeneration on the anisotropic distribution of token representations.

3 Method

We try to improve the model in three ways based on the baseline code and dataset of DSTC11 Track 5 (Zhao et al., 2023). First, we employ augmentation techniques, such as contextual word embedding and back translation, to enrich the training data. Second, we test various large-size language models and fine-tune the model using augmented data. Finally, we implement the contrastive search method for response generation to our model and check the feasibility.

3.1 Data Augmentation

For data augmentation, we basically augmented user-system conversations, and desirable system responses. We only augmented user reviews and FAQs when we do named entity recognition-based augmentation. Back translation, contextual word embedding, and ontology-based methods can preserve the entities and meaning of the sentences as much as possible. Consequently, we applied those methods for augmenting the training data of our task. Table 1 summarizes the data augmentation results using each methodology. The first two sentences are examples of back translation results. For contextual word embeddings, we utilized two ways: inserting and substituting words considering the meaning of the sentence while ensuring they retain coherence. The experiments were performed by augmenting the data using each method to increase it by a factor of $\times 1$ and $\times 2$, respectively. Next, word concept-based methods are also used, which replace sets of synonyms for words that are contained in sentences. Finally, a named entity recognition-based method replaces entity names with external knowledge. As presented in Table 1, these models provide suitable alternative sentences through grammatical variations, word order changes, and word substitutions while preserving the overall meaning. Details of each methodology are described below.

Back Translation We used En-De, and En-Ru language pairs to obtain new round-trip translated data because they are easy to implement (Ma, 2019), many studies have been conducted on these language pairs, and there is also a huge dataset for these language pairs(Barrault et al., 2019). Furthermore, as Murthy et al. (2019) proved that translation between languages with similar word orders works better than translation between languages that do not, we used these language pairs. In order to confirm the superior performance improvement effect of back translation to a language with similar word order, a back translation experiment via Korean was also conducted. For back translation implementation, fairseq wmt19 transformer (Ng et al., 2019) En$\rightarrow$De, De$\rightarrow$En, En$\rightarrow$Ru, and Ru$\rightarrow$En models were used. To implement back translation via Korean, the OPUS-MT (Tiedemann, 2020) En$\rightarrow$Ko, Ko$\rightarrow$En model was used.

Contextual Word Embedding Research has proven that the semantic fidelity of augmented data using the BERT model and back translation model is the highest (Kumar et al., 2020), and because the quality of data generated by RoBERTa is better than that of BERT, the contextual word embedding method was applied using RoBERTa to augment the data.

Ontology-based augmentation Ontology-based data augmentation is a common strategy to improve performance in insufficient data environments. We used two strategies for text enrichment: word concept-based and named entity recognition-based augmentation. Word concept-based augmentation comprises two steps: pre-processing and replacing words in target sentences.

Pre-processing When dealing with large-size models, the effect of several data omissions or inaccuracies tends to be less pronounced. However, when

Augmentation Method	Sentence
Original text	Before you do that can you confirm whether Kymmoy offers good views?
Back translation(wmt19-En-De-En)	Can you confirm in advance whether Kymmoy offers good prospects?
Back translation(wmt19-En-Ru-En)	Before you do, can you confirm that Kimmoy offers good views?
Contextual word embedding(insert) ×1	Before you can do all that can you also confirm whether Kymmoy offers any good views?
Contextual word embedding(insert) ×2	Before you do that though can you confirm first whether Mr Kymmoy offers good political views? Before that you do that can can you confirm whether Kymmoy offers you good political views?
Contextual word embedding(substitute) ×1	Before we do that should we confirm whether Kymmoy offers similar results?
Contextual word embedding(substitute) ×2	Before you answer that can you confirm if this offers good value? Before you answer that can we confirm that Kymmoy offers any value?
Word concept-based	Before you do that can you check whether Kymmoy provides good views?
Named Entity Recognition-based	Before you do that can you confirm whether Madera offers good views?

Table 1: Various data augmentation methods and their examples.

Generation method	Response
Label	Acorn guest house's reviews all agree that the hotel is a place of peace and quiet, without any loud or disruptive patrons.
Contrastive Search	The Acorn Guest House gets high marks for their quietness with guests saying it was quiet and relaxing. Would you like to know more about them?
Baseline	Yes, the Acorn Guest House is quiet. Past guests have said that they were not disturbed by any loud or disruptive patrons.
Label	It is not clear if there is a big room available, but one customer said the room they got was spacious while another said the one they got was cramped.
Contrastive Search	The Avalon gets mixed reviews on room size. Some say the rooms are spacious while others say they felt cramped and small. Would you like me to find another place with larger rooms?
Baseline	According to the reviews I have on hand for that location, the opinions are mixed. Half found the rooms cramped and small, while half found them spacious and enjoyable.

Table 2: Example of contrastive search response. The parameters are set as $\alpha = 0.6, k = 4$.

dealing with small datasets, it becomes crucial to increase the precision of the modeling process. Accordingly, we propose a semi-automated approach to data cleaning and correction.

Word concept-based augmentation Next, we propose to augment the data by leveraging the concept of words and their relationships. With WordNet, you can replace a wide range of slot values without any restrictions on part-of-speech. It is an effective method to enhance the quality and consistency of data by ensuring semantic preservation and contextual considerations.

Named Entity Recognition-based augmentation

We choose hotel and restaurant names as the main entities in our oriented task. As entities are captured from the existing training logs, labels, and knowledge data, they are replaced with values from other external entity sets.

3.2 Large Language Model

We used the DeBERTa-v3 model for the detection and selection task, because it showed almost the best performance as a single LLM in NLU benchmarks such as MNLI, SQuAD, and GLUE (He et al., 2021), and was easy to implement. BART and T5 models are used for the generation task because they showed superior performance on sentence generation-related tasks (Lewis et al., 2020; Raffel et al., 2020), and they were easy to implement too. Furthermore, we used each one's large version, because recent studies have shown that the performance of PLMs can be improved by increasing the size of the model parameters (Kaplan et al., 2020). Many experimental results have proved that the large version of the language model is better than the base model too. We also apply the fine-tuning technique using augmented data for better performance.

3.3 Contrastive Search

We try the contrastive search for encoder–decoder models such as BART and T5 for the response generation sub-task. In contrastive search, there are two main parameters as given in Equation 1.

$$x_t = \underset{v \in V^{(k)}}{\arg\max}\{(1-\alpha) \times p_\theta(v|x_{<t}) \\ -\alpha \times (max\{s(h_v, h_{x_j}) : 1 \leq j \leq t-1\})\} \quad (1)$$

Name	CWE Substitute, Insert	BT	Ontology Based	Task#2	Task#3	CS	Total Data Size
M1	.	.	.	DeBERTa-v3-base	BART-base	.	$\times 1$
M2	.	.	.	DeBERTa-v3-base	T5-base	.	$\times 1$
M3	.	.	.	DeBERTa-v3-base	T5-base	O	$\times 1$
M4	.	.	.	DeBERTa-v3-large	BART-large	.	$\times 1$
M5	.	.	Wordnet	DeBERTa-v3-base	BART-base	.	$\times 2$
M6	.	En-De	.	DeBERTa-v3-base	BART-base	.	$\times 2$
M7	.	En-Ru	.	DeBERTa-v3-base	BART-base	.	$\times 2$
M8	.	En-Ko	.	DeBERTa-v3-base	BART-base	.	$\times 2$
M9	RoBERTa-large($\times 1$)	.	.	DeBERTa-v3-base	BART-base	.	$\times 3$
M10	RoBERTa-base($\times 2$)	.	.	DeBERTa-v3-base	BART-base	.	$\times 5$
M11	RoBERTa-large($\times 2$)	.	.	DeBERTa-v3-base	BART-base	.	$\times 5$
M12	RoBERTa-large($\times 3$)	.	.	DeBERTa-v3-base	BART-base	.	$\times 7$
M13	RoBERTa-large($\times 3$)	.	.	DeBERTa-v3-large	BART-large	.	$\times 7$
M14	RoBERTa-large($\times 3$)	En-De, En-Ru	.	DeBERTa-v3-base	BART-base	.	$\times 9$
M15	**RoBERTa-large($\times 3$)**	**En-De, En-Ru**	.	**DeBERTa-v3-large**	**BART-large**	.	**$\times 9$**
M16	RoBERTa-large($\times 3$)	En-De, En-Ru	.	DeBERTa-v3-large	BART-large	O	$\times 9$
M17	RoBERTa-large($\times 3$)	En-De, En-Ru	Named Entity	DeBERTa-v3-base	BART-base	.	$\times 10$
M18	RoBERTa-large($\times 3$)	En-De, En-Ru	Named Entity	DeBERTa-v3-large	BART-large	.	$\times 10$

Table 3: Experiments conducted by our team. CWE stands for Contextual Word Embedding. BT stands for Back Translation. LM stands for Language Model, and it is composed of DeBERTa-v3, and BART. In case of T5-base in Task#3, we used DeBERTa-v3 base-size for the detection and selection task and T5 base-size for the generation task. CS stands for Contrastive Search. $\times 10$ means the overall training data size is ten times bigger than the original data. M1 is the baseline in DSTC11 Track 5 and our best model is M15 highlighted on the yellow background.

The top-k parameter denotes the number of $k \in N$ prediction candidates with the highest probability from the model's probability distribution. Penalty alpha $\alpha \in [0, 1]$ is a weight factor for model confidence and degeneration penalty. The closer the value is to 1, the higher is the weight given to the denaturation penalty term. We tested various cases for the penalty alpha and top-k parameters and double-checked the objective measures and subjective quality of the generated responses manually. Table 2 lists the output of the contrastive search on response generation on the BART-base model.

4 Experiments and Results

Table 3 summarizes the experiments list conducted while participating in this competition. In case of the WordNet-based data augmentation method, when used with a base-size language model, the performance was lower than the baseline, so it was not used for large-size language model experiments. T5 also had poorer performance than BART when using the base size, so large-size T5 was not used for other experiments. We cannot discover the correlation between response generation scores and others. For this reason, we check the subjective quality of generation output. We designed the experiment setting focusing on em accuracy. Table 5 lists the types of hyperparameters used in contrastive search and the performance change according to each value. Although not all cases were tested, the performance was the best when the value of penalty alpha was 0.4 and the top-k value was 4. Therefore, those values were used in subsequent tests related to the contrastive search.

4.1 Objective Evaluation

Table 4 lists the experiments submitted by our team and the performance of each. The first four results represent the performance results on the validation set, while the next four results represent the performance results on the test set. We have the following observations: 1) Using a large-size language model plays a direct role in improving performance. In all tasks, the overall performance of models with large-size models is better than models with base models. In the case of the detection task, since the performance of the base-size model itself was very high at 99.84(The performance of the published baseline was also 99.95), it was difficult to derive much performance improvement even when using the large-size model. 2) Contrastive search is not appropriate for generating responses for task-oriented conversation. There have been performance degradations in basic as well as large-size models with contrastive search. 3) Ontology-based augmentation was not that effective. Ontology-based models

	Base	Valid set				Test set			
		Base-NE (id 2)	Large (id 0)	Large-NE (id 3)	Large-CS (id 1)	Base-NE (id 2)	Large (id 0)	Large-NE (id 3)	Large-CS (id 1)
$Detect_{f1}$	99.84	99.86	99.86	**99.88**	99.86	99.75	**99.80**	99.79	**99.80**
$Select_{f1}$	84.51	84.35	**86.00**	**86.00**	**86.00**	74.92	**81.90**	81.87	**81.90**
$Select_{emacc}$	44.81	47.84	**49.88**	49.86	**49.88**	44.10	**51.30**	51.27	**51.30**
Gen_{bleu}	10.09	10.04	10.54	9.53	**10.90**	9.46	**10.29**	9.51	4.45
Gen_{meteor}	17.44	17.50	**17.99**	17.56	17.30	16.68	**17.64**	17.37	11.39
Gen_{rouge_1}	35.82	35.91	**36.77**	35.71	36.04	34.08	**35.87**	35.18	28.01
Gen_{rouge_2}	14.36	14.53	**15.06**	14.12	14.63	13.66	**14.79**	13.79	10.46
Gen_{rouge_1}	28.28	28.37	**29.20**	27.80	28.70	26.67	**28.22**	27.05	22.76

Table 4: Our final entries submitted on DSTC11 Track 5, and their performance. Our best model is entry id 0. Context word embedding augmentation using RoBERTa-large was applied, and En-De and En-Ru back translations were applied. 'Base' and 'Large' refer to the base-size language model and the large-size language model respectively. NE stands for Named entity recognition-based data augmentation and CS stands for Contrastive Search.

exhibit a slight performance decrease in generation and selection tasks. Augmented data using this method appears to have acted as noise because, among many sets of augmented data, new entities are added to only one set of augmented data. Ontology-based augmentation using synonym sets tends to work at the token level. Therefore, it is necessary to impose numerous constraints to ensure semantic maintenance as the quality of the data may contain biases. However, in this study, we have overlooked this process. Named entity recognition-based augmentation should have been applied to the entire augmented data set rather than just one set of data.

Table 6 presents the change in performance according to various back translations. Data augmentation using En-De back translation and En-Ru Back translation improved performance respectively, and using both methods together performed better than using them separately, as it showed 85.77 of F1 score and 43.92 of EM accuracy score. Furthermore, En-Ko, a language pair with dissimilar word order, was not very helpful in improving performance.

Table 7 lists various experiments using contextual word embedding data augmentation and their

	Case 1	Case 2	Case 3	Case 4	Case 5
$Penalty\ \alpha$	0.8	0.6	0.6	0.5	0.4
$Top - k$	4	5	4	4	4
Gen_{bleu}	4.45	8.45	8.53	8.91	**9.02**
Gen_{meteor}	14.81	16.57	16.62	17.08	**17.29**
Gen_{rouge_1}	27.56	33.84	34.05	35.05	**35.44**
Gen_{rouge_2}	7.27	12.67	12.71	13.58	**13.84**
Gen_{rouge_1}	19.73	26.49	26.64	27.49	**27.77**

Table 5: Performance variation according to some combinations of hyperparameters used in the contrastive search.

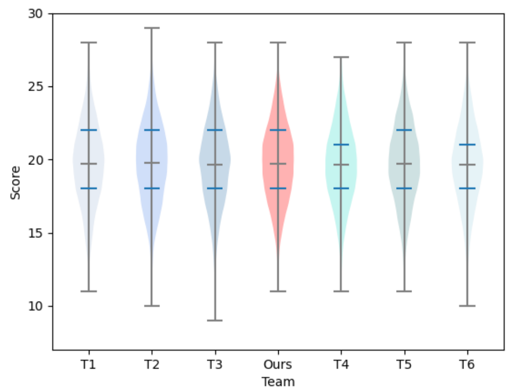

Figure 2: Comparison of human evaluation for the top-6 teams. The vertical axis is the sum of scores for each response.

results. Empirically, the performance tended to improve as more data was augmented using contextual word embedding and when large-size language models were used. The experimental setting which applied CWE×3 augmentation with the large-size model showed the best performance as a 84.43 of F1 score and 49.6 of EM accuracy score.

4.2 Human Evaluation

Conventional reference-based metrics such as BLEU and ROUGE are well-known and widely used as automatic evaluation methods for natural language generation systems. Specifically, the automatic evaluation methods are mainly based on n-grams. BLEU is based on n-gram precision, whereas ROUGE is based on recall. An automatic evaluation method is quick, inexpensive, and language-independent (Papineni et al., 2002). However, they have been shown to have a relatively low correlation with human judgments, especially for

	baseline	En-De BT	En-Ru BT	En-De BT, En-Ru BT	En-Ko BT
$Detect_{f1}$	99.84	99.93	99.91	**99.95**	99.91
$Select_{f1}$	81.08	84.39	83.20	**85.77**	81.07
$Select_{emacc}$	37.27	41.10	41.43	**43.92**	37.02
Gen_{bleu}	9.89	10.22	10.32	**10.50**	10.15
Gen_{meteor}	**17.92**	17.86	17.85	17.72	17.57
Gen_{rouge_1}	35.99	**36.65**	36.46	36.31	35.99
Gen_{rouge_2}	14.38	14.88	**14.89**	**14.89**	14.63
Gen_{rouge_l}	28.19	**28.86**	28.75	28.80	28.49

Table 6: Experiments related to back translation and their performance.

CWE multiple RoBERTa model size LM size (fine-tune)	Baseline	×1 large base	×2 large base	×2 base base	×3 large base	×3 large large	No aug. large
$Detect_{f1}$	99.84	99.91	99.81	99.79	99.86	99.86	**99.93**
$Select_{f1}$	81.08	81.95	82.75	83.74	83.46	**84.43**	83.97
$Select_{emacc}$	37.27	43.29	43.62	42.40	44.08	**49.6**	41.76
Gen_{bleu}	9.89	10.30	10.34	**10.64**	10.30	9.79	10.28
Gen_{meteor}	**17.92**	17.66	17.59	17.78	17.76	17.45	17.87
Gen_{rouge_1}	35.99	36.26	36.18	**36.62**	36.44	35.89	36.30
Gen_{rouge_2}	14.38	14.84	14.83	14.99	14.90	14.29	**15.04**
Gen_{rouge_l}	28.19	28.61	28.50	**28.83**	28.72	27.94	28.71

Table 7: Performance of context word embedding augmentation as multiple increments in various settings of language model size.

Speaker	Text
	... Omit previous dialog ...
User	Also I need to find an expensive chinese restaurant book the table for 3 at 16:45 on tuesday
System	Is there a particular part of town you would like to go to?
User	I need that in the north please.
System	I have 2 chinese restaurants in the north, the hakka and the hotpot, both expensive. Phone number is 22.568988 and 22.3366552 respectively. Would you like me to book reservations for you?
User	I love to try different beers. Do they have lots of different beer options at either Hakka or Hotpot?
Speaker	**Response**
Label	Yes, they do have a lot of beer options at Hakka. One past guest mentioned a good selection of high end beer.
Ours	The Hakka has a good selection of high-end beers to choose from, while the Hotpot has a limited selection of alcoholic drinks.
Response 1	Yes, both Hakka and The Hotpot have a good selection of high-end beers. Would you like me to book a reservation for you?
Response 2	Yes, Hakka does serve alcoholic beverages. The Hotpot does not. Would you like me to book reservations for you?
Response 3	The Hakka restaurant does serve alcoholic beverages, but I am not sure what beer they serve. Would you like me to book reservations for you?
Response 4	Both Hakka and Hotpot have mixed reviews on the selection of beers. Would you like me to book a reservation for you?
Response 5	Hakka has a great selection of high-end beers to choose from. Would you like me to book reservations for you?
Response 6	The Hakka has a great selection of high-end beers to choose from. Would you like me to book a reservation for you?

Table 8: Example of conversations and responses that scored well on human evaluation. Our response contains accurate facts as well as pertinent extra details. Blue : Correct, Violet : Additional info., Orange : Unclear, Red : Wrong info.

tasks that require creativity and diversity (Yang Liu, 2023). Simply put, even if you have two sentences that differ by only one token, their meaning, and usage can be completely different. Therefore, including human evaluation to measure performance is essential. In a qualitative evaluation by humans, factors such as creativity, spontaneity, relevance, fluency, naturalness, and appropriateness will be given higher weightage than automaticity. We placed seventh in the automatic evaluation but second overall through the human evaluation. We believe that we were able to receive a good evaluation because accuracy and appropriateness are the two key factors in human evaluation in DSTC11 Track 5.

Tables 8 and 9 present examples of our responses being more appropriate than others. Since the DSTC11 organizing committee revealed the team id, entry id, entry information and contents of the top 7 teams as a result of human evaluation, the responses of other teams corresponding to

Speaker	Text
User	What can you tell me about the Lovell Lodge?
System	It is a hotel in the north part of town in the moderate price range with a 2 rating, free internet, and free parking
User	Thanks, could you give me the address and phone number?
System	No problem. The Lovell Lodge is located at 365 Milton Road and their phone number is 22.3425478. Is there anything else I can help you with?
User	**Does this place have a strong shower?**
Speaker	**Response**
Label	According to a review I found, the water pressure in the showers at Lovell Lodge is low. Would you like me to look for another option?
Ours	**No**, Lovell Lodge doesn't seem to have a strong shower. Past guests have said that the bathroom was not as clean as they would have liked and the water pressure was low.
Response 1	According to the reviews I have on hand for that location, the opinions are mixed. Half found the bathrooms exceptionally clean, while half thought they weren't as clean as they would have liked.
Response 2	The Lovell Lodge does have a strong shower according to one guest, but I am not sure how strong it is. Is there anything else I can help you with?
Response 3	The Lovell Lodge has exceptionally clean bathrooms according to most guests, but one guest did say that the bathroom was not as clean as they would have liked.
Response 4	The Lovell Lodge has a strong shower according to one guest. Do you want to make a reservation there?
Response 5	According to the reviews I have on hand for that location, the water pressure there is low and the bathroom wasn't as clean as they would have liked. Sorry to disappoint!
Response 6	The Lovell Lodge does have a weak shower. One reviewer said that the water pressure was low and the bathroom wasn't as clean as she would have liked. Is there anything else I can help you with?

Table 9: Another example of conversations and responses that scored well during human evaluation. Our response clearly answers 'yes/no' to the question and also provides extra information to the user.

our response sentence were extracted and written in Tables 8 and 9. In Table 8, our model gave an appropriate answer to the user's last question and provided additional useful information to the user. Other answers provided incorrect or unclear information or provided correct answers only without providing additional helpful information. In Table 9, we are the only ones who responded with a direct answer of "No", including correct information. A question that starts with the auxiliary verb "Do" means that the question must be answered using yes/no, and questions using 'either A or B' need to be answered as a comparison. In the examples in Tables 8 and 9, it can be seen that our model answered these satisfactorily.

Figure 2 depicts the score of the human evaluation from the DSTC11 Track 5 official release. The y-axis represents the sum of scores regarding the accuracy and appropriateness parts of the three raters. We received mostly higher scores than others and rarely had low scores.

5 Conclusion

In DSTC11 Track 5, we applied data augmentation methods that retain meaning and used a large-size language model. As a result, we obtained the seventh place in the objective evaluation and second place in the final human evaluation. Through various and gradual experimental innovations in the designs, it was found that data augmentation using contextual word embedding and back-translation technique play a major role in improving performance. Using the DeBERTa-v3-large and BART-large language models was also crucial. It was also found that ontology-based augmentation and contrastive search were not very helpful for this task.

6 Future Works

Our experiments have left open various questions and possibilities for future work. Owing to the limitations of the competition period, only empirical results of data augmentation and the use of language models were included in this study. Research on finding the combination of various corpus augmentation methods, finding the optimal amount of augmented data, and finding the best combination of language models needs to be conducted. It is also necessary to test other data augmentation methods that retain meaning, such as data augmentation using ChatGPT and other ontology-based methods similar to WordNet. We hope that these works will be addressed in the future.

References

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, N. Tepper, and Naama Zwerdling. 2020. Do not have enough data? deep learning to the rescue! In *AAAI Conference on Artificial Intelligence*.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Markus Bayer, Marc-André Kaufhold, Björn Buchhold, Marcel Keller, Jörg Dallmeyer, and Christian Reuter. 2022a. Data augmentation in natural language processing: a novel text generation approach for long and short text classifiers. *International Journal of Machine Learning and Cybernetics*, 14(1):135–150.

Markus Bayer, Marc-André Kaufhold, and Christian Reuter. 2022b. A survey on data augmentation for text classification. *ACM Computing Surveys*, 55(7):1–39.

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. In *Advances in Neural Information Processing Systems*, volume 13. MIT Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Seokhwan Kim, Mihail Eric, Karthik Gopalakrishnan, Behnam Hedayatnia, Yang Liu, and Dilek Hakkani-Tur. 2020. Beyond domain APIs: Task-oriented conversational modeling with unstructured knowledge access. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 278–289, 1st virtual meeting. Association for Computational Linguistics.

Seokhwan Kim, Yang Liu, Di Jin, Alexandros Papangelis, Karthik Gopalakrishnan, Behnam Hedayatnia, and Dilek Hakkani-Tür. 2021. "how robust ru?": Evaluating task-oriented dialogue systems on spoken conversations. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1147–1154. IEEE.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. 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 452–457, New Orleans, Louisiana. Association for Computational Linguistics.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26, Suzhou, China. Association for Computational Linguistics.

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

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Son Luu, Kiet Nguyen, and Ngan Nguyen. 2020. Empirical study of text augmentation on social media text in Vietnamese. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, pages 462–470, Hanoi, Vietnam. Association for Computational Linguistics.

Edward Ma. 2019. Nlp augmentation. https://github.com/makcedward/nlpaug.

Haitao Mi, Qiyu Ren, Yinpei Dai, Yifan He, Jian Sun, Yongbin Li, Jing Zheng, and Peng Xu. 2021. Towards generalized models for beyond domain api task-oriented dialogue. In *AAAI-21 DSTC9 Workshop*.

Kenton Murray and David Chiang. 2018. Correcting length bias in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 212–223, Brussels, Belgium. Association for Computational Linguistics.

Rudra Murthy, Anoop Kunchukuttan, and Pushpak Bhattacharyya. 2019. Addressing word-order divergence in multilingual neural machine translation for extremely low resource languages. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3868–3873, Minneapolis, Minnesota. Association for Computational Linguistics.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook fair's wmt19 news translation task submission. *arXiv preprint arXiv:1907.06616*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. 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. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Yuanlong Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. 2017. Generating high-quality and informative conversation responses with sequence-to-sequence models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2210–2219, Copenhagen, Denmark. Association for Computational Linguistics.

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022. A contrastive framework for neural text generation. In *Advances in Neural Information Processing Systems*.

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

William Yang Wang and Diyi Yang. 2015. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using #petpeeve tweets. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2557–2563, Lisbon, Portugal. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *International Conference on Learning Representations*.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Conditional bert contextual augmentation. In *Computational Science – ICCS 2019*, pages 84–95, Cham. Springer International Publishing.

Xinnuo Xu, Guoyin Wang, Young-Bum Kim, and Sungjin Lee. 2021. AugNLG: Few-shot natural language generation using self-trained data augmentation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1183–1195, Online. Association for Computational Linguistics.

Yilin Yang, Liang Huang, and Mingbo Ma. 2018. Breaking the beam search curse: A study of (re-)scoring methods and stopping criteria for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3054–3059, Brussels, Belgium. Association for Computational Linguistics.

Yichong Xu Shuohang Wang Ruochen Xu Chenguang Zhu Yang Liu, Dan Iter. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Chao Zhao, Spandana Gella, Seokhwan Kim, Di Jin, Devamanyu Hazarika, Alexandros Papangelis, Behnam Hedayatnia, Mahdi Namazifar, Yang Liu, and Dilek Hakkani-Tur. 2023. "what do others think?": Task-oriented conversational modeling with subjective knowledge.