Concept-based Persona Expansion for Improving Diversity of Persona-Grounded Dialogue

Donghyun Kim[1] **Youbin Ahn**[1] **Chanhee Lee**[1] **Wongyu Kim**[1]
Kyong-Ho Lee[1] **Donghoon Shin**[2] **Yeonsoo Lee**[2]
[1]Department of Computer Science, Yonsei University
[2]NCSOFT
{dhkim92,ybahn,heederer,rladnjsrb9999,khlee89}@yonsei.ac.kr
{dhshin,yeonsoo}@ncsoft.com

Abstract

A persona-grounded dialogue model aims to improve the quality of responses to promote user engagement. However, because the given personas are mostly short and limited to only a few informative words, it is challenging to utilize them to generate diverse responses. To tackle this problem, we propose a novel persona expansion framework, **C**oncept-based **P**ersona e**X**pansion (**CPX**). CPX takes the original persona as input and generates expanded personas that contain conceptually rich content. We constitute CPX with two task modules: 1) Concept Extractor and 2) Sentence Generator. To train these modules, we exploit the duality of two tasks with a commonsense dataset consisting of a concept set and the corresponding sentences which contain the given concepts. Extensive experiments on persona expansion and response generation show that our work sufficiently contributes to improving the quality of responses in diversity and richness.

1 Introduction

A persona-grounded dialogue model aims to generate more human-like and engaging responses based on given traits called persona (Zhang et al., 2018a). As efforts of this research line, many recent works have explored various approaches to improving the quality of persona-based responses (Liu et al., 2020; Song et al., 2020; Kim et al., 2020).

In spite of these efforts, there remain some limitations in persona-grounded dialogue models. They are in need of generating more diverse responses. Due to the predefined personas being mostly short and limited to only a few informative words, the responses based on these personas tend to be generic and monotonous. To tackle this issue, COMPAC (Majumder et al., 2020) expands the predefined personas with a commonsense knowledge graph about events, ATOMIC (Sap et al., 2019). With the expanded personas, the dialogue agent generates more diverse responses which contain inferential knowledge stemming from the original persona.

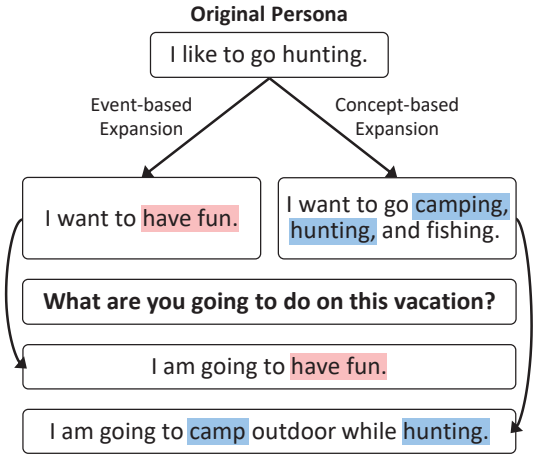

Figure 1: An example of responses generated by different persona expansion strategies for the same query.

However, there is room for improvement in the expansion strategy of COMPAC, which takes an event-based approach as means for persona expansion, which captures causal inferences from personas. The event-based persona expansion is done with the utilization of ATOMIC as a training dataset for the COMET commonsense transformer (Bosselut et al., 2019). As an example, in Figure 1, the expanded persona of the agent (i.e., *I want to have fun*) can be obtained from the given original persona (i.e., *I like to go hunting*) through causal inference. Then given the question "*What are you going to do on this vacation?*", the response based on the event-based expansion strategy is given as "*I am going to have fun*". The generated response is suitable for a given query in terms of relevance. Nevertheless, the response is still monotonic because the expanded persona lacks content.

On the other hand, in the manner of the concept-based expansion strategy, the expanded persona of the agent (i.e., *I want to go camping, hunting, and*

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 3471–3481
May 2-6, 2023 ©2023 Association for Computational Linguistics

fishing) can be obtained through semantic reasoning between concepts (*hunt* → *camping*, *fishing*). Then given the same question, the response based on the expanded persona by concept-based strategy is "*I am going to camp outdoor while hunting*", which is more conceptually diverse and rich.

In general, the diversity of responses is directly correlated with the interlocutor's feeling of being involved in a conversation. Further, in order for the interlocutor to be attracted, a more significant amount of information needs to be delivered. In other words, even in responses that convey the same meaning, the presence of words to enrich the content makes a difference in diversity. Hence, it is crucial for original personas to be expanded to have rich content to improve the diversity of persona-based responses. As demonstrated in Figure 1, by utilizing various concepts in expansion, it is possible to generate personas with richer content than the existing expansion approach.

In this paper, we propose a novel **C**oncept-based **P**ersona e**X**pansion framework, called **CPX**. First, the concept extractor extracts relevant concepts from the original persona sentence, which consists of the constituent words and their semantically related terms. Then, the sentence generator leverages the extracted concepts to generate expanded personas via generative commonsense reasoning. In training, we exploit the duality of CommonGen (Lin et al., 2020) dataset to train the two opposing task modules that consist of the CPX framework: 1) Concept Extractor and 2) Sentence Generator. The experimental results demonstrate that our framework outperforms the baseline models in terms of the expanded persona's diversity. We also show that the persona-grounded dialogue model employing our expansion strategy generate more engaging and diverse responses.

The contributions of our work are as follows:

- We propose a novel concept-based persona expansion framework, CPX, to generate expanded personas with rich content.
- We adopt the duality of concept extraction and sentence generation to constitute the framework for the proposed expansion strategy.
- Through the proposed framework, it is possible to augment the persona-grounded dataset to improve the diversity of responses.

2 Related Work

2.1 Persona-Grounded Dialogue

A persona-grounded dialogue model aims to generate more engaging, human-oriented responses by using some personal characteristics of an agent (Zhang et al., 2018a). As a data-driven approach, Welleck et al. (2019) propose the elaborately constructed dialogue inference dataset, DNLI. Some studies show that fine-tuning the pre-trained language models on the persona-grounded dataset can improve the quality of dialogues (Wolf et al., 2019; Golovanov et al., 2019). Furthermore, recent studies have explored sophisticated neural architectures for better persona-based responses, such as endowing mutual-persona (Liu et al., 2020), multi-stage framework (Song et al., 2020), and self-consciousness modeling (Kim et al., 2020).

Despite the aforementioned efforts, diversity remains limited due to the deficient information in predefined personas. Majumder et al. (2020) adopt the event-based persona expansion strategy for generating more engaging responses. They construct a large number of expanded personas by COMET (Bosselut et al., 2019). Their expansion strategy increases the number of personas, but the expanded personas are still simple in content. That is, while the range of responses that can be generated based on the persona has stretched, the diversity and richness are still limited. To solve this problem, we propose a novel concept-based persona expansion framework that contributes to increasing the semantic richness of expanded personas.

2.2 Generative Commonsense Reasoning

CommonGen task (Lin et al., 2020) aims to generate sentences describing an everyday scenario from a given set of concepts. Formally, the input is a concept set defined as $x = \{c_1, c_2, ..., c_k\} \in X$, and the expected output is a sentence $y \in Y$ that describes a common scenario in our daily life, containing all input concepts. The goal of this task is to learn the reasoning ability between concepts and sentences by injecting relational commonsense knowledge into a language model. We exploit this sentence generation ability in our persona expansion framework (i.e., $f : X \rightarrow Y$). Meanwhile, the extractor model can learn the ability to extract concepts from sentences using an inversely aligned dataset. We leverage the concept extraction ability based on the duality of the tasks in our framework (i.e., $g : Y \rightarrow X$).

Original Concept Set	Human References	Added Concept Set
{cucumber, salad, tomato}	"make a salad of cucumber and tomato" "cooked asparagus and a green salad with cucumbers, tomatoes and chick peas"	{asparagus, green}
{athlete, championship, win}	"olympic athlete has won only medal of the championships" "athlete on his way to winning sport at the championships" "athlete is all smiles after winning the championship after stadium"	{olympic, sport, stadium}
{coast, sea, wave, weather}	"incoming waves of the sea on the coast in foggy weather" "storm clouds in bad weather over rough sea with breaking waves off the coast"	{foggy, storm}

Table 1: Examples of instance pairs (Original Concept Set–Human References) in CommonGen dataset and augmented concept set (Added Concept Set).

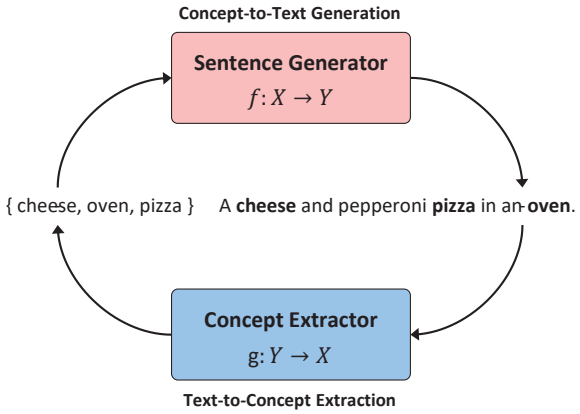

Figure 2: Duality of task modules: 1) Concept Extractor and 2) Sentence Generator. We leverage the CommonGen dataset to train each task module.

	Train	Dev	Test
# Concept Sets	32,651	993	1,497
- Size = 3	25,020	493	-
- Size = 4	4,240	250	747
- Size = 5	3,391	250	750
# Sentences	67,389	4,018	7,644
- Unique@3	49,459	1,814	-
- Unique@4	8,109	1,135	-
- Unique@5	1,488	1,062	-

Table 2: Statistics of CommonGen dataset. The upper row summarizes the numbers of concept sets by size, and the lower is the number of unique sentences corresponding to concept sets with N elements.

3 Task Definition

Our goal is to generate expanded personas to solve the lack of information of a given original persona in PERSONA-CHAT denoted as D_{PC}. Formally, given the original persona p_i, a set of k concepts $C_i = \{c_1, c_2, ..., c_k\}$ is extracted from p_i, where P is a set of all personas in D_{PC}. Then, an expanded persona p_i^+ is generated, using C_i. We perform this expansion for all personas in P, constructing augmented PERSONA-CHAT D_{PC}^+ that includes all expanded personas P^+. Finally, D_{PC}^+ is the augmented dataset that can enhance the dialogue in a data-driven manner.

4 Data Preparation

4.1 Training Data for Framework

As shown in Figure 2, we leverage the duality of two tasks in the CPX framework. The generator and extractor are trained with the original and inversed CommonGen dataset, respectively. The original instance pairs in CommonGen consist of one or more sentences corresponding to each concept set. In other words, it is designed so that a model learns the ability to generate multiple sentences, including given concepts. This dataset configuration can help increase the diversity of the sentences produced by the generator. We also construct the inversely aligned dataset to inject the ability to extract concepts from the sentence into the extractor.

4.2 Concept Augmentation

In order for our expansion strategy to work successfully, the diversity of the extracted concepts and the generated sentences through each task module must be guaranteed. For this reason, we augment a concept set before the training instead of using the dataset as it is. The model of Feng et al. (2021) is referred to in the concept set augmentation, but the details are different. We use an off-the-shelf NLP tool spaCy[1] to extract various words as candidate concepts from the source sentences. Then, we calculate the average BERTScore (Zhang et al., 2019) between the candidate concepts and the source sentences. Finally, 1 to 5 candidate concepts with the highest score are selected as *Added Concept Set*. Some examples of augmentation are illustrated in Table 1. Formally, we denote the original concept set as C^{ori}, and the added concept set as C^{add}. C^+

[1] https://spacy.io/, The used version is 3.1.4.

denotes the union of C^{ori} and C^{add}. For example, if C^{ori} and C^{add} aligned with the same sentence are {*building, cloud, sky*} and {*blue, city*} respectively, then C^{+} is {*building, cloud, sky, blue, city*}.

5 CPX Framework

The CPX framework consists of two task modules: 1) **Concept Extractor** – extracting concepts from the original persona; 2) **Sentence Generator** – generate multiple sentences from the given concepts. Inspired by Xia et al. (2017), we utilize the duality of the tasks to train these modules. However, since we just exploit the advantage of the task duality, two modules are independently trained without sharing parameters based on probabilistic duality.

5.1 Concept Extractor

The concept extractor aims to take the concepts that comprise the given sentence. Our model achieves the capability by training on the inverted version of the CommonGen dataset. Formally, given the input sentence $y \in Y$, and the target concept set to extract is a $x = \{c_1, c_2, ..., c_k\} \in X$. The concept extractor aims to find an objective function $g : Y \rightarrow X$,

$$g(Y; \theta_{ext}) \triangleq argmax_{x \in X} P(x|Y; \theta_{ext}) \quad (1)$$

where θ_{ext} is a trainable parameter. We leverage BERT (Devlin et al., 2019) based model as the extractor. We train the concept extractor in three ways with C^{ori}, C^{add}, and C^{+} as output X to compare the performance according to the extracted concept set. First, the purpose of the extractor Ext^{ori} trained with C^{ori} is to extract only the concepts contained in a given sentence. Second, the purpose of the extractor Ext^{add} trained with C^{add} is to extract the concepts not contained in a given sentence. By utilizing these hidden concepts, it is possible to generate expanded personas with concepts not included in the original persona but semantically correlated. Finally, the purpose of the extractor Ext^{+} trained with C^{+} is to extract not only the concepts contained in a given sentence but also the ones not contained but semantically related. The experimental results of the effect of each extractor on the persona expansion performance are described in Section 6.2.2.

5.2 Sentence Generator

The goal of the sentence generator is precisely the same as the original CommonGen task, generating sentences by inferring the underlying relational knowledge among concepts. Formally, given the input concept set defined as $x = \{c_1, c_2, ..., c_k\} \in X$, the target is to generate a sentence $y \in Y$ containing all input concepts. The sentence generator aims to find an optimal objective function $f : X \rightarrow Y$,

$$f(X; \theta_{gen}) \triangleq argmax_{y \in Y} P(y|X; \theta_{gen}) \quad (2)$$

where θ_{gen} is a trainable parameter. In our setting, we exploit the pre-trained models as the generator, specifically BART (Lewis et al., 2020) and T5 (Raffel et al., 2020), which are the state-of-the-art sequence-to-sequence transformer (Vaswani et al., 2017) language models. As reported by Feng et al. (2021), the use of augmented concept sets at the training phase improves the length and richness of generated sentences. Therefore, we use the integrated concept set C^{+} as input X to train the sentence generator.

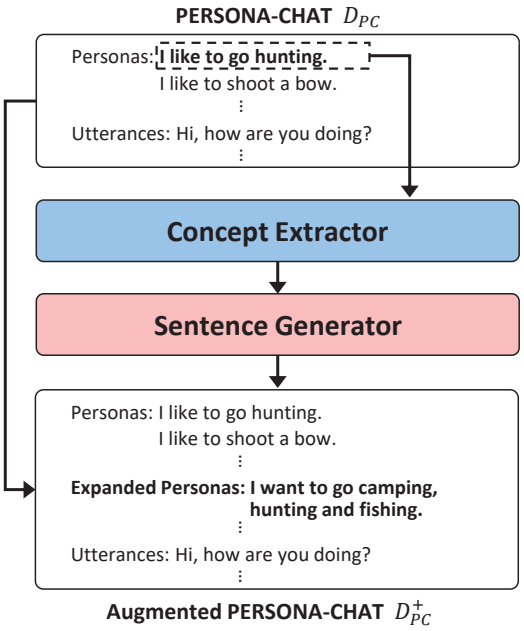

Figure 3: The overall workflow of CPX framework. Given a persona-grounded dialogue dataset D_{PC}, CPX consisting of the extractor and generator generates expanded personas. Final output D^{+}_{PC} is the augmented dialogue dataset.

5.3 Control of Persona Expansion

To constrain the form of expanded personas, we use four types of prompts: *I want*, *I need*, *I feel*, and *I am*.[2] These prompts can control the generation form of the expanded persona and represent an

[2]These four types of prompts were determined by referring to Majumder et al. (2020) and the verb distribution of personas in PERSONA-CHAT.

agent's desire, intent, emotion, and status, respectively. Also, according to our in-depth analysis of PERSONA-CHAT, most of the personas describe the first-person subject, so we fix the subject of the prompt as "*I*". Accordingly, the constrained decoder output sequence for generating the expanded persona is as follows:

$$[I]\,[want|need|feel|am] \tag{3}$$

Consequently, we obtain four expanded personas per original persona. The expanded persona set P^+ is added to the original dataset D_{PC} to construct the augmented dataset D^+_{PC}. The experimental results on response generation using D^+_{PC} will be described in Section 6.3.3.

6 Experiments

To validate the performance of CPX, we conducted two experiments: 1) persona expansion and 2) response generation.

6.1 Experimental Setup

6.1.1 Dataset

We carried out our experiments on the PERSONA-CHAT, converted to the ConvAI2 benchmark version (Dinan et al., 2020). The dataset consists of 17,878/1,000 multi-turn dialogues and 1,155/100 profiles for Train/Valid set. We utilized the unique personas extracted from all profiles for persona expansion experiments. Also, we utilized the original and augmented PERSONA-CHAT datasets to demonstrate the impact of the expansion strategy on response generation.

6.1.2 Baselines

Two types of expansion baselines are considered: 1) **paraphrasing** and 2) **transformer** trained with commonsense knowledge.

• **MANUAL PARAPHRASING** (Zhang et al., 2018a): We used manually paraphrased personas provided with the original PERSONA-CHAT, where workers rephrased the original personas to remove trivial word overlaps.

• **AUTOMATIC PARAPHRASING** (Xie et al., 2020): To paraphrase the personas in an automated manner, we leveraged the existing paraphrasing system based on back-translation. We generated the paraphrased persona by exploiting the pre-trained En-Fr and Fr-En translation models.[3]

[3] https://github.com/google-research/uda

• **COMET** (Bosselut et al., 2019): COMET is a transformer-based model that generates commonsense expansions of a given world event by training on a commonsense knowledge graph such as ATOMIC. We generated expanded personas for four relations (i.e., xWant, xNeed, xReact, xAttr) that are suitable for representing an agent's traits.

In our setting, the automatic paraphrasing model generated one expanded persona for each original persona, and the transformer-based model generated four expanded personas, one for each relation.

6.2 Persona Expansion

In this section, we compared the quality of personas expanded by CPX and baselines. We further reported CPX's performance according to the concept extractor and sentence generator.

6.2.1 Evaluation Metrics

Automatic Evaluation To evaluate the quality of expanded personas, we employed four metrics. (1) Distinct-n (**Dist-n**) (Li et al., 2016) measures the diversity of sentences by calculating the ratio of distinct words against total n-grams. (2) Entropy-n (**Ent-n**) (Zhang et al., 2018b) measures the entropy obtained via the n-gram distribution in a sentence. (3) Also, we report the length of the sentence (**Length**) to measure the amount of content conveyed by the persona. (4) Finally, we employ the BERTScore ($\mathbf{F_{BERT}}$) (Zhang et al., 2019) to measure how semantically relevant the expanded persona is to the original persona.

Human Evaluation We conducted a human evaluation with 100 random samples. We hired three human annotators proficient in domain language through a third-party company. Annotators knew nothing of the system they were evaluating. Also, annotators were properly compensated for their labor. Human annotators evaluated the quality of expanded persona sentences on three criteria. The metrics used for human evaluation are as follows: (1) **Engagement** measures whether the expanded persona is engaging or interesting. (2) **Diversity** measures whether the expanded persona is diverse and informative. (3) **Relevance** measures whether the expanded persona is semantically relevant to the original persona. The scoring range from 1 to 5, with lower indicating poor and higher indicating better. Further, we conducted a pairwise comparison between personas expanded by CPX and personas expanded by COMET.

Expansion Model	Automatic Evaluation						Human Evaluation		
	Dist-1	Dist-2	Ent-1	Ent-2	Length	F_{BERT}	Engagement	Diversity	Relevance
Original	0.08	0.34	5.54	8.63	7.57	-	-	-	-
Manual Paraphrasing	0.16	0.48	6.02	8.47	6.49	-	-	-	-
Automatic Paraphrasing	0.07	0.39	5.79	8.54	7.64	**0.90**	3.08	2.78	**3.98**
COMET	0.04	0.16	4.43	5.92	6.17	0.81	2.76	2.43	3.04
CPX_{BART}	0.19	**0.53**	6.06	8.94	11.57	0.85	4.08	4.14	3.85
CPX_{T5}	**0.20**	0.51	**6.17**	**9.01**	**12.07**	0.86	**4.12**	**4.22**	3.66

Table 3: Automatic and human evaluation results on persona expansion. The extractor used for CPX is Ext^+ trained with C^+. The best results are bolded, and the second-best are underlined.

CPX vs. COMET	Desire (I want)			Intent (I need)			Emotion (I feel)			Status (I am)			Average	
Metrics	win	loss	*k*	win	loss	*k*	win	loss	*k*	win	loss	*k*	win	loss
Engagement	**87.7**	12.3	0.61	**85.7**	14.3	0.66	**91.2**	5.7	0.68	**82.3**	17.7	0.63	**86.4**	13.0
Diversity	**90.3**	9.7	0.72	**89.0**	5.7	0.68	**92.7**	4.3	0.70	**94.3**	4.2	0.69	**91.6**	6.0
Relevance	**52.1**	47.4	0.52	**62.7**	33.3	0.54	43.5	**56.5**	0.48	49.2	**50.8**	0.45	**51.9**	47.0

Table 4: Pairwise comparison results between personas per each relation (prompt) expanded by CPX_{T5} vs. COMET. All numbers are in percentages, and ties are not indicated in the table. The values of Fleiss' kappa *k* (Fleiss, 1971) for all results are in $0.4 < k < 0.8$, indicating moderate agreement among the annotators.

6.2.2 Persona Expansion Results

Analysis on Automatic Evaluation Automatic evaluation results are reported on the left of Table 3. CPX outperforms the baselines in all automatic evaluation metrics in diversity. The personas expanded by CPX were not only long (Length) but also rich in content, being composed of different and unique words (Dist-1/2, Ent-1/2). Also, CPX showed a diversity similar to or better than manual and automatic paraphrasing. In particular, we demonstrated the effectiveness of CPX's concept-based approach by surpassing immensely on all metrics of the comparative model COMET. For the BERTScore, which indicates relevance to the original persona, CPX performed comparatively lower than automatic paraphrasing, i.e., machine translation. This is because the back-translated personas have almost the same meaning as the original persona. On the other hand, CPX scored higher than the comparative generative-based model COMET. Judging from these results, the concepts extracted by CPX were semantically related to the original persona, and the generator effectively generated the expanded persona with sufficient commonsense.

Analysis on Human Evaluation Human evaluation results along the three criteria are depicted on the right of Table 3. The annotators evaluated that CPX outperformed all baselines in terms of Engagement and Diversity except for Relevance, for which CPX made the second-best result but was still close to the machine translation. For accurate evaluation with the target model COMET, we pairwisely compared the personas that each model expanded for four relations, as shown in Table 4. First, CPX was superior in all metrics in the evaluation of the persona forms of "*I want*" and "*I need*", which represent the agent's desire and intent. More specifically, CPX was overwhelmingly superior in Engagement and Diversity. This is because the personas generated by CPX contain richer content than COMET. This means that CPX learned the commonsense relationship between concepts well and used it effectively to generate expanded personas. On the other hand, in terms of Relevance to the original persona, CPX showed lower performance in the remaining two persona forms of "*I feel*" and "*I am*", which represent the agent's emotion and status. For this result, we analyzed that it is because the too-long sentences generated by CPX are more likely to contain concepts unrelated to the original persona.

Extractor	Generator	Engagement	Diversity	Relevance
Ext^{ori}	BART	3.55	3.72	3.66
Ext^{add}		3.16	3.24	2.98
Ext^{+}		**4.08**	**4.14**	**3.85**
Ext^{ori}	T5	3.68	3.85	**3.70**
Ext^{add}		3.20	3.32	2.86
Ext^{+}		**4.12**	**4.22**	3.66

Table 5: Human evaluation results on impact of concept extractor. The best results are bolded.

Expansion Model	Automatic Evaluation					Human Evaluation		
	PPL	BLEU	Dist-1	Dist-2	F_{BERT}	Fluency	Engagement	Diversity
Original	20.24	1.28	0.04	0.19	0.09	3.12	2.44	2.32
Manual Paraphrasing	19.76	1.43	0.13	0.24	0.12	3.20	3.02	3.17
Automatic Paraphrasing	20.38	1.51	0.11	0.27	0.11	3.13	2.78	2.89
COMET	19.87	2.66	0.18	0.31	0.13	3.24	3.05	3.11
CPX_{BART}	<u>19.68</u>	<u>3.18</u>	<u>0.32</u>	<u>0.79</u>	<u>0.14</u>	**3.30**	<u>3.67</u>	<u>3.48</u>
CPX_{T5}	**19.44**	**3.27**	**0.34**	**0.82**	**0.16**	<u>3.28</u>	**3.85**	**3.70**

Table 6: Automatic and human evaluation results on response generation utilizing expanded personas by each model. The generative model used in response generation experiment is GPT-2 (Wolf et al., 2019). The best results are bolded, and the second-best are underlined.

Impact of Concept Extactor We analyzed the impacts of the extracted concept on the CPX framework. We expanded the persona and conducted a human evaluation using the extractors Ext^{ori}, Ext^{add}, and Ext^{+} learned with the three concept sets of C^{ori}, C^{add}, and C^{+}, respectively. As reported in Table 5, the overall performance was the best when the extractor trained with C^{+} was used. In particular, when the Ext^{add} extractor trained using only C^{add}, it showed a deficient value in Relevance. It shows that not only the diversity of the extracted concepts from the sentence but also the relevance to the original persona is essential to concept-based persona expansion.

6.2.3 Case Study

Table 7 shows an example of personas expanded by CPX and other baseline models. The paraphrased personas were semantically identical to the original persona. COMET generated expanded personas that can be inferred from the original persona. However, the expanded personas were still short and lacking in content. On the other hand, CPX generated expanded personas with rich content, including extracted concepts.

Original Persona: I like to remodel homes.
Extracted Concepts: remodel, home, country, house, repair
PARAPHRASING: • MANUAL: I love to redesign houses. • AUTOMATIC: I like to renovate houses.
COMET: • xWant: I want to buy a new home. • xNeed: I need to buy a house. • xReact: I feel happy. • xAttr: I am a homeowner.
CPX: • Desire: I want to live in a country house with a large yard. • Intent: I need to a hammer and a saw to repair my old house. • Emotion: I feel comfortable in country home. • Status: I am busy repairing solar panels on the roof of home.

Table 7: Examples of expanded personas.

6.3 Response Generation

We conducted an experiment on response generation using the original and augmented dataset expanded by different strategies.

6.3.1 Dialogue Model

The dialogue model used in our response generation experiment is GPT-2 (Wolf et al., 2019) just concatenating all persona sentences along with dialog history. In the case of paraphrasing, the training dataset was constructed by concatenating all personas. On the other hand, in the case of COMET and CPX, the number of expanded personas is large. Therefore, training datasets were constructed using only the expanded personas with the highest score, each per original persona, by utilizing the RoBERTa (Liu et al., 2019) based NLI model pretrained with the DNLI dataset.

6.3.2 Evaluation Metrics

We automatically evaluated the response generation using the metrics used in the persona expansion experiment. We also adopted widely used metrics **PPL** and **BLEU** (Papineni et al., 2002) to measure the quality of responses. **Fluency**, which measures whether the generated responses are fluent, was additionally used for human evaluation.

6.3.3 Response Generation Results

Analysis on Automatic Evaluation Automatic evaluation results of response generation are reported on the left of Table 6. The dialogue model trained on the dataset augmented by CPX outperforms other comparative models in all evaluation metrics. In particular, it significantly outperformed other models in terms of diversity. It shows that our expansion strategy improves the diversity of

Personas: I have two dogs. I like to work on vintage cars. My favorite music is country.
Query: What do you usually do in your spare time?
Original: I just relax with my dogs.
Manual: I love vintage car. How about you?
Automatic: I like country music.
COMET: I want to take care of my dogs.
CPX: I like to play frisbees with my dog in a nearby park.

Table 8: Examples of generated responses.

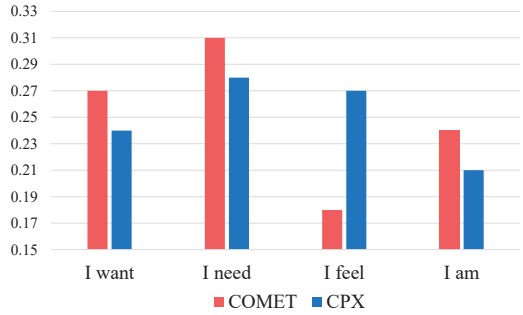

Figure 4: Distribution of each relation (prompt) of the expanded persona selected by the NLI model.

responses based on the language understanding ability of the pre-trained language model. Since pre-trained language models like GPT-2 are trained with large data, it is important to be provided with a rich source (i.e., expanded personas) to generate responses that include various concepts rather than a problem of lack of fine-tuning data.

Analysis on Human Evaluation Human evaluation results are shown on the right of Table 6. A dialogue dataset augmented through CPX significantly improved the performance of the dialogue model for all human evaluation metrics. On the other hand, the dataset augmented by COMET did not achieve significant performance improvement compared to other models. The difference in performance between the two expansion strategies was particularly large in Engagement and Diversity. It means that the concept-based expansion strategy is more suitable for improving the diversity and richness of the persona-based responses.

6.3.4 Case Study

Table 8 depicts examples of responses generated by models trained according to paraphrasing or expansion strategies. The responses generated by CPX were the richest in content. The paraphrasing-based models responded just at the level of simply copying personas. COMET generated a response containing "*take care*" that could be inferred from "*I have two dogs.*" based on its causal reasoning ability. Nevertheless, it did not significantly improve the diversity of the generated responses. On the other hand, CPX utilized the concepts such as "*frisbee*" and "*park*" that are commonly related to the original persona "*I have two dogs.*" to generate a richer utterance. We found that improving the conceptual diversity of personas enriched persona-based responses and made them more engaging.

7 Discussions

Biases in Persona Selection While constructing the dataset for model training, we identified the bias in selecting the expanded personas. As shown in Figure 4, among the four types of personas expanded by COMET, the ratio of each type selected through the NLI model is uneven. While the personas expanded by CPX had been selected more evenly, the distribution was still somewhat biased. This suggests that it is likely for persona expansion models to generate more biased responses through poor selection. It also means that the approach that leverages NLI-based similarities between expanded personas and predefined personas, queries, or gold responses, in the persona selection can also introduce some bias. In order to generate a richer and unbiased persona-based response, it is necessary to study a method of considering the fairness of this expanded persona selection process.

Dual Use of Dataset In this study, we utilized CommonGen dataset for the generative commonsense reasoning task to train the proposed framework. This dual use of the dataset causes some concerns that need to be discussed. First, indiscriminate use of commonsense can lead to the problem of deceiving users or disclosing information by making it difficult for users to recognize that their conversation partner is a chatbot (Gros et al., 2021). Next, the CommonGen dataset may contain biases because it was constructed by human crowd workers. These biases may cause unintended "safety" problems where the dialogue model generates aggressive and harmful responses (Dinan et al., 2022). Fortunately, we found no such case while checking extended personas and generated responses in the CPX framework. Nevertheless, these issues deserve careful consideration in future works leveraging a similar dual use approach.

8 Conclusions and Future Work

In this paper, we proposed the concept-based persona expansion framework to improve the semantic diversity of the persona-grounded dialogue. We expanded personas by extracting and utilizing semantically related concepts through concept-based commonsense reasoning, making persona-grounded dialogues more engaging. During the experiments, we identified some less relevant expansion cases. Therefore, we have a plan to develop a method of preventing potential overexpansion. We also found that in order to effectively reflect the diversely expanded personas in response generation, it was necessary to resolve the bias in the selection process. From this perspective, we will conduct experiments and studies for models that can generate more effective persona-based responses through a bias-free selection process while maintaining the diversity of expanded personas.

Limitations

In this study, we tried to show the importance of the diversity of expanded personas in order to generate more engaging responses. Therefore, the narratives in the paper and the results reported in the tables were focused on diversity. Due to the nature of the proposed concept-based expansion strategy, the expanded personas are generated by using similar but not different concepts, which are not included in the original persona. Also, we intended that the expanded personas be distinct from the original persona as much as possible for the diversity that can be gained from various concepts. For these reasons, we decided to decrease relevance slightly as a trade-off while increasing diversity.

Ethical Considerations

In this study, we utilized the CommonGen and PERSONA-CHAT datasets which contain crowdsourced work by human annotators. Although we did not find any notable cases during this study, the dual use of these datasets may produce results that contain unintended linguistic and cultural biases.

Acknowledgements

We thank all anonymous reviewers for their valuable comments on this work. This work was supported by NCSOFT NLP Center. Kyong-Ho Lee is the corresponding author.

References

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Gavin Abercrombie, A. Bergman, Shannon Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. 2022. SafetyKit: First aid for measuring safety in open-domain conversational systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4113–4133, Dublin, Ireland. Association for Computational Linguistics.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2020. The second conversational intelligence challenge (convai2). In *The NeurIPS'18 Competition: From Machine Learning to Intelligent Conversations*, pages 187–208. Springer.

Steven Y. Feng, Jessica Huynh, Chaitanya Prasad Narisetty, Eduard Hovy, and Varun Gangal. 2021. SAPPHIRE: Approaches for enhanced concept-to-text generation. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 212–225, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Sergey Golovanov, Rauf Kurbanov, Sergey Nikolenko, Kyryl Truskovskyi, Alexander Tselousov, and Thomas Wolf. 2019. Large-scale transfer learning for natural language generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6053–6058, Florence, Italy. Association for Computational Linguistics.

David Gros, Yu Li, and Zhou Yu. 2021. The R-U-a-robot dataset: Helping avoid chatbot deception by detecting user questions about human or non-human identity. 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 6999–7013, Online. Association for Computational Linguistics.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2020. Will I sound like me? improving persona consistency in dialogues through pragmatic self-consciousness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 904–916, Online. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, Online. Association for Computational Linguistics.

Qian Liu, Yihong Chen, Bei Chen, Jian-Guang Lou, Zixuan Chen, Bin Zhou, and Dongmei Zhang. 2020. You impress me: Dialogue generation via mutual persona perception. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1417–1427, Online. 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*.

Bodhisattwa Prasad Majumder, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Julian McAuley. 2020. Like hiking? you probably enjoy nature: Persona-grounded dialog with commonsense expansions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9194–9206, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035.

Haoyu Song, Yan Wang, Wei-Nan Zhang, Xiaojiang Liu, and Ting Liu. 2020. Generate, delete and rewrite: A three-stage framework for improving persona consistency of dialogue generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5821–5831, 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.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741, Florence, Italy. Association for Computational Linguistics.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Yingce Xia, Tao Qin, Wei Chen, Jiang Bian, Nenghai Yu, and Tie-Yan Liu. 2017. Dual supervised learning. In *International conference on machine learning*, pages 3789–3798. PMLR.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33:6256–6268.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018a. 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)*, pages 2204–2213,

Melbourne, Australia. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018b. Generating informative and diverse conversational responses via adversarial information maximization. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 1815–1825.

A Implementation Details

A.1 Persona Expansion

CPX was implemented with HuggingFace's Transformers library.[4] The Concept Extractor initialized from the publicly available BERT-based-uncased[5] model with 12 layers and 768 hidden sizes. We use an Adam optimizer (Kingma and Ba, 2014) and a learning rate of 3e-5. The training of the Concept Extractor was conducted on an Nvidia RTX3090 24G GPU with a batch size of 16. The Sentence Generator initialized from the BART-base[6] and T5-base[7] model. We use an Adam optimizer, and the learning rates are 3e-6 and 5e-6, respectively. The Sentence Generator was trained on an Nvidia RTX3090 24G GPU with a batch size of 8.

A.2 Response Generation

We utilized the repositories and implementation details of GPT-2[8] for response generation. We adjusted some of the details of the model and trained in a single-turn dialogue setting.

B Human Evaluation Protocol

We hired three well-educated annotators from a third-party company to conduct human evaluations. The annotators were given the original persona and the expanded persona pairs to evaluate persona expansion. Also, annotators were given the original personas, the query, and the generated response pairs to evaluate response generation. Each annotator evaluated 100 samples, each sample worth \$0.1. The evaluation was conducted in a double-blind fashion. The human evaluation metrics include: (1) Engagement, which measures whether the expanded persona or generated response is engaging and interesting; (2) Diversity, which measures whether the expanded persona or generated response is diverse (3) Fluency, which measures whether the generated response is fluent; (4) Relevance, which measures whether the expanded persona is semantically relevant to the original persona. The scoring range of the metrics is 1 to 5. The specific scoring criteria for human annotation are shown in Table 9.

Engagement
1-2: (very) Simple and meaningless
3: Semantically moderate
4-5: (very) Interesting and want to keep the conversation
Diversity
1-2: (very) Generic and short in length
3: Conceptually moderate
4-5: (very) Informative and contain various concepts
Fluency
1-2: (very) Hard to read or syntactically incorrect
3: Grammatically correct
4-5: (very) Fluent and easy to understand
Relevance
1-2: (very) Unsuitable for given query or persona
3: Relevant to given query or persona
4-5: (very) Suitable and reflect well given query or persona

Table 9: Scoring criteria of human evaluation.

[4] https://github.com/huggingface/transformers
[5] https://huggingface.co/bert-base-uncased
[6] https://huggingface.co/facebook/bart-base
[7] https://huggingface.co/t5-base
[8] https://github.com/huggingface/transfer-learning-conv-ai