# Improving Persona Consistency of Dialogue Generation by Constructing Negative Word Set

Zhenfeng Han[1], Sai Zhang[1] and Xiaowang Zhang[1,†]

[1]College of Intelligence and Computing, Tianjin University, Tianjin, 300350, China

## Abstract

Maintaining consistent personas is essential for dialogue models. However, dialogue models can generate fluent but inconsistent responses with persona. We observed that some inconsistent responses often contain similar but inconsistent words. In this poster, we propose a method that uses unlikelihood loss to separate semantics of similar but inconsistent words. To get such words, we leverage Word2Vec to construct the negative word set. And ConceptNet is used to remove consistent noise words from negative word set and add antonyms. Experiments demonstrate that our method improves the persona consistency of dialogue generation.

## Keywords

Consistent persona, Unlikelihood, ConceptNet

## 1. Introduction

With the success of existing dialogue models on generating human-like responses, dialogue models are required to express their own personality. Zhang et al. [1] introduce a persona-conditioned dialogue dataset PersonaChat to build persona consistent dialogue models. However, the best performing generative models trained on PersonaChat such as GPT2 [2] still generate fluent but inconsistent responses. The reason is that dialogue models are trained with standard maximum likelihood loss, which lacks the constraint of persona consistency.

Unlikelihood training is a technique developed for removal of repetition in language model completions. Li et al. [3] use unlikelihood to solve the persona consistent issue of dialogue models. However, they just consider the whole sentence and ignore the keywords. We observed that most inconsistent responses are caused by similar but inconsistent words. As shown in fig. 1, the generated response of GPT2 model is inconsistent with persona due to the word “20”. The word “20” is similar to “26” but is inconsistent considering the fourth persona description.

In this poster, we construct negative word set to separate semantics of similar but inconsistent words. Firstly, we obtain coarse negative word set by Word2Vec. Secondly, we use ConceptNet [4] to remove synonyms and add antonyms. Thirdly, we use the unlikelihood loss to compute the loss of negative word set, which assigns a low probability to inconsistent words. The experiments show that our method can generate more consistent responses.

*ISWC'22: The 21th International Semantic Web Conference, October 23-27,2022,Hangzhou*

*Corresponding author.

zhenfenghan@tju.edu.cn (Z. Han); zhenfenghan@tju.edu.cn (S. Zhang); xiaowangzhang@tju.edu.cn (X. Zhang)

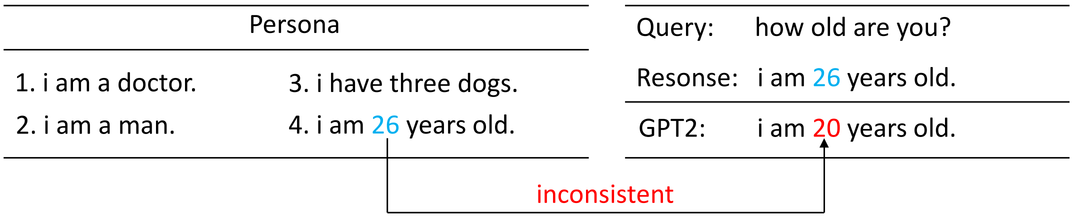

**Figure 1:** The response generated by GPT2 is inconsistent with the persona.

# 2. Approach

## 2.1. Problem Definition

Our task is to train a generative model to generate a persona consistent response. Formally, given a set of persona texts $P = \{P_1, P_2, \dots, P_m\}$, and an query $Q$, to generate a response $R$ which should consistent with persona. Here $P_i$, $Q$ and $R$ are sentences, which are consist of some words, such as $R = \{r_1, r_2, \dots, r_n\}$, where $n$ is the length of $R$.

## 2.2. Training Loss

### 2.2.1. Likelihood Loss

Likelihood training is commonly used in text generation models. Some pre-trained generative models trained with likelihood can generate fluent and meaningful responses. For a sample $\{P, Q, R\}$, the likelihood uses maximum likelihood estimation (MLE) to compute loss:

$$L_{MLE} = -\log\left(p_\theta(R \mid P, Q)\right) = -\sum_{i=1}^{|R|} \log\left(p_\theta\left(r_i \mid P, Q, R_{<i}\right)\right) \tag{1}$$

where $r_i$ is current word needed to be predicted, $R_{<i}$ are previous words before $r_i$ and $p_\theta(r_i \mid P, Q, R_{<i})$ represents the probability of $r_i$ predicted by model conditional on $P, Q, R_{<i}$.

### 2.2.2. Unlikelihood Loss

Likelihood training increases the probability of true word and decreases the probability of all other words. On the contrary, unlikelihood training decreases the probability of negative words. The unlikelihood (UL) loss can be defined as:

$$L_{UL} = -\log\left(1 - p_\theta(R \mid P, Q)\right) = -\sum_{i=1}^{|R|} \sum_{c \in C^i} \log\left(1 - p_\theta\left(c \mid P, Q, R_{<i}\right)\right) \tag{2}$$

where $C^i$ is the negative word set of current word $r_i$.

## 2.3. Constructing Negative Word Set

### 2.3.1. Negative Word Set

MLE leverages the previous context of the word to predict the current word, which results in words with similar contexts having similar semantics. Therefore, the model might generate similar but inconsistent words. One solution is separating the semantics of similar words, which can be done by UL training. The core of UL training is to construct negative word set. Unlike Welleck et al. [5], our negative word set contains inconsistent words with current word conditional on persona. As shown in fig. 1, the generated word "20" is inconsistent with "26" in persona. The negative word set of "26" maybe contains "20", "25" and "30" et al.

### 2.3.2. Word2Vec

The challenge is how to construct negative word set used for UL loss. Word2Vec, a method of learning word embedding, follow the distributional Hypothesis: words that occur in the same contexts tend to have similar meanings. Word2Vec leverages the previous and following context to predict current word, which is similar to MLE. We learning word embedding of all words on dataset by Word2Vec. we approximatively regard similar words computed by cosine similarity of word embedding as negative word set. For example, the negative word set of "man" computed by Word2Vec contains "male", "boy", and "girl" et al.

### 2.3.3. ConceptNet

We also observed that there are some noise words in the negative word set constructed by Word2Vec. The synonym, hyponym, and hypernym have similar context but are consistent with the word, so we should remove them from the coarse negative word set. Fortunately, ConceptNet [4], a knowledge graph containing common sense knowledge, provides these three relations of one word. For example, "male" is a synonym of "man" and "dog" is a hyponym for "pet". Besides, ConceptNet also provides antonyms that can be added to negative word set. For example, "man" and "woman" is a pair of antonym, and they have similar context but opposite semantics.

## 2.4. Training Model

We use GPT2 [2] as our basic model because it shows strong performance in dialogue generation. During the training phase, we combine the UL loss with the MLE loss as follows:

$$L = L_{MLE} + L_{UL} \tag{3}$$

where $L_{MLE}$ aims to promote true words, training model to assign the highest probabilities to such words. On the other hand, $L_{UL}$ focuses on negative words, so that the model can learn to rank negative words lower than true words effectively.

**Table 1**
Results of automatic(on the left) and human evaluations(on the right).

| Method | Consi.$\uparrow$ | Contr.$\downarrow$ | PPL$\downarrow$ | Consi.$\uparrow$ | Contr.$\downarrow$ | Fluc.$\uparrow$ |
|---|---|---|---|---|---|---|
| MLE baseline | 61.8% | 14.2% | 13.9 | 51.0% | 22.7% | 2.65 |
| **+UL** | **63.2%** | **12.3%** | **13.7** | **52.3%** | **21.0%** | **2.71** |
| **+UL+ConceptNet** | **63.9%** | **11.2%** | **13.7** | **52.6%** | **19.7%** | **2.72** |

## 3. Experiments

We verify our method on PersonaChat [1]. For automatic evaluation, we employ a classification model to evaluate the persona-consistency of the generated responses. The results contain three categories: consistent (Consi.), contradictory (Contr.), and neutral. We use perplexity (PPL) to measure the fluency of responses. For human evaluation, we randomly select 100 samples per method and ask three professional annotators to evaluate the quality of these samples. Annotators also label generated responses as consistent (Consi.), contradictory (Contr.), and neutral with persona. The fluency (Flue.) of responses is rated on a 3-scale, with higher scores indicating better fluency.

Table 1 shows that the model trained by UL loss achieves better results on all metrics than the base model. Higher consistent ratio and lower contradictory ratio indicate our method can separate semantics of similar but inconsistent words. The consistency of responses is further improved after using ConceptNet, which means the knowledge such as synonyms provided by ConceptNet is useful to construct higher-quality negative word set.

## 4. Conclusion

In this poster, we propose a method to construct negative word set for unlikelihood training to separate semantics of similar but inconsistent words. Experiments demonstrate that our method can improve persona consistency of dialogue generation. In future work, we are interested in leveraging more efficient loss and constructing more appropriate data to improve the persona consistency of dialogue generation.

## References

[1] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, J. Weston, 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, Melbourne, Australia, Association for Computational Linguistics, 2018, pp. 2204–2213.

[2] S. Radford, J. Wu, R. Child, Language models are unsupervised multitask learners, OpenAI, 2018.

[3] M. Li, S. Roller, I. Kulikov, S. Welleck, Y. Boureau, K. Cho, J. Weston, Don't say that! making inconsistent dialogue unlikely with unlikelihood training, in: Proceedings of the 58th

Annual Meeting of the Association for Computational Linguistics, Online, Association for Computational Linguistics, 2020, pp. 4715–4728.

[4] R. Speer, J. Chin, C. Havasi, Conceptnet 5.5: An open multilingual graph of general knowledge, in: Proceedings of the 31st AAAI Conference on Artificial Intelligence, San Francisco, California, USA, AAAI Press, 2017, pp. 4444–4451.

[5] S. Welleck, I. Kulikov, S. Roller, E. Dinan, K. Cho, J. Weston, Neural text generation with unlikelihood training, in: Proceedings of the 8th International Conference on Learning Representations, Addis Ababa, Ethiopia, OpenReview.net, 2020.