Learning to Abstract for Memory-augmented Conversational Response Generation

Zhiliang Tian,[1,3*] **Wei Bi,**[2] **Xiaopeng Li,**[1,3] **Nevin L. Zhang**[1,3]
[1]Department of Computer Science and Engineering,
The Hong Kong University of Science and Technology, Hong Kong
[2]Tencent AI Lab, Shenzhen, China
[3]HKUST-Xiaoi Joint Lab, Hong Kong
ztianac@cse.ust.hk victoriabi@tencent.com {xlibo,lzhang}@cse.ust.hk

Abstract

Neural generative models for open-domain chit-chat conversations have become an active area of research in recent years. A critical issue with most existing generative models is that the generated responses lack informativeness and diversity. A few researchers attempt to leverage the results of retrieval models to strengthen the generative models, but these models are limited by the quality of the retrieval results. In this work, we propose a memory-augmented generative model, which learns to abstract from the training corpus and saves the useful information to the memory to assist the response generation. Our model clusters query-response samples, extracts characteristics of each cluster, and learns to utilize these characteristics for response generation. Experimental results show that our model outperforms other competitive baselines.

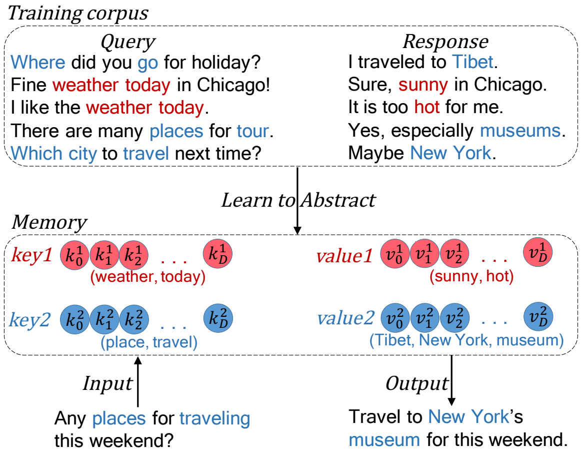
Training corpus
Query
Where did you go for holiday?
Fine weather today in Chicago!
I like the weather today.
There are many places for tour.
Which city to travel next time?
Response
I traveled to Tibet.
Sure, sunny in Chicago.
It is too hot for me.
Yes, especially museums.
Maybe New York.
Learn to Abstract
Memory
key1 k_0^1 k_1^1 k_2^1 ... k_D^1
(weather, today)
value1 v_0^1 v_1^1 v_2^1 ... v_D^1
(sunny, hot)
key2 k_0^2 k_1^2 k_2^2 ... k_D^2
(place, travel)
value2 v_0^2 v_1^2 v_2^2 ... v_D^2
(Tibet, New York, museum)
Input
Any places for traveling this weekend?
Output
Travel to New York's museum for this weekend.

Figure 1: An example of abstracting training corpus and memorizing their characteristics in the form of key vectors and value vectors. Red and blue indicate two clusters. The input query matches the blue one and generates the response assisted by information collected from the last two training samples.

1 Introduction

Automatic human-computer dialogue / conversation is a core topic in natural language processing. There is a boom in research on open-domain chit-chat dialogue systems due to the availability of vast conversational data online. Most existing models of dialogue systems can be divided into retrieval-based models and generative models.

Given a query, retrieval-based (Ji *et al.*, 2014) models search for the most similar query stored in the training corpus and directly copy its corresponding response as the result. These models cannot create new replies customized for the given queries. Generative models (Shang *et al.*, 2015) learn a query-response mapping to generate responses by maximizing $P(r|q)$, where q is the input query and r is the response. The most popular generative model is the Sequence-to-Sequence (Seq2Seq) model (Sutskever *et al.*, 2014), which generates new utterances tailored for queries and achieves high coherence between queries and generated utterances. However, existing generative models often generate uninformative and universal responses (Li *et al.*, 2016a).

To address these issues, several researchers leverage retrieved results R to augment the information used in generative models. Such methods are called retrieval-augmented generative models and their objectives are to maximize $P(r|q, R)$, where R is one or a few (at most 3 in practice) retrieved results. Particularly, some researchers (Li *et al.*, 2017; Zhuang *et al.*, 2017; Song *et al.*, 2018) build the combination of retrieval and generative models, which retrieve one or a few responses r^+, and then feeds both the query q and r^+ into the generative model to maximize $P(r|q, R = r^+)$. It enriches generated responses by informatively retrieved responses but can only utilize a limited

[*] Work done while Zhiliang Tian was collaborating with Tencent AI Lab.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3816–3825
Florence, Italy, July 28 - August 2, 2019. ©2019 Association for Computational Linguistics