Response-Anticipated Memory for On-Demand Knowledge Integration in Response Generation

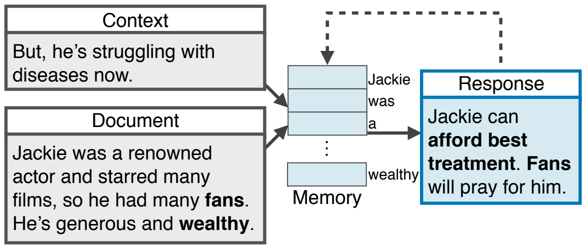
Zhiliang Tian,*[1,4] Wei Bi,†[2] Dongkyu Lee,[1] Lanqing Xue,[1]
Yiping Song,[3] Xiaojiang Liu,[2] Nevin L. Zhang[1,4]
[1]Department of Computer Science and Engineering,
The Hong Kong University of Science and Technology, Hong Kong SAR, China
[2]Tencent AI Lab, Shenzhen, China
[3]Department of Computer Science School of EECS, Peking University, Beijing, China
[4]HKUST Xiao-i Robot Joint Lab, Hong Kong SAR, China
{ztianac,dleear,lxueaa,lzhang}@cse.ust.hk
{victoriabi,kieranliu}@tencent.com songyiping@pku.edu.cn

Abstract

Neural conversation models are known to generate appropriate but non-informative responses in general. A scenario where informativeness can be significantly enhanced is Conversing by Reading (CbR), where conversations take place with respect to a given external document. In previous work, the external document is utilized by (1) creating a context-aware document memory that integrates information from the document and the conversational context, and then (2) generating responses referring to the memory. In this paper, we propose to create the document memory with some anticipated responses in mind. This is achieved using a teacher-student framework. The teacher is given the external document, the context, and the ground-truth response, and learns how to build a response-aware document memory from three sources of information. The student learns to construct a response-anticipated document memory from the first two sources, and the teacher's insight on memory creation. Empirical results show that our model outperforms the previous state-of-the-art for the CbR task.

1 Introduction

Neural conversation models have achieved promising performance in response generation. However, it is widely observed that the generated responses lack sufficient content and information (Li et al., 2016a). One way to address this issue is to integrate various external information into conversation models. Examples of external information include document topics (Xing et al., 2017), commonsense knowledge graphs (Zhou et al., 2018), and domain-specific knowledge bases (Yang et al., 2019). Conversing by reading (CbR) (Qin et al., 2019) is a recently proposed scenario where external information can be ingested to conversations. In CbR, conversations take place with reference to a document. The key problem in CbR is to learn how to integrate information from the external document into response generation on demand.



Context
But, he's struggling with diseases now.
Document
Jackie was a renowned actor and starred many films, so he had many **fans**. He's generous and **wealthy**.
Jackie
was
a
⋮
wealthy
Memory
Response
Jackie can **afford best treatment**. **Fans** will pray for him.

Figure 1: A motivating example of constructing a response-anticipated document memory for response generation. Details are provided in the introduction.

To exploit knowledge from documents for conversations, a conventional way is to extend the sequence-to-sequence (Seq2Seq) model (Sutskever et al., 2014) with Memory Networks (Sukhbaatar et al., 2015), which store knowledge representations accessible to their decoder (Ghazvininejad et al., 2018; Parthasarathi and Pineau, 2018). Dinan et al. (2018) propose to encode the dialogue context as well as a set of retrieved knowledge by Transformer (Vaswani et al., 2017) to construct the memory. However, these methods only use sentence-level representations of the documents in the memory, which cannot pinpoint accurate token-level document information.

To discover token-level document information, researchers borrow models from other generation tasks, which are adept at extracting segments of sentences for given questions. Moghe et al. (2018) explore the pointer generator network (See et al., 2017) for abstractive summarization and the bi-directional attention flow model (Seo et al., 2017), which is a QA model to predict a span of the

*This work was partially done when Zhiliang Tian was an intern at Tencent AI Lab.

†Corresponding author

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 650–659
July 5 - 10, 2020. ©2020 Association for Computational Linguistics