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

Motivation

- Task: conversations take place with respect to a given external document(CbR)
- The key problem in CbR is to learn how to integrate information from the external document into response generation on demand.
- Some previous introduce seq2seq with memory to store knowledge from the document, or learn to predict a span of the document to be contained in the response

Motivation

- However, unlike QA or Machine Reading Comprehension, which requires the model to extract answers from documents, CbR expects to output a general utterance relevant to both context and document.
- the document and the context have no topic overlap, thus we cannot pinpoint document information from the context. Also we can hardly acquire helpful information from context-document interaction

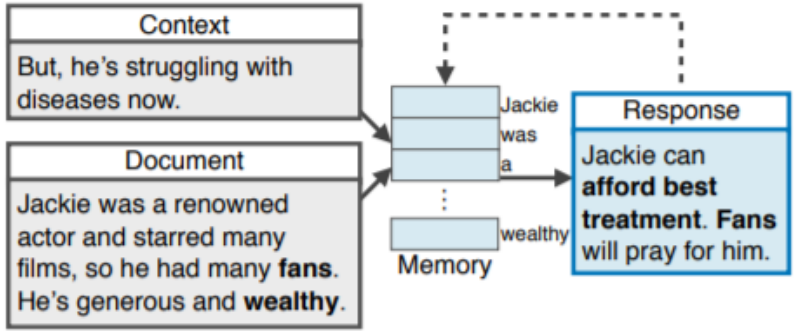

Figure 1: A motivating example of constructing a response-anticipated document memory for response generation. Details are provided in the introduction.

Framework

- Our goal 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 . The student learns to construct a response-anticipated document memory from the first two sources.

Framework

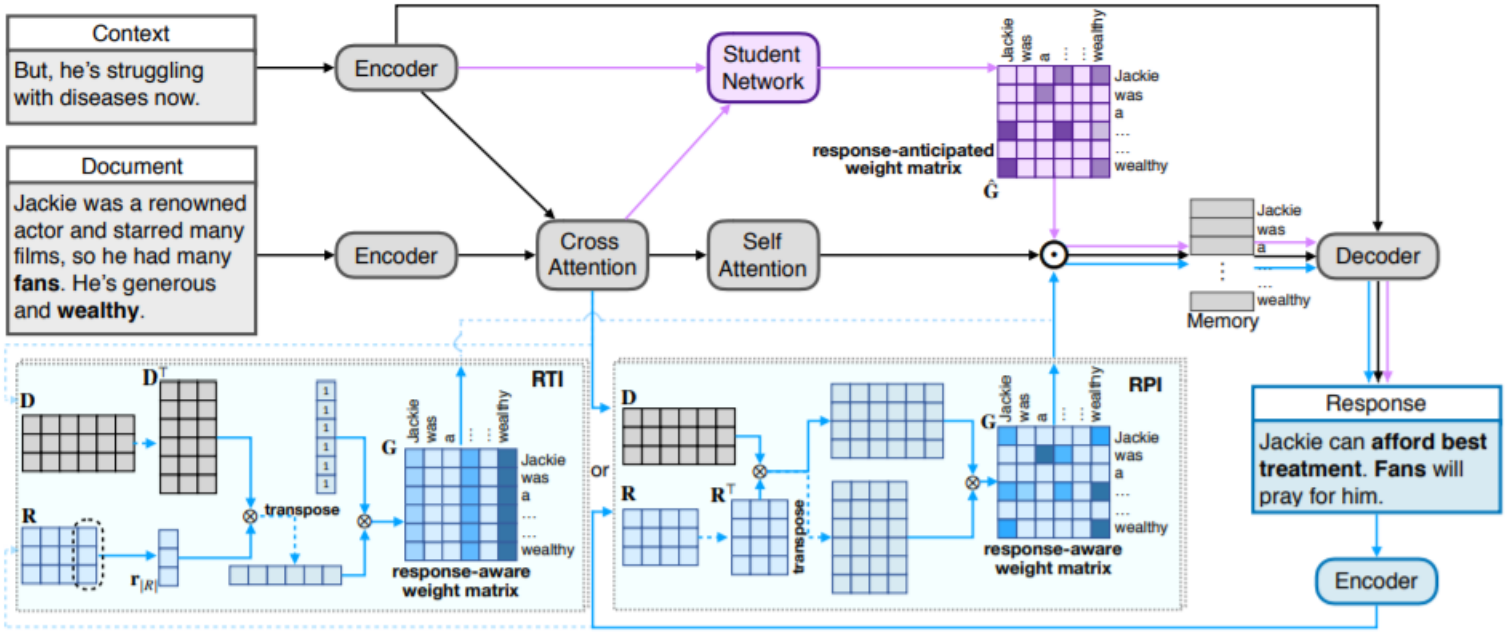

Figure 2: The architecture of our model. Blocks and lines in gray color compose the base model. Blue and gray parts compose the teacher model, while purple parts compose the student model. All components work for training, while only the student model and the decoder works for inference. In the response-aware/anticipated weight matrix, darker grids indicate higher weights. ($\otimes$: matrix multiplication; $\odot$: element-wise matrix multiplication.)

Framework

- The CbR task provides a conversation context X and a document D as inputs, requiring the model to generate a response R to X by referring to D
- We use |X|,|D| and |R| to denote the number tokens in X,D and R, respectively.
- The teacher model learns a response-aware document memory $\boldsymbol{M}$

Used in our base conversation model. Specifically, we construct a response-aware weight matrix $\boldsymbol{G} \in \mathbb{R}^{|D|\times|D|}$ to measure correlation between context-aware document representations and response representations. We impose $\boldsymbol{G}$ on the memory matrix $\boldsymbol{M}$

Base Model

- Input encoder: We use two bi-directional LSTM encoders to extract token-level representations of the document D and the context X.
- Memory construction: we build the document memory $\boldsymbol{M} \in \mathbb{R}^{|D| \times k}$

which will be used in the decoder. A cross-attention layer is first applied to the outputs of the two encoders to integrate information from the context to the document. Then we obtain document representation $\boldsymbol{D} = [\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_{|D|}]$, we then use a self-attention layer:

$$\mathbf{M} = \text{SelfAttn}(\mathbf{D}) = \mathbf{A}\mathbf{D}^T, \mathbf{A} = \text{softmax}(\mathbf{D}^T\mathbf{D}) \quad (1)$$

Base Model

- Output decoder: We use an attentional recurrent decoder to generate response tokens by attending to the memory M.
- $\boldsymbol{e_{t-1}}$ is the word embedding at time step $t-1$

$$\mathbf{z}_t = \mathrm{GRU}(\mathbf{e}_{t-1}, \mathbf{h}_{t-1}), \quad (2)$$

$$\mathbf{h}_t = \mathbf{W_1}[\mathbf{z}_t; \mathrm{CrossAttn}(\mathbf{z}_t, \mathbf{M})] \quad (3)$$

Teacher Model

- To ingest accurate memory information for response generation, our teacher model builds a response aware weight matrix $\boldsymbol{G} \in \mathbb{R}^{|D|\times|D|}$ given D and R.
- we describe how to modify the memory matrix $\boldsymbol{M}$ when $\boldsymbol{G}$ is given. To facilitate response awareness, we update the attention weight matrix $\boldsymbol{A}$ by element-wise multiplying $\boldsymbol{G}$,

$$\mathbf{A} = \text{softmax}(\mathbf{D}^T\mathbf{D}), \widetilde{\mathbf{M}} = (\mathbf{G} \odot \mathbf{A})\mathbf{D}^T. \quad (4)$$

- We describe two methods to construct $\boldsymbol{G}$, (1) We measure the response-aware token importance (RTI),(2) We measure the response-aware pairwise importance (RPI) of each token pair (i, j),

Response-Aware Token Importance (RTI)

- We denote the response-aware token importance of document tokens as $\boldsymbol{\beta} \in \mathbb{R}^{|D|}$, and measure it by response R and context-aware token representation $\boldsymbol{D}$. To obtain β, we first apply an encoder to obtain the token-level representations of the response: $[\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_{|D|}]$, and use its last hidden state $\boldsymbol{r}_{|R|}$ as the sentence representation.
- The response-aware token importance of token i is defined as the similarity between its context-aware token representation $\boldsymbol{d_i}$ and the response representation $\boldsymbol{r}_{|R|}$

$$\beta_i = \mathbf{d}_i^T \mathbf{r}_{|R|}, \quad \mathbf{G} = \mathbf{1}\boldsymbol{\beta}^T, \qquad (5)$$

Response-Aware Token Importance (RTI)

- Recall that the document contains a large amount of noise information in CbR. Thus the attention distributions may become long-tailed due to the existence of many redundant document tokens
- We construct a binary weight vector based on $\boldsymbol{\beta}$. We keep the weight of each element as 1 with the probability of β_i . If the weight of a token turns to 0, this token is deactivated in calculating the attention distributions.

where $g(\boldsymbol{\beta})$ is defined as:

$$\begin{cases} g(\beta_i) = \text{GumbelSoftmax}(\beta_i) & \text{Training,} \\ g(\beta_i) \sim \text{Bernoulli}(\beta_i) & \text{Prediction.} \end{cases} \quad (7)$$

Response-Aware Token Importance (RTI)

- The teacher model maximizes the log-likelihood of responses generated by the response-aware memory constructed with $\boldsymbol{\beta}$

$$\boldsymbol{\beta} = f_{\theta_t}^{t}(D, X, R), \mathcal{J}_t = \mathop{\mathbb{E}}_{D,X,R\sim\mathcal{D}} \log P_{\phi}(R|D, X, \boldsymbol{\beta}), \tag{8}$$

Response-Aware Pairwise Importance (RPI)

- Instead of using token importance, we can construct $\boldsymbol{G}$ by the pairwise importance of each token pair. After obtaining the response representations $[\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_{|D|}]$, we can calculate the similarity of each $\boldsymbol{d}_i$ towards all $\boldsymbol{r}_j$, denoted as $\boldsymbol{n}_i \in \mathbb{R}^{|R|}$:

$$\mathbf{n}_i = [\mathbf{r}_1, \ldots, \mathbf{r}_{|R|}]^T \mathbf{d}_i, \mathbf{B}_{ij} = \mathbf{n}_i^T \mathbf{n}_j, \mathbf{G} = \mathbf{B}. \tag{9}$$

- Compared with response-aware token importance, response-aware pairwise importance allows different values of different index (i, j)'s in $\boldsymbol{G}$

Student Model

- The student model learns to construct a response anticipated weight matrix $\boldsymbol{G}$ without access to the ground-truth R. If we employ RTI, the estimated target is $\boldsymbol{\beta}$, if we use RPI, the estimated target is $\boldsymbol{B}$.
- Given $\boldsymbol{D}$ and $\boldsymbol{X}$ as inputs, we apply a bilinear attention layer to obtain a hidden representation matrix $\boldsymbol{H}$. We then use a two layer MLP with ReLU to estimate $\boldsymbol{\beta}$**.** (we combine two attention outputs by $\boldsymbol{W}_a$ to estimate $\boldsymbol{B}$ in RPI)

$$\mathbf{H} = \text{softmax}(\mathbf{D}^T\mathbf{W}\mathbf{X})\mathbf{X}^T, \tag{10}$$

$$\begin{cases} \hat{\beta} = \text{MLP}(\mathbf{H}) & \text{for RTI,} \\ \hat{\mathbf{B}} = \mathbf{H}\mathbf{W}_\mathbf{a}\mathbf{H}^T & \text{for RPI.} \end{cases} \tag{11}$$

Student Model

- The objective function of the student model is to maximize the log-likelihood of generating response as well as diminish the gap of weighting vector between the student model and the teacher model.

$$\hat{\boldsymbol{\beta}} = f_{\theta_s}^{s}(D, X), \tag{12}$$

$$\mathcal{J}_s = \mathop{\mathbb{E}}_{D,X,R\sim\mathcal{D}} \log P_{\phi}(R|D, X, \hat{\boldsymbol{\beta}}) - \lambda \mathcal{L}_{\text{MSE}}(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}),$$

Model Training

- We first train the teacher model until it converges, and then train the student model with the use of $\boldsymbol{\beta}$ or $\boldsymbol{B}$ from the converged teacher model
- we repeat the above processes iteratively
- In the training of the teacher model, we fix parameters in θ_s; for the student model, we fix θ_t.

Experiments

- We use the dataset for the CbR task released by “Conversing by reading: Contentful neural conversation with on-demand machine reading”. In total, we have 2.3M/13k/1.5k samples for training/testing/validation

Experiments-Baselines

- Seq2seq
- MenNet: A knowledge-grounded conversation model that uses a memory network to store knowledge
- GLKS: It applies a global knowledge selector in encoding and a local selector on every decoding step.
- CMR: . Conversation with Machine Reading, the state-of-art model in CbR. Here we further copy mechanism.

Experiments-Metrices

- Grounding: We measure the relevance between documents and generated responses to reveal the effectiveness of responses exploiting the document.
- Define #overlap as the number of tokens in both the document D and the generated response $\hat{R}$ but not in context X:

$$\#\text{overlap} = |(D \cap \hat{R}) \backslash X \backslash S|, \quad (13)$$

$$\text{P} = \frac{\#\text{overlap}}{|\hat{R} \backslash S|}, \text{R} = \frac{\#\text{overlap}}{|D \backslash S|}, \quad (14)$$

Experiments-Metrices

- We further propose to measure the effectiveness of exploiting the document information considering the ground-truth.Define $\#overlap_{GT}$ as number of tokens in document D, the generated response $\hat{R}$ and the ground truth R but not in contexts X.

$$\#\text{overlap}_{GT} = |(D \cap \hat{R} \cap R) \backslash X \backslash S|, \quad (15)$$

$$\text{P}_{GT} = \frac{\#\text{overlap}_{GT}}{|\hat{R} \backslash S|}, \text{R}_{GT} = \frac{\#\text{overlap}_{GT}}{|D \backslash S|}, \quad (16)$$

Experiments-Automatic results

	Appropriateness			Grounding						Informativeness			
	NIST	BLEU	Meteor	P	R	F1	P_{GT}	R_{GT}	$F1_{GT}$	Ent4	Dist1	Dist2	Len
Human	2.650	3.13%	8.31%	2.89%	0.45%	0.78%	0.44%	0.09%	0.14%	10.445	0.167	0.670	18.8
Seq2Seq	2.223	1.09%	7.34%	1.20%	0.05%	0.10%	0.89%	0.05%	0.09%	9.745	0.023	0.174	15.9
MemNet	2.185	1.10%	7.31%	1.25%	0.06%	0.12%	0.91%	0.05%	0.10%	9.821	0.035	0.226	15.5
GLKS	2.413	1.34%	7.61%	2.47%	0.13%	0.24%	0.84%	0.05%	0.10%	9.715	0.034	0.213	15.3
CMR	2.238	1.38%	7.46%	3.39%	0.20%	0.38%	0.91%	0.05%	0.10%	9.887	0.052	0.283	15.2
CMR+Copy	2.155	1.41%	7.39%	5.37%	0.28%	0.54%	0.92%	0.06%	0.11%	9.798	0.044	0.266	14.4
RAM_T	**2.510**	**1.43%**	**7.74%**	4.46%	0.26%	0.49%	**1.04%**	**0.08%**	**0.15%**	**9.900**	**0.053**	**0.290**	15.1
RAM_P	2.353	1.40%	7.59%	3.89%	0.21%	0.41%	0.97%	0.07%	0.13%	9.891	0.049	0.279	14.9
RAM_T+Copy	2.467	1.41%	7.64%	**6.14%**	**0.32%**	**0.61%**	0.65%	0.04%	0.08%	9.813	0.045	0.265	14.9
RAM_P+Copy	2.342	1.41%	7.51%	5.83%	0.30%	0.57%	0.84%	0.06%	0.10%	9.798	0.045	0.267	14.6

Table 1: Automatic evaluation results on all competing methods. *Len* denotes the length of the generated responses.

Experiments-Case

	Case 1	Case 2
Document	fa premier league was the fourth season of the competition, since its formation in 1992. due to the decision to reduce the number of clubs in the premier league from 22 to 20, only two clubs were promoted instead of the usual three , middlesbrough and bolton wanderers.	darko milicic. darko milicic (serbian cyrillic. serbian pronunciation. born june 20, 1985) is a serbian former professional basketball player . he is 2.13 m (7 ft 0 in) , and played center .
Context	at least we qualified for a european competition we're capable of winning now	that darko milicic, who was drafted 2nd overall in the 2003.nba draft is currently an apple farmer in serbia.
Seq2Seq	i do n't really need to take a time and a bit more and i think he 's saying it was n't in an accident .	he is so happy when i 'm not in 0ame universe as the first time.
MemNet	i am not saying i was a kid .	you know what ? is there anything in a book ?
GLKS	i have a pretty good chance of being the first person i know !	i think a lot of people are still able to get in a hour
CMR	well , at what point do you think about how they are getting play for ?	i remember my comment on my post. and i am not sure why but my point is that he has the best score that will always get a good
CMR+Copy	they are , but not the same as the first one .	he also played the same game, is there title to be a team
RAM_T	i think we have num teams playing the premier league team. in my opinion he was not a good player, but the united kingdom was in the europa	i love him the next time i play for num years, so that is probably the only option i understand.
RAM_T +Copy	they are the best player in the world.	he also played the second one, but that doesn't mean it was num years ago.

Figure 3: Test samples with generated responses of all models. A colored word in the responses indicate that it has similar words with documents or contexts, which are marked in the same color.