Retrieval-guided Dialogue Response Generation via a Matching-to-Generation Framework

Deng Cai[1*] **Yan Wang**[2*] **Wei Bi**[2] **Zhaopeng Tu**[2] **Xiaojiang Liu**[2] **Shuming Shi**[2]
[1]The Chinese University of Hong Kong, [2]Tencent AI Lab
thisisjcykcd@gmail.com
{brandenwang,victoriabi,zptu,kieranliu,shumingshi}@tencent.com

Motivation

- Skeleton-then-response framework has shown promising results for dialogue generation task
- How to precisely extract a skeleton and how to effectively train a retrieval-guided response generator

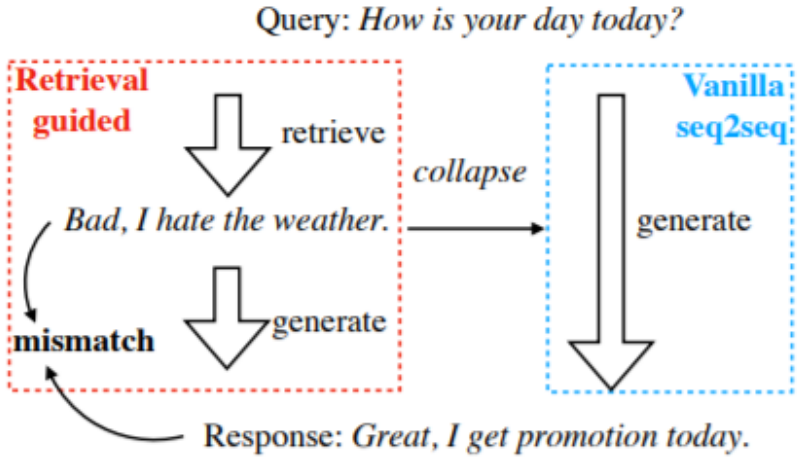


Figure 1: The common problem for training a retrieval-guided generation model in previous work. The model is forced to neglect the retrieved response even though it is a proper response, due to the mismatch between the retrieved response and the target response

frame

- propose an interpretable matching model for matching skeleton extraction.
- train skeleton-guided response generator

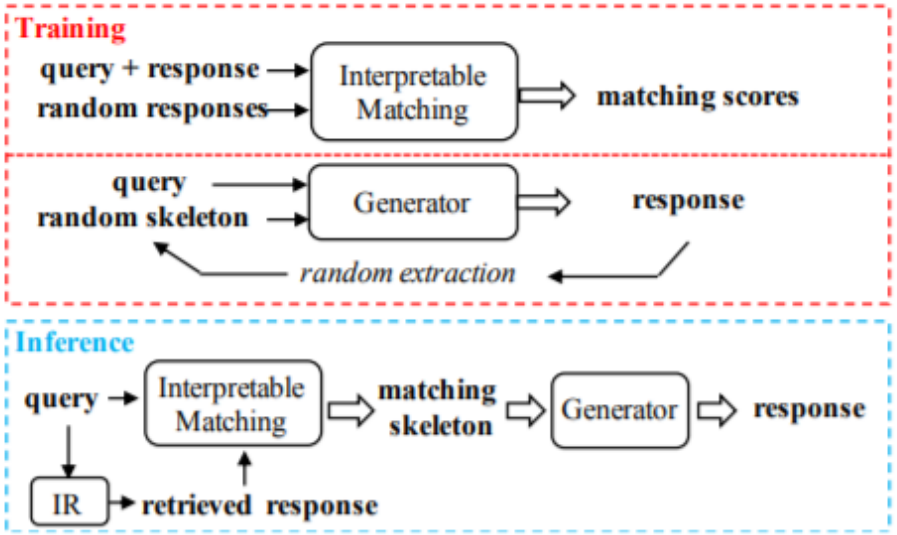
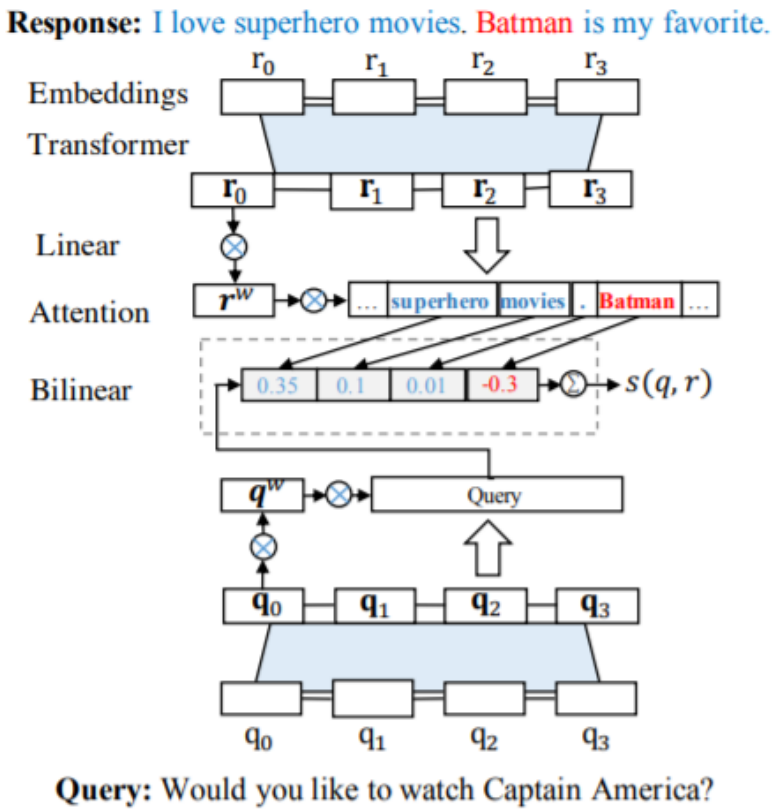


Figure 2: Flow charts during training and inference.

model

- **Interpretable matching model**



The goal of the interpretable matching model is to reveal token-level matching information between a query-response pair, thus we can choose a matching skeleton

model

- **Interpretable matching model**
- For a query $q = (q_1, q_2, q_3, \dots, q_n)$ and a response $r = (r_1, r_2, \dots, r_m)$ where n,m are query length and the response length, respectively
- We first insert a special token at the beginning of each input sentence,
- Use a transformer encoder to get hidden state vectors $\boldsymbol{q}_0, \boldsymbol{q}_1, \dots, \boldsymbol{q}_n$ and $\boldsymbol{r}_0, \boldsymbol{r}_1, \dots, \boldsymbol{r}_m$, where $\boldsymbol{q}_0$ and $\boldsymbol{r}_0$ are considered as the aggregate summary for the query and the response , respectively

model

- **Interpretable matching model**
- First, the sequence-level summary is projected to another vector:

$$\mathbf{r}^w = W^w \mathbf{r}_0 + b^w$$

- Use attention to compute the score between weight vector $\boldsymbol{r^w}$ and the token representation $\boldsymbol{r_i}$:

$$\omega_i = \frac{\exp(\mathbf{r}^w \cdot \mathbf{r}_i)}{\sum_{k=1}^{m} \exp(\mathbf{r}^w \cdot \mathbf{r}_k)}$$

- Calculate the response representation by weighted sum:

$$\mathbf{x}_r = \sum_{k=1}^{m} \omega_i (\mathbf{r}_i + \mathbf{e}_{r_i})$$

model

- **Interpretable matching model**

  The pair-wised score can is calculated by a bilinear function of $\boldsymbol{x}_q$ and

- $\boldsymbol{x}_r$:

$$s(q,r) = \mathbf{x}_q{}^T W^s \mathbf{x}_r$$

- further discussion:

$$\begin{aligned} s(q,r) &= \mathbf{x}_q{}^T W^s \mathbf{x}_r \\ &= \mathbf{x}_q{}^T W^s \sum_{k=1}^{m} \omega_k (\mathbf{r}_k + \mathbf{e}_{r_k}) \\ &= \sum_{k=1}^{m} \omega_k \mathbf{x}_q{}^T W^s (\mathbf{r}_k + \mathbf{e}_{r_k}) \end{aligned}$$

$$s(q,r) = \sum_{k=1}^{m} \omega_k s_k$$

- $s(q,r)$ can be interpreted as the weighted sum of w_k and s_k, where s_k and w_k are the **local matching score** and **local importance score**

model

- **Skeleton-guided Response Generator**
- To ensure the skeleton-guided response generator does make use of the input skeleton, we extract the training skeleton from the ground-truth response by some randomized strategies:
- For a given pair (q, r), we randomly generate a training skeleton through the following procedure:

- All stop words in r are masked in advance. The rest tokens are masked at a mask rate γ. 90% of the time, γ is set to 0.7. 10% of the time, γ is uniformly sampled in the range of $[0, 1]$.

- Instead of always replacing the masked token with a special placeholder token, 20% of time, we replace the token with a random word uniformly sampled from the total vocabulary.

model

- **Skeleton-guided Response Generator**
- The model consists one encoder for query q, one encoder for skeleton s, one decoder for response r.

model

- **Training**
- The matching model and the response generator are trained separately.
- At each training mini-batch, we randomly sample M query-response pairs, then compute the matching score with $s(q, r)$ between all combinations of queries and responses, we can form a scoring matrix:$\boldsymbol{S} \in \mathbb{R}^{M \times M}$ where $\boldsymbol{S}_{i,j}$ is the score between i-th query and the j-th response:

$$L(\theta) = -\sum_{k=1}^{M} \log \text{softmax}(\mathbf{S}_{k:})_k$$

Experiment

- Dataset:a single-turn conversation dataset collected from popular Chinese social websites such as Douban and Weibo
- retrieval system: a publicly available chatbot API

Experiment

- Existing automatic metrics such as BELU and METEOR cannot authentically reflect te quality of dialog response
- The main evaluation is done by human annotators, including informativeness,relevance and fluency. Ezch aspect is rated on a five-point scale

Models	Informativeness	Relevance	Fluency	Dist-1(%)	Dist-2(%)
Retrieval	2.65 (0.90)†	2.58 (0.86)	2.96 (0.72)	**49.10**	**84.19**
Seq2Seq	2.01 (0.65)	2.58 (0.53)	2.71 (0.43)	30.38	54.52
Seq2Seq-MMI	2.47 (0.70)	2.79 (0.67)	2.99 (0.61)	30.98	62.85
RetrieveNRefine^{++}	2.30 (0.79)	2.62 (0.63)	2.82 (0.51)	29.83	61.07
EditVec	2.29 (0.61)	2.62 (0.60)	2.83 (0.47)	35.30	67.57
Skeleton-Lex	2.45 (0.61)	2.80 (0.56)	2.99 (0.46)	25.70	56.61
Ours	**2.69** (0.87)	**3.11** (0.55)	**3.20** (0.55)	**49.01**	**80.36**

Experiment

- Quality of extracted skeleton:

Skeletons	Info.	Relevance	Fluency
Ours	2.69	3.11	3.20
Lexical	2.62	2.92	3.05
keywords	2.56	2.90	3.03
PMI	2.53	2.88	3.02

- PMI: for any word in response, we compute the sum of PMIs between it and all words in query
- Keyword: preserve the most informative words in retrived response

Experiment

- Case

Query	喜欢长头发的女生 I like girls with long hair
Retrieved response	可以 把 头发 后面 汤 大 卷 , 前面 是 直 刘海 , 这样 看 的 很 可爱 You can try perming the hair , hair with curly ends and straight bangs is very cute.
Ours	我也喜欢长**头发**的女生,**卷**了**直刘海**, 头发就不好看了 I also like girl with long **hair**, I think once the **straight bangs** are **curled**, it doesn't look good any more.
Seq2Seq-MMI	我也喜欢长头发的女生 I also like girls with long hair.
RetrieveNRefine^{++}	我喜欢长头发的女生 I like girls with long hair.
EditVec	我也喜欢长头发的女生 I also like girls with long hair.
Skeleton-Lex	我喜欢长头发的女生 I like girls with long hair.
Query	我提现怎么还没到 Why hasn't my withdrawal arrived yet?
Retrieved response	就是 提现 到 你 绑定 的 银行卡 上 了 Withdrawal is in the bound bank card .
Ours	你**提现**的时候**绑定**了**银行卡**吗 Have you **bounded** the **bank card** when **withdrawing** ?
Seq2Seq-MMI	我提现也到了 My withdrawal has arrived too.
RetrieveNRefine^{++}	你要支付宝干嘛 Why do you need Alipay?
EditVec	你是提现的吗 Do you want to withdraw?
Skeleton-Lex	你不是已经到了吗 Haven't you arrived ?
Query	我月经不太规律 I have irregular periods.
Retrieved response	去 看 医生 啊 . 最 好看 中医 , 挺准 的 , 别 不好意思 Go to see a doctor , best see a traditional Chinese doctor . Pretty accurate , don't be shy.
Ours	看**医生**吧,**最**简单的方法就是**中医**调理一下了 See a **doctor**, the **easiest** way is with **Chinese medicine** recuperation.
Seq2Seq-MMI	我也不规律 I am irregular too.
RetrieveNRefine^{++}	有啥不规律的 What is irregular?
EditVec	有啥不好意思的 Why you are shy?
Skeleton-Lex	我也不规律 I am irregular too.

Inspire

- For a given pair (x,y), find some input sentences which is most similar to x in training corpus , then use these sentences to guide generation