posterID:P6-08-SLP

The 2nd FutureDial Challenge: Dialog Systems with Retrieval Augmented Generation

Yucheng Cai, Si Chen, Yuxuan Wu, Yi Huang, Junlan Feng, Zhijian Ou
Speech Processing and Machine Intelligence (SPMI) Lab, Tsinghua University, China Mobile Research Institute, China
fengjunlan@chinamobile.com, ozj@tsinghua.edu.cn

Motivation

The 2nd FutureDial challenge (Futuredial-RAG)

- Conversational Datasets are extremely important for developing dialog systems.
- Retrieval Augmented Generation (RAG) helps to **import knowledge and reduce hallucination.**
- The first **human-to-human real-life customer service dialog datasets** featuring retrieval augmented generation.

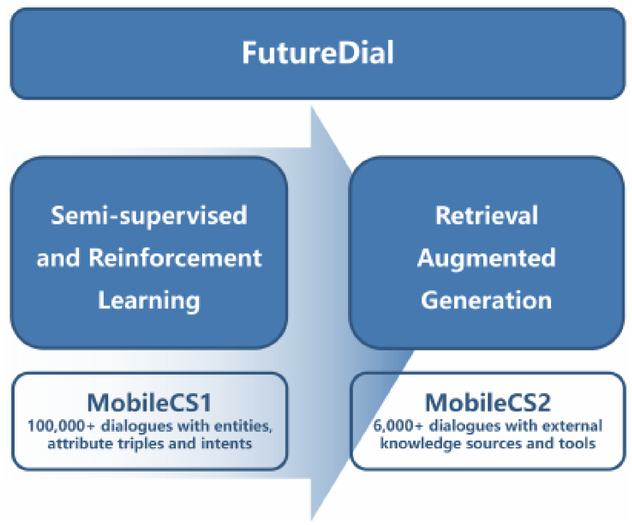

Fig. 1. Overview of the FutureDial-RAG Challenge: Dialog Systems with Retrieval Augmented Generation.

Related Works

- Traditional dialog datasets use **wizard-of-oz** simulation like MultiWOZ.
- The MobileCS2 dataset is annotated from **real-life** customer service dialogs.
- Our dataset contains **ground-truth knowledge** annotation, while other RAG-related datasets (like Wizard of Wikipedia) do not.

Challenge Dataset

MobileCS2 overview:

✓Contains 2751 carefully annotated **real-life** China mobile customer service dialogs
✓Remove the **sensitive personal information** and translated into **English**
✓Restructure the annotation to build **the user, FAQ, and product** database


```
"20287056060747740601898 8": {
  "log": [
    {
      "user": "Manual",
      "system": "Seasons change, but our affection for you never does. Meeting you is the best moment! Hello, what can I assist you with? Just let me know!",
      "api_query": "",
      "api_result": ""
    },
    {
      "user": "I have claimed a monthly card from Bilibili, where can I redeem it?",
      "system": "Hello, please wait a moment, I will check for you. Bilibili Anytime Watch Membership. Dear, the message for claiming has been sent to you. You can claim according to the message prompt.",
      "api_query": "search for user information",
      "api_result": "Business Name: \"Bilibili Casual Viewing Membership, Business Details: $19.9 RMB/month, includes 15GB domestic (excluding Hong Kong, Macao, and Taiwan) Bilibili targeted data. After users subscribe, they can claim a Bilibili benefit each month on the unified online page.\""
    },
    {
      "user": "Correct.",
      "system": "Dear, once you claim this, it will automatically go under this phone number. When you log in to B site, you can use this phone number as your account. Dear, do you have any other questions?",
      "api_query": "",
      "api_result": ""
    }
  ],
  "local_kb": [
    "Business Name: \"Bilibili Casual Viewing Membership, Business Details: $19.9 RMB/month, includes 15GB domestic (excluding Hong Kong, Macao, and Taiwan) Bilibili targeted data. After users subscribe, they can claim a Bilibili benefit each month on the unified online page.\""
  ]
```


Fig. 2. An example of annotated dialogs.

Table 1. Statistics of the MobileCS2 dataset.

	Train	Dev	Test	Total
Dialogs	1926	412	413	2751
Turns	16120	3246	3240	22606
Knowledge-retrieval-needed	4314	808	817	5939

Two challenge tracks based on the challenge dataset:

➢Track1: **Information retrieval** based on knowledge bases and dialog context
➢Track2: Dialog systems with **retrieval augmented generation**

Annotation details:

➢Each turn, the **api_query** and the corresponding **api-result** are annotated.
➢Api_query to KB: Search for user information→user KB, [QA]→FAQ KB, Search for products information→product KB

Main_class	api_query	Description
QA	[QA]	Consult the FAQ manual, which includes a collection of commonly asked questions such as recent promotional packages and general business regulations.
NULL	-	Based on the contextual information, customer service personnel can successfully complete the conversation without the need for additional inquiries.
API-Inquiry	Search for products information	Inquire about the current business information of the mobile company, such as specific packages, data plans, etc.
	Search for user information	Inquire about the services that the user currently possesses, including the current package, current monthly fee, and current data usage.
	Search for other information	Inquire about other key information used to complete the dialog. For example, inquiring about text messages regarding excessive data usage alerts sent by the mobile company in the historical trajectory, querying the address of the business hall, etc.
API-Cancel	Cancel business	Revoke a certain service currently possessed by the user.
API-Handle	Handle business	Process a new service for the user.
API-Verification	Verify identity	Send verification codes, passwords, or other related customer service verification operations to the user.

Baseline

Setting:

- Dialog X with T turns:
  $u_1, r_1, \cdots, u_T, r_T$
- Knowledge base KB contains 3 parts:
  $KB_X \triangleq KB_{user} \cup KB_{FAQ} \cup KB_{product}$
- Context $c_t \triangleq u_1 \oplus r_1 \oplus \cdots \oplus u_{t-1} \oplus r_{t-1} \oplus u_t$

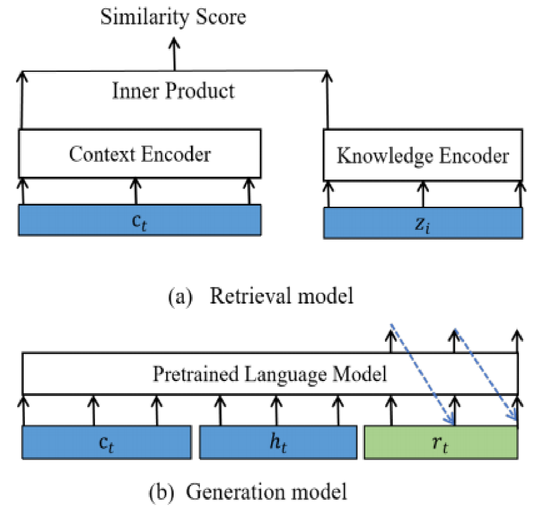

Fig. 3. Overview of our baseline systems: (a) the retrieval model, (b) the generation model.

Approach:

- Get the retrieval probability of each knowledge piece z_i in the KB_X:
  $p_\eta(z_i \mid c_t) \propto \exp\left(\text{Encoder}_p(z_i)^\top \text{Encoder}_c(c_t)\right)$
- Train the retriever with loss (Z_+ is the annotated api-result):

$$\mathcal{L}_{ret} = -\frac{1}{|Z_+|} \sum_{z \in Z_+} \log \frac{p_\eta(z \mid c_t)}{p_\eta(z \mid c_t) + \sum_{i=1, z_i \neq z}^{K} p_\eta(z_i \mid c_t)}$$

- Train the generator to maximize the auto-regressive generation probability:

$$p_\theta(r_t \mid c_t, h_t) = \prod_{l=1}^{|r_t|} p_\theta(y^l \mid c_t, h_t, y^1, \ldots, y^{l-1})$$

where h_t is the retrieved knowledge and y^l the l-th token of r_t

Evaluation Results

Pretrained models: retriever (BGE) ;generator (GPT2-chinese)

Metrics for retriever:

recall@1, recall@5, recall@20, score = recall@1 + recall@5 + recall@20

Table 3. Baseline results for the retrieval task (Track1).

recall@1	recall@5	recall@20	Score
0.225	0.387	0.573	1.185

Metrics for generator:

➢BLEU-4: 4-gram overlap between real responses and generated responses
➢BERTScore: semantic similarity between real responses and generated responses
➢Inform: how often generated responses cover the requested information by the user
➢Score = 0.5 ∗ (BLEU/100 + BERT Score) + Inform

Table 4. Baseline results for the response generation task (Track2).

BLEU-4	BERTScore	Inform	Score
14.54	0.639	0.092	0.484

Analysis of the baseline results:

➢ Retrieval: the retrieval task is **difficult** as the **recall@20 is low**
➢ Generation: the **low inform** result is possibly because the low **recall@1**; building a RAG system is difficult as the **Score** is low

Conclusion

➢ We present the FutureDial-RAG challenge to **promote the study of RAG** for dialog systems
➢ We build the MobileCS2 dataset, **a real-life customer service** datasets with nearly 3000 high-quality annotated dialogs.
➢ We **build a baseline and design evaluation metrics**. The baseline results show that MobileCS2 is challenging.