CONVERSATIONAL QUERY REWRITING WITH SELF-SUPERVISED LEARNING

Hang Liu, Meng Chen, Youzheng Wu, Xiaodong He, Bowen Zhou*

JD AI, Beijing, China

{liuhang55, chenmeng20, wuyouzheng1, xiaodong.he, bowen.zhou}@jd.com

ABSTRACT

Context modeling plays a critical role in building multi-turn dialogue systems. Conversational Query Rewriting (CQR) aims to simplify the multi-turn dialogue modeling into a single-turn problem by explicitly rewriting the conversational query into a self-contained utterance. However, existing approaches rely on massive supervised training data, which is labor-intensive to annotate. And the detection of the omitted important information from context can be further improved. Besides, intent consistency constraint between contextual query and rewritten query is also ignored. To tackle these issues, we first propose to construct a large-scale CQR dataset automatically via self-supervised learning, which does not need human annotation. Then we introduce a novel CQR model **Teresa** based on Transformer, which is enhanced by self-attentive keywords detection and intent consistency constraint. Finally, we conduct extensive experiments on two public datasets. Experimental results demonstrate that our proposed model outperforms existing CQR baselines significantly, and also prove the effectiveness of self-supervised learning on improving the CQR performance.

Index Terms— conversational query rewriting, self-supervised learning, multi-turn dialogue

1. INTRODUCTION

Building conversational bots has attracted increasing attention due to the promising potentials on applications like virtual assistants [1] or customer service systems [2]. With the development of deep learning, both the task-oriented dialogue and open-domain conversation have made remarkable progress in recent years [3, 4, 5]. However, multi-turn dialogue modeling still remains extremely challenging. One major reason is people tend to use co-reference and ellipsis in daily conversations [6], which leaves the utterances paragmatically incomplete if they are separated from context. According to previous research, this phenomenon exists in more than 60% conversations [7]. Taking the conversation in Table 1 for example, the key information of *Bluetooth headphones* is omitted in the Q_2. To help the conversational bots understand the incomplete utterances, we rewrite Q_2 to R_2.

---

*Corresponding author.

Turn	Utterance (*Translation*)
Q_1	请问Mix3可以连接蓝牙耳机吗? *Can Mix3 connect to Bluetooth headphones?*
A_1	可以的 *Yes, Mix3 can.*
Q_2	小米8可以连接吗? *Can Mi8 connect it?*
R_2	小米8可以连接蓝牙耳机吗? *Can Mi8 connect to Bluetooth headphones?*

Table 1. An example of contextual query rewriting. The incomplete query Q_2 is rewritten into R_2 by our proposed model. Mix3 and Mi8 are model names of cellphone.

Previous works [7, 8, 9, 10, 11, 12] formulate the conversational context understanding as a query rewriting problem, transforming a user utterance with anaphora or ellipsis into a new utterance where the left-out or referred expressions are automatically generated from the dialogue context. Usually an end-to-end sequence-to-sequence model with copy mechanism is applied for this task. Su et al. [8] proposed a Transformer-based generative model with pointer network. To locate the omitted information from context, Song et al. [9] employed a multi-task learning framework by taking sequence labeling as an auxiliary task. Pan et al. [7] proposed a Pick-and-Combine model to decompose the task into a cascaded process. The picking stage predicts the omitted words and the combining stage rewrites the query. As the training process of generative model needs massive rewriting pairs, all above works construct their datasets by manual annotation.

Although tremendous progress has been made, we argue that the following aspects can be further improved. First, collecting large-scale supervised data is extremely time-consuming and labor-intensive, which becomes the bottleneck of neural models. Second, the sequence labeling task tends to focus on entities and may ignore other important information, such as verb and adjective words. However, the omitted information is usually text spans which are not limited to entity words. Third, previous works lack intent consistency constraint between contextual query and rewritten query, which leaves the generation under-constrained.

To tackle above issues, in this paper, we propose a novel

Transformer-based qu**e**ry **re**writing model, equipped by **S**elf-Attentive Keywords Detection (SAKD) and Intent Consistency Constr**a**int (ICC), namely **Teresa**. Specifically, SAKD utilizes the self-attention weights of words to build a graph network on encoder to represent relevance between words. Then TextRank [13] algorithm is adopted to calculate each word's importance, which guides the copy mechanism during generation. As to ICC, we first obtain the intent representations of contextual query and rewritten query with the same encoder, then force their distributions on intent to keep consistent by Kullback-Leibler divergence loss [14]. Lastly, we propose to construct the CQR training data automatically from raw dialogue corpus with self-supervised learning (SSL) [15], which does not need manual annotation. Extensive experiments are performed on two public datasets. And experimental results demonstrate the superiority of our proposed model compared with state-of-the-art baselines.

2. METHODOLOGY

We denote a conversation session $s = \{u_1, u_2, ..., u_t\}$ with t utterances. Given $q = u_t$ is the incomplete query and $c = \{u_1, ..., u_{t-1}\}$ is the context, our goal is to learn a rewriting model $g(c, q)$ to generate a context-independent query r, which has the same meaning with q but recovers all co-referenced and omitted information. r could be equivalent to q when q is already self-contained without context c.

2.1. Self-Supervised Learning

Generally, the supervised learning (SL) is trained over a specific task with a large manually labeled dataset. Differently, self-supervised learning (SSL), also known as self-supervision, is an emerging solution to such cases where data labeling is automated, and human interaction is eliminated. In SSL, the learning model trains itself by leveraging one part of the data to predict the other part and generate labels accurately. In the end, this learning method converts an unsupervised learning problem into a supervised one. While Computer Vision is making amazing progress on SSL only in the last few years [16], SSL has been a trend in NLP research recently. Especially for the representation learning in NLP (e.g. Skip-Gram [17] and BERT [18]), various *pre-training tasks* are proposed in the self-supervised formulations, such as Neighbor Word Prediction, Masked Language Modeling, and Next Sentence Prediction etc.

Inspired by SSL, as Figure 1 shows, we propose to construct the training sample (c, q, r) from raw dialogue corpus automatically, by corrupting the normal query r into incomplete query q. Suppose there exist common text spans between context c and normal query r, we can construct the incomplete q by treating the common text spans by following two approaches: (1) removing the common text spans from r directly, (2) if the common text spans are noun phrases in
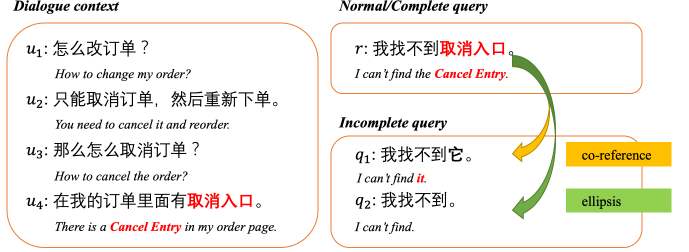

Fig. 1. Dataset construction based on SSL.

r, we replace them with pronouns in 50% of time. The first approach is designed to cover the ellipsis situation, and the second approach is to cover the co-reference scenario. To improve the quality of the constructed dataset, we require the common text spans to contain at least one word of noun, verb or adjective. And only informative queries (queries with at least 10 characters for Chinese in this work) are processed. With the constructed dataset above, the CQR task can be formulated as a typical SSL problem. We first corrupt the normal query r, then force the model to recover it.

We also follow the trendy *pre-train and fine-tuning* two-stage learning paradigm to train our CQR model. The pre-training stage is based on SSL with auto-generated data, and the fine-tuning stage is based on SL with annotated data.

2.2. Model

Transformer-based Generative Model. Figure 2 shows the overall architecture of our proposed model Teresa, which is in the encoder-decoder framework [19]. Both the encoder and decoder are based on the transformer model [20]. To learn dependency between context c and query q, the input context and query are packed together with a segment token [SEP]. For each token w_i, the input embedding is the sum of token, position and segment embedding where segment embedding indicates if each token comes from the context or query. Then transformer encoder is leveraged to produce a sequence of hidden states H.

The transformer decoder is applied to generate rewritten query r, in which the copy mechanism [21] is utilized to copy important words from the context c and query q. Each layer l of decoder is composed of three sub-layers. The first sub-layer is a multi-head self-attention layer M^l. The second sub-layer is an encoder-decoder interaction layer. And the third sub-layer is a feed-forward layer. Inspired by Su et al [8], we calculate the context representation C^l and query representation Q^l separately in the encoder-decoder interaction sub-layer, and C^l and Q^l are concatenated as input of the feed-forward sub-layer to obtain the final decoder hidden state S^l. At each time step t, the decoding probability $P(w)$ is computed by fusing the information from c, q and last decoding layer hidden state S_t. The copy mechanism is used to predict the next target word according to $P(w)$, which is computed
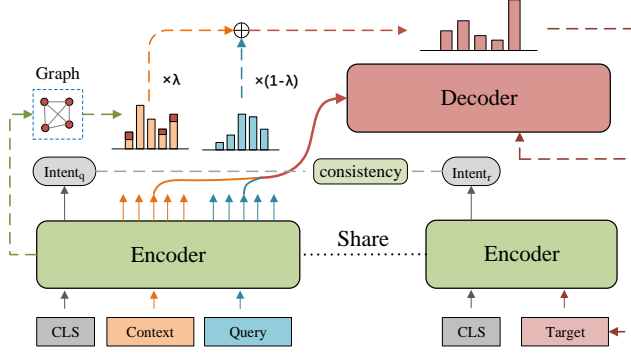
Fig. 2. The framework of our propose model **Teresa**.

as follows:

$$P(r_t = w) = \lambda \sum_{i:(w_i=w,\in c)} a_{t,i} + (1-\lambda) \sum_{j:(w_j=w,\in q)} a'_{t,j} \quad (1)$$

$$a_t = Attention(M_t, H_c) \quad (2)$$

$$a'_t = Attention(M_t, H_q) \quad (3)$$

$$\lambda = \sigma(w_S^T S_t + w_C^T C_t + w_Q^T Q_t) \quad (4)$$

where w_S, w_C, and w_Q are trainable parameters. σ is the sigmoid function to output a value between 0 and 1. λ is a learning coefficient to decide whether to copy token from c or q. Note that all tokens in r can only be copied either from the context c or query q. The rewriting model is trained by maximizing the log-likelihood of the output rewritten query.

$$\mathcal{L}_{NLL} = -\frac{1}{T} \sum_{t=0}^{T} log P(r_t) \quad (5)$$

Self-Attentive Keywords Detection. To facilitate carrying the omitted important information from context into the rewritten query, we propose to enhance the copy mechanism by a novel Self-Attentive Keywords Detection module (SAKD). Inspired by Xu et al [22], we build a graph network by stacking a self-attention layer over the output of context encoder. Formally, let $G = (V, D)$ be a directed graph where vertices V is word set from context and edge $D_{i,j}$ represents the self-attention weight from word w_i to the word w_j. Then the TextRank algorithm [13] is adopted to calculate word importance $score$ based on graph G.

The word importance score can be seen as a prior information to indicate the salient information in dialogue context. It is incorporated into copy mechanism as an extra input to calculate the attention weights a_t of context words. To further ensure that important information is extracted by copy mechanism, the Kullback-Leibler (KL) divergence is adopt as an auxiliary loss to force the distribution of attention weights close to the prior importance $score$.

$$a_t = softmax(Attention(M_t, H_c) + w_{score}^T score) \quad (6)$$

$$\mathcal{L}_{SAKD} = KL(\frac{1}{T} \sum_{t=0}^{T} a_t, score) \quad (7)$$

Intent Consistency Constraint. Intent matters to query understanding in dialogue. We argue that the rewritten query should be consistent with the contextual query in the intent dimension. Therefore, we propose a novel Intent Consistency Constraint (ICC) to guide the rewriting process. In this module, the latent intent recognition task is equipped to learn the corresponding intent representation for the given contextual query. The latent intent recognition shares encoder parameters with the rewriting model. A special classification token [CLS] is inserted in front of the input sequence to collect the intent information of original query in the context. Similarly, the corresponding intent representation for the rewritten query can also be collected by the content of itself, because the rewritten query is self-contained. Then another KL divergence loss is adopted to keep the intent distributions consistency between the contextual query and the rewritten query.

To sum up, the total objective of our proposed model is to minimize the integrated loss:

$$\mathcal{L} = \mathcal{L}_{NLL} + \mathcal{L}_{SAKD} + KL(f(H_{CLS}^{c,q}), f(H_{CLS}^r)) \quad (8)$$

where f is a function to map text representations into intent distributions.

3. EXPERIMENTS

3.1. Datasets and Metrics

We carry out extensive experiments on two public datasets. First, we construct a new CQR dataset from scratch based on large-scale raw dialogue corpus JDDC [23]. The JDDC corpus contains more than 1 million real multi-turn conversations between users and customer service staffs in E-commerce scenario. The average turn number of dialogues is 20, indicating the contextual dependency is very common in the dialogues. By applying the SSL approach mentioned in Section 2.1, we generate the pre-training data *JDDC-CQR-10M*, which includes about 10 million *(c, q, r)* triplets. We generate the positive samples by SSL, and the negative samples by random sampling. To compare with the SL approach, we also annotate another 146,000 triplets manually, namely *JDDC-CQR-146K*. The ratio of positive and negative samples is 1:1 in above two datasets. The positive sample means r is different from q, and the negative sample means r is the same as q. For context c, we keep at most 5 utterances. Second, to compare with previous CQR models, we also conduct experiments on a public CQR dataset *Restoration-200K*, which was collected from open-domain conversations and manually annotated by [7]. We split both the *JDDC-CQR-146K* and *Restoration-200K* into train/dev/test sets. The *JDDC-CQR-10M* is only used for pre-training in our experiment.

For evaluation, we choose three automatic evaluation metrics of BLEU-4 [24], ROUGE-L [25], and Exact Match by

Model	B4	RG-L	EM(-)	EM(+)
JDDC-CQR-146K				
T-Ptr-λ [8]	73.34	84.71	96.55	32.40
PAC [7]	70.78	83.52	87.18	28.24
MLR [9]	68.53	81.12	92.65	19.08
Teresa w/ SL	73.71	84.90	96.60	33.07
Teresa w/ SSL	78.34	87.68	94.94	47.36
Teresa w/ SSL+SL	**79.62**	**88.78**	**97.59**	**50.82**
w/o SAKD	79.35	88.61	97.36	50.00
w/o ICC	78.81	88.25	97.07	48.69
Restoration-200K				
T-Ptr-λ [8]	74.73	88.65	86.63	53.63
PAC [7]	73.69	86.66	82.23	46.27
MLR [9]	71.99	86.74	82.42	48.01
Teresa w/ SL	**74.82**	**88.69**	**87.49**	**54.46**

Table 2. The experimental results on *JDDC-CQR-146K* and *Restoration-200K* datasets. **B4** and **RG-L** stand for BLEU-4 and ROUGE-L respectively. **EM(+)** and **EM(-)** represent EM percentage for positive and negative samples.

following previous works [8]. BLEU and ROUGE are widely used in generation tasks to measure the lexical similarity between generated utterance and ground-truth. Exact Match is a very strict metric which requires the generated utterance to be the same as the ground-truth.
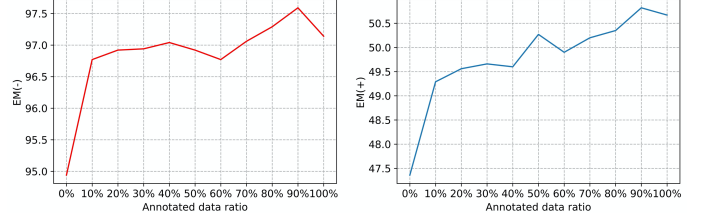
3.2. Baselines

Our baselines are as follows: (1) **T-Ptr-λ** [8]. This is a Pointer-Generator Network based on Transformer, which only copies words from context or query during generation. (2) **PAC** [7]. Pick-and-Combine (PAC) is a cascaded model that first identifies omitted words in context based on BERT [18], then appends the omitted words to the query as the input of Pointer-Generator Network. (3) **MLR** [9]. MLR is a two-stage CQR model with multi-task learning, which trains the sequence labeling task and query rewriting task jointly. All three baseline models are trained with the annotated training set (*aka. supervised learning*) in the following experiments.

For our proposed model **Teresa**, both the encoder and decoder consist of 6 layers of transformer block. The embedding dimension is set to 256 and the attention head number is 8. It is optimized with the Adam optimizer. The initial learning rate is 0.5 and batch-size is 64. Beam search is used for decoding and beam size is 4. For all the baselines, we follow the same experimental settings in the corresponding papers.

3.3. Experimental Results

Table 2 shows the experimental results on *JDDC-CQR-146K* and *Restoration-200K*. We can obtain following interesting conclusions: (1) For both two CQR datasets, with only the an-

Fig. 3. Performance analysis of fine-tuning experiments.

notated training data, Teresa w/ SL outperforms above three baselines on all metrics. It indicates the superiority of our proposed model. (2) Even with the pre-training data only, Terera w/ SSL has already outperformed Terera w/ SL significantly on **B4**, **RG-L** and **EM(+)**, which proves the effectiveness of SSL. For **EM(-)**, we argue it may be slightly effected by the random negative samples. (3) By utilizing the *pre-train and fine-tune* paradigm, Teresa w/ SSL+SL makes further improvement and obtains the best performance.

To figure out the contributions of SAKD and ICC, we conduct two groups of ablation study on *JDDC-CQR-146K*. From Table 2, it's observed that, by removing SAKD and ICC separately, both the performance drops notably, which demonstrates the necessity and rationality of each module.

Figure 3 illustrates the performance when we fine-tune Teresa with different percentages of annotated data. The two curves show that the performance improves very fast when adding only 10% of annotated data. Then it starts to saturate even adding more annotated data. This indicates much less of annotated data is needed with the help of SSL. We tried to plug our CQR model in the dialogue system and the experiments show that CQR can facilitate downstream tasks too.

4. CONCLUSIONS

In this paper, we propose a novel transformer-based generative model (denoted as Teresa) for conversational query rewriting, which is equipped by a novel self-attentive keywords detection module and an auxiliary intent consistency constraint. To address the time-consuming data annotation issue, we propose to construct the CQR training data via self-supervised learning automatically. Experiments on two CQR datasets demonstrate the superiority of SSL and the competitiveness of our proposed model. In the future, we will explore integrating the CQR task into pre-training stage of Pre-trained Language Model, and provide an universal pre-trained CQR model for various dialogue tasks.

5. ACKNOWLEDGEMENT

6. REFERENCES

[1] Heung-Yeung Shum, Xiao-dong He, and Di Li, "From eliza to xiaoice: challenges and opportunities with social chatbots," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, no. 1, pp. 10–26, 2018.

[2] Feng-Lin Li, Minghui Qiu, Haiqing Chen, Xiongwei Wang, Xing Gao, Jun Huang, Juwei Ren, Zhongzhou Zhao, Weipeng Zhao, Lei Wang, et al., "Alime assist: An intelligent assistant for creating an innovative e-commerce experience," in *CIKM*, 2017, pp. 2495–2498.

[3] Xiujun Li, Yun-Nung Chen, Lihong Li, Jianfeng Gao, and Asli Celikyilmaz, "End-to-end task-completion neural dialogue systems," in *IJCNLP*, 2017, pp. 733–743.

[4] Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan, "A neural network approach to context-sensitive generation of conversational responses," in *NAACL*, 2015, pp. 196–205.

[5] Qian Chen and Wen Wang, "Sequential attention-based network for noetic end-to-end response selection," *arXiv preprint arXiv:1901.02609*, 2019.

[6] Jaime G Carbonell, "Discourse pragmatics and ellipsis resolution in task-oriented natural language interfaces," in *ACL*, 1983, pp. 164–168.

[7] Zhufeng Pan, Kun Bai, Yan Wang, Lianqiang Zhou, and Xiaojiang Liu, "Improving open-domain dialogue systems via multi-turn incomplete utterance restoration," in *EMNLP-IJCNLP*, 2019, pp. 1824–1833.

[8] Hui Su, Xiaoyu Shen, Rongzhi Zhang, Fei Sun, Pengwei Hu, Cheng Niu, and Jie Zhou, "Improving multi-turn dialogue modelling with utterance rewriter," in *ACL*, 2019, pp. 22–31.

[9] Shuangyong Song, Chao Wang, Qianqian Xie, Xinxing Zu, Huan Chen, and Haiqing Chen, "A two-stage conversational query rewriting model with multi-task learning," in *WWW*, 2020, pp. 6–7.

[10] Xiyuan Zhang, Chengxi Li, Dian Yu, Samuel Davidson, and Zhou Yu, "Filling conversation ellipsis for better social dialog understanding," in *AAAI*, 2020, vol. 34, pp. 9587–9595.

[11] Qian Liu, Bei Chen, Jian-Guang Lou, Bin Zhou, and Dongmei Zhang, "Incomplete utterance rewriting as semantic segmentation," in *EMNLP*, 2020, pp. 2846–2857.

[12] Shi Yu, Jiahua Liu, Jingqin Yang, Chenyan Xiong, Paul Bennett, Jianfeng Gao, and Zhiyuan Liu, "Few-shot generative conversational query rewriting," in *SIGIR*, 2020, pp. 1933–1936.

[13] Rada Mihalcea and Paul Tarau, "Textrank: Bringing order into text," in *EMNLP*, 2004, pp. 404–411.

[14] Solomon Kullback, *Information theory and statistics*, Courier Corporation, 1997.

[15] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Zhaoyu Wang, Li Mian, Jing Zhang, and Jie Tang, "Self-supervised learning: Generative or contrastive," *arXiv preprint arXiv:2006.08218*, 2020.

[16] Longlong Jing and Yingli Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[17] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL*, 2019, pp. 4171–4186.

[19] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," in *ICLR*, 2015.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *NIPS*, 2017, pp. 6000–6010.

[21] Jiatao Gu, Zhengdong Lu, Hang Li, and Victor OK Li, "Incorporating copying mechanism in sequence-to-sequence learning," in *ACL*, 2016, pp. 1631–1640.

[22] Song Xu, Haoran Li, Peng Yuan, Youzheng Wu, Xiaodong He, and Bowen Zhou, "Self-attention guided copy mechanism for abstractive summarization," in *ACL*, 2020, pp. 1355–1362.

[23] Meng Chen, Ruixue Liu, Lei Shen, Shaozu Yuan, Jingyan Zhou, Youzheng Wu, Xiaodong He, and Bowen Zhou, "The jddc corpus: A large-scale multi-turn chinese dialogue dataset for e-commerce customer service," in *LREC*, 2020, pp. 459–466.

[24] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318.

[25] Chin-Yew Lin, "ROUGE: A package for automatic evaluation of summaries," in *ACL*, 2004, pp. 74–81.