Improving Open-Domain Dialogue Systems via Multi-Turn Incomplete Utterance Restoration

Zhufeng Pan[1,2*] **Kun Bai**[2*] **Yan Wang**[2 †] **Lianqiang Zhou**[2] **Xiaojiang Liu**[2]
[1]University of California, Los Angeles [2]Tencent AI Lab
panzhufeng@cs.ucla.edu
{bookerbai, brandenwang, tomcatzhou, kieranliu}@tencent.com

Abstract

In multi-turn dialogue, utterances do not always take the full form of sentences. These incomplete utterances will greatly reduce the performance of open-domain dialogue systems. Restoring more incomplete utterances from context could potentially help the systems generate more relevant responses. To facilitate the study of incomplete utterance restoration for open-domain dialogue systems, a large-scale multi-turn dataset *Restoration-200K*[1] is collected and manually labeled with the explicit relation between an utterance and its context. We also propose a "pick-and-combine" model to restore the incomplete utterance from its context. Experimental results demonstrate that the annotated dataset and the proposed approach significantly boost the response quality of both single-turn and multi-turn dialogue systems.

1 Introduction

Dialogue systems have attracted increasing attention due to the promising potentials on applications like virtual assistants or customer support systems (Hauswald et al., 2015; Poulami Debnath, 2018). However, studies (Carbonell, 1983) show that users of dialogue systems tend to use succinct language which often omits entities or concepts made in previous utterances. (also known as non-sentential utterances, (Fernández et al., 2005, 2007)). To make appropriate responses, dialogue systems must be equipped with the ability to understand these incomplete utterances.

Take Example 1 in Table 1 for instance, contents in parentheses are information omitted in the utterance. Humans are capable of comprehending those incomplete utterances based on previous utterances. For example, A_3 means what kind of *dessert* matches B's taste, instead of what kind of *shop* B likes. Failing to understand this utterance would be a disaster for dialogue systems to generate a relevant and coherent response. According to our survey (details in Table 2), in about 60% conversations, fully comprehending current utterance depends on previous context. We will refer conversation history (A_1 to B_2 in above example) as *previous utterances*, the utterance to be restored (A_3) as *original utterance*, and the complete form of A_3 as *restored utterance*.

Studies show that restoring the incomplete questions could help question-answering systems better understand users' intention (Raghu et al., 2015; Kumar and Joshi, 2017). It inspires us to improve the performance of open-domain dialogue systems via incomplete utterance restoration. However, most existing multi-turn dialogue datasets only provide sets of utterances, without any information about relations between utterances. In other words, they lack the necessary supervisions to restore the incomplete utterance.

To make dialogue systems better understand incomplete utterances, we collect a multi-turn conversation dataset from internet communities, and each of the conversations contains at least six utterances. Then we hire an annotation team to (1) label whether an utterance is related to its context or not, and (2) restore an incomplete utterance to a complete and context-free form based on its context. Finally, we get a high-quality and large-scale dataset with 200K annotated conversations. Such a dataset offers a new way of modelling utterance relations and improving the context-understanding ability of dialogue systems. We hope it would benefit the research of context understanding for multi-turn dialogue systems in the future.

With the annotated dataset above, we first attempt to restore the incomplete utterance using

[*] Both authors contributed equally to the work. The work was conducted when Zhufeng Pan was interning at Tencent AI Lab.

[†] Corresponding author

[1] The dataset is available at: https://ai.tencent.com/ailab/nlp/dialogue/datasets/

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 1824–1833, Hong Kong, China, November 3–7, 2019.

	Example 1	Example 2	Example 3
A_1	我能在巴黎哪个地方学做甜品？ Where can I learn to make dessert in Paris?	今天买了一堆**桌游**有爱**玩**的可以一起 I bought a bunch of **board game**. Welcome anybody who also likes to **play** it	我们一起过个情人节吧 Shall we spend Valentine's Day together
B_1	为什么(你想学做甜品)啊？ Why (do you want to learn to make dessert)?	我比较喜欢卡卡颂和现代艺术 I like Carcassonne and Modern Art	头像都一样在一起吧。 Let's date since we have the same avatar
A_2	因为我想在(巴黎)这儿开个甜品店 Because I want to open a dessert shop here (in Paris)	听说过不过没买 Heard of it. But I haven't bought it	在一起不错的选择 Dating is a good choice
B_2	(在巴黎开甜品店)不错啊，我很喜欢**甜点**！ (Opening a dessert shop in Paris) Sounds great, I love dessert!	我有 No problem, I have	赞 Cool
A_3	你最喜欢哪一种? Which kind matches your taste most?	一起啊 Let's do it together	人呢 Where are you?
Label	1	1	0
Reference	你最喜欢哪一种**甜品**? Which kind **of dessert** matches your taste most?	一起**玩桌游**啊 Let's **play board game** together	人呢 Where are you?

Table 1: Examples from the dataset. If annotators judge the fifth utterance (A_3) omits concepts or entities made in previous utterances like Example 1 and 2, they rewrite the utterance as ground truth reference.

two vanilla models, Sequence-to-Sequence model (Seq2Seq) and Pointer Generative Network. Then a cascaded pick-and-combine model is further proposed to first "Pick" omitted words from context and then "Combine" them with the incomplete utterance. To better evaluate the restoration performance, an evaluation metric is also designed. In the experiment, both automatic metrics and human evaluation show that the proposed approach could achieve promising results and significantly improve the response relevance of both single-turn dialogue systems and multi-turn dialogue systems.

The proposed approach enables single-turn dialogue systems with the capability to comprehend the dialogue context. It also facilitates multi-turn dialogue systems to model the relation between the query (original utterance) and context (previous utterances) explicitly in a supervised manner, in contrast to modelling the relation implicitly and without extra guidance (Serban et al., 2016; Wu et al., 2017).

Our contributions are summarized as below:

1) A large-scale Chinese dataset with 200K multi-turn conversations are collected and manually labeled with the explicit relations between an utterance and its context.

2) A cascaded pick-and-combine model is proposed, which achieves promising results on both automatic metrics and human evaluation.

3) Experimental results demonstrate that the incomplete utterance restoration model could be complementary to existing dialogue systems and is conducive for improving response quality.

In the remaining part, we first describe the collected dataset in Section 2. In Section 3, several models and a new metric are presented for incomplete utterance restoration. Section 4 shows experimental results on both automatic metrics and human evaluation for the proposed method. In Section 5 the proposed model is applied to dialogue systems to evaluate its effectiveness. Section 6 introduces related work. We conclude our work and discuss future directions in Section 7.

2 Restoration-200K

2.1 Data Collection

We collect open-domain dialogues from Douban group[2], a well-known Chinese online community which is a common data source for dialogue systems (Wu et al., 2017). To determine if one utterance is complete or not, we take four utterances before it as the conversation context. Crawled conversations with less than six (extra one to assist annotation) utterances are filtered out. For conversations with more than six utterances, only the first six are reserved. We construct a conversation by the identification of reply tags in comments under each post.

2.2 Data Annotation

To ensure the annotation quality, five professional data annotators are hired to annotate the dataset instead of using crowd-sourcing platforms like MTurk. It took six months to finish the annotation. As shown in Figure 1, annotators first discern whether the original utterance (the fifth utterance in conversation) omits concepts or entities made in previous utterances. Since the boundary

[2]https://www.douban.com/group

	train	val	test
# conversations	194k	5k	5k
Incomplete ratio (%)	60.1	59.4	58.8
Vocabulary size	80k	15k	15k
Avg. context length	25.9	25.8	25.7
Avg. utterance length	8.62	8.53	8.60
Avg. reference length	12.4	12.3	12.4

Table 2: Statistics of Restoration-200K. The incomplete ratio refers to the ratio of conversations that contains the incomplete utterance. Vocabulary size is counted after Chinese word segmentation and the average length is counted on character level.

between yes and no is vague under certain circumstances, annotators are allowed to skip and discard the instance if it is hard to make a decision. The sixth utterance is also provided to assist annotators in comprehending the conversation. When restoration is needed, they rewrite the original utterance to a context-free restored utterance which contains all necessary information for utterance understanding, as shown in the last row of Table 1.

To reduce the diversity of rewritten sentences, annotators are instructed to use words from previous utterances wherever possible. A small commonly used word list is provided, and annotators could use those words to ensure the fluency of rewritten sentences. In other words, all words in the restored utterance come from previous and original utterances or the extra word list. Our survey shows only about 4.8% of conversations that need to be restored in the dataset do not satisfy such condition. All of them are discarded as well.

2.3 Data Statistics

The statistics of train, validation and test set are shown in Table 2. Kumar and Joshi (2016) collect 6K incomplete questions for QA systems, where each utterance only consists of 3.52 words on average[3]. In comparison, *Restoration-200K* has a longer utterance length, which indicates covering a broader range of topics and more informative contents.

3 Methodology

In this section, we try to tackle the incomplete utterance restoration problem by using the vanilla Seq2Seq model with attention and pointer generative network. Since these two models are easily dominated by a simple copy mechanism that directly regenerates the original utterance as the restored utterance, we further propose a cascaded pick-and-combine (PAC) model to restore the original utterance.

[3]The average sentence length is estimated based on the released test set.

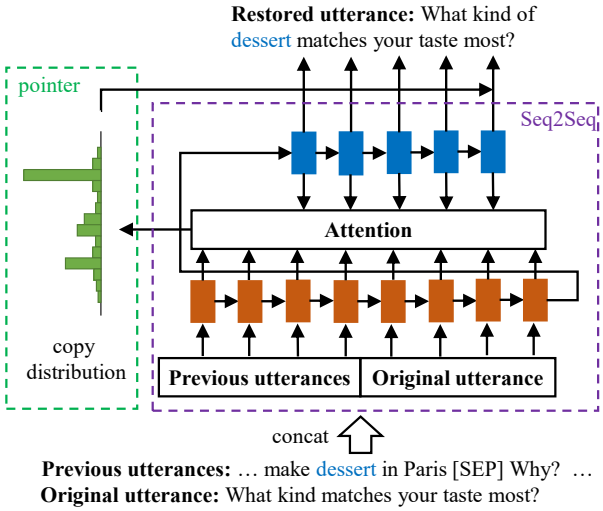

Figure 1: The architecture of Seq2Seq model and pointer generative network.

3.1 Vanilla Models

Seq2Seq Model with Attention: the Seq2Seq framework has been widely used in sequence generation tasks. The encoder encodes the input sequence into a vector representation c and the decoder generates the target sequence based on representation vector c, and all the previous words. As shown in Figure 1, to restore the incomplete utterance, we concatenate previous utterances C and the original utterance X as the input sequence. By inserting a special token $[sep]$ to separate C and X, the input sequence will be: $[c_1, \ldots, c_{t1}, [sep], x_1, \ldots x_{t2}]$.

Pointer Generative Network: as analyzed in Section 2.2, the incomplete utterance restoration problem has a unique characteristic: most generated words come from previous utterances or original utterance. To exploit this constraint, we propose using the pointer generative network to directly copy words from the input sequence. It has the capability of copying words from source sequence by directly taking the attention score a_k as prediction probability, as shown in the green part of Figure 1. The generation distribution P_{gen} can be calculated as: $P_{gen} = f(h_t^*, s_t, y_t)$, $p_{gen} = g(h_t^*, s_t, y_t)$. And the final vocabulary distribution

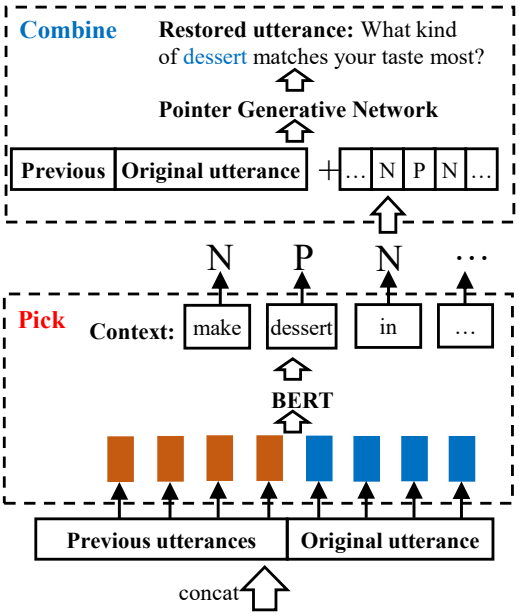

Figure 2: The pick stage predicts which words in previous utterances are omitted by the original utterance. P represents the label of positive, namely the omitted words. In combine stage, the picking result is appended to original conversation as extra guidance for the complete utterance sequence generation.

P_{vocab} is defined as:

$$P_{vocab} = p_{gen} * P_{gen} + (1 - p_{gen}) * P_{copy}$$

3.2 Pick-and-Combine Model

For most utterances to be restored in the corpus, the reference only differs from the original utterance in few words. Our study shows on average only 17.7% words in previous utterances overlap with the restored utterance, while 100% words in the original utterance are included in the restored utterance. Such *unbalanced* probability makes models mentioned above tend to simply regenerate the original utterance to maximize the conditional probability of the generated sentence during the beam-search process. In other words, the Seq2Seq model and pointer generative network tend to regenerate the original utterance and cannot effectively restore the original utterance. The third conversation in Table 5 is a typical example.

To mitigate this problem, we propose to decompose the incomplete utterance restoration task into a cascaded process. The first stage is the *Pick* process that identifies omitted words in previous utterances. The second is the *Combine* stage that restores the original utterance based on the identified omitted words.

Pick: inspired by recent advances in transfer learning for language representation, we fine-tune the pre-trained deep bidirectional transformers for language understanding model (BERT) (Devlin et al., 2019) as a classifier to select omitted words from previous utterances. Specifically, instead of discrete classifications, we first formulate this word selection problem as a sequence tagging problem. Each word in previous utterances will be determined to be positive P (should be identified as an omitted word) or negative N (not an omitted word).

Combine: the combine stage is straightforward. The selected omitted words are appended to the input sequence as extra guidance. And the two sequences are taken as input of the pointer generative network mentioned in Section 3.1. We also tried to fine-tune the BERT to directly generate the restored utterance, but the result is far from satisfactory. Therefore, we directly append selected words to the original input sequence as extra guidance for generating restored sentences.

3.3 Task Evaluation

BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) are commonly used metrics in machine translation and summarization. However, for the task of incomplete utterance restoration, these metrics do not differentiate words in the original utterance from those in previous utterances, and each gram is equally important. The statistics in Table 2 show that most words in the reference overlap with words in the original utterance. Thus, a simple copy mechanism which regenerates the original utterance as restored utterance would achieve a high score on these metrics.

To alleviate this issue, we propose the *restoration score* to evaluate the performance of incomplete utterance restoration model. This metric focuses on n-grams that contain at least one restored word, excluding other n-grams. Specifically, the n-gram restoration precision, recall, and F-score can be calculated as:

$$p_n = \frac{|\{\text{restored n-grams}\} \cap \{\text{n-grams in ref}\}|}{|\{\text{restored n-grams}\}|}$$

$$r_n = \frac{|\{\text{restored n-grams}\} \cap \{\text{n-grams in ref}\}|}{|\{\text{n-grams in ref}\}|}$$

$$f_n = 2 \cdot \frac{p_n \cdot r_n}{p_n + r_n}$$

Model	p_1	r_1	f_1	p_2	r_2	f_2	p_3	r_3	f_3	B_1	B_2	R_1	R_2
Syntactic	67.4	37.2	47.9	53.9	30.3	38.8	45.3	25.3	32.5	84.1	81.2	89.3	80.6
Seq2Seq	65.5	40.8	50.3	52.2	32.6	40.1	43.6	27.0	33.4	84.9	81.7	88.8	80.3
Pointer	66.6	40.4	50.3	54.0	33.1	41.1	**45.9**	28.1	34.9	84.7	81.7	89.0	80.9
PAC	**70.5**	**58.1**	**63.7**	**55.4**	**45.1**	**49.7**	45.2	**36.6**	**40.4**	**89.9**	**86.3**	**91.6**	**82.8**

Table 3: p, r, f here represent the restoration precision, recall and F-score that we propose in Section 3.3. B_n represents n-gram BLEU score and R_n represents n-gram ROUGE score.

where "restored n-grams" refer to the n-grams in restored utterance that contain at least one restored words, and "n-grams in ref" refer to the n-grams in reference that contain at least one restored words, $|X|$ refers to the number of elements in set X.

4 Experiment

4.1 Compared methods

We compare the performance of the following methods on the collected dataset:

- **Syntactic:** model proposed by Kumar (Kumar and Joshi, 2016). Specifically, they replace each out-of-vocabulary word in the corpus with a numbered unknown token. The number is based on its relative position in the conversation. The model itself is a Seq2Seq model with attention.
- **Seq2Seq:** the Seq2Seq model with attention introduced in Section 3.1.
- **Pointer:** the pointer generative network introduced in Section 3.1.
- **PAC:** the pick-and-combine model introduced in Section 3.2.

We adopt one layer unidirectional LSTM as the encoder and decoder of each model. During training, the vocabulary size is set to 10K. The size of each mini-batch is 64. Parameters are updated by Adam algorithm (P and Ba, 2014) with the betas set to 0.9 and 0.999 and the eps set to 1e-8. The learning rate is 0.25 and the clipping threshold of gradients is 0.1. The dropout rate is set to 0.5. The word embedding size, encoder hidden size and decoder hidden size are all set to 512. During the inference stage, the checkpoint with smallest validation loss is chosen and the beam-search size is set to 5 for all methods. For the pick stage in the proposed approach, we use a BERT model trained on 200G high-quality news data from Tencent AI Lab.

Method	Quality	Fluency
Syntactic	2.25	**2.58**
Seq2Seq	2.45	2.37
Pointer	2.53	2.50
PAC	**2.89**	2.51

Table 4: Human evaluation on the restoration quality and language fluency. The quality score adopts a 4-point scale, and fluency score adopts a 3-point scale. A higher score is better for both.

4.2 Evaluation metrics

The n-gram restoration score proposed in Section 3.3 is adopted as the automatic evaluation metrics. Then human evaluation is further conducted. Specifically, human annotators are instructed to evaluate the quality and fluency of restored utterances. The quality score evaluates if the restored utterances can be better understood without context than the original utterance. A restored utterance should be scored 4 if it can be perfectly understood without previous utterances, 3 if it is not perfect but still better than the original utterance, 2 if it hasn't been restored or it cannot help better the understanding, and 1 if it is worse than the original utterance. The fluency score is used to evaluate the fluency of a restored utterance. A 3-point scale is adopted and higher is better. We randomly select three hundred examples from the test set for human evaluation. Five professional annotators who have at least one year's experience on text annotation participate in the experiment.

4.3 Results

Automatic Metrics

Scores of each method on the automatic metric are shown on Table 3. PAC model achieves the best performance on the restoration precision, recall, and F-score from 1-gram to 3-grams. The restoration recall reflects how often does the model choose to restore the original utterance. As discussed in Section 3.2, Seq2Seq model and pointer

	Example 1	Example 2	Example 3
A_1	我这个样子也来相个亲喃 I am here for a blind date	晚上了在上**夜班** It's late. I am working in **night shift**	谁给你拍的 Who took it for you?
B_1	**会煮饭**加100分 **Being able to cook** will add 100 points for you	早安世界 Morning, the world	朋友抓拍 My friend took the snapshot
A_2	这个技能竟然100分 I can't believe this skill is worth 100 points	早 Morning	抓拍的够写意给了别人脑补的空间 The snapshot is so atmospheric that people could free their imagination
B_2	我喜欢**这个技能** I like **the skill**	乃是要通宵么 Are you going to work overnight?	**要的就是这个效果** **That is what I want**
A_3	我也喜欢希望能做好 I also like it and hope I could do well	上到8点钟 Until 8 a.m.	你确定吗我怎么觉得其实是没拍好 Are you sure? Why do I feel it was just a mistake
Reference	我也喜欢**会煮饭这个技能**希望能做好 I also like **the cooking skill** and hope I could do well	上**夜班**到8点钟 **The night shift** is until 8 a.m.	你确定**要的就是这种效果**吗我怎么觉得其实是没拍好 Are you sure **that is what you want**? Why do I feel it was just a mistake
PAC	我也喜欢**煮饭这个技能**希望能做好 I also like **the cooking skill** and hope I could do well	**夜班通宵**上到8点钟 **The night shift** requires **working overnight** until 8 a.m.	你确定**要的就是这种效果**吗我怎么觉得其实是没拍好 Are you sure **that is what you want**? Why do I feel it was just a mistake
Pointer	我也喜欢**这个技能**希望能做好 I also like **the skill** and hope I could do well	上到8点钟**通宵** **Working overnight** until 8 a.m.	你确定吗我怎么觉得其实是没拍好 Are you sure? Why do I feel it was just a mistake
Syntactic	我也喜欢**这个技能100分**希望能做好 I also like **the skill 100 points** and hope I could do well	上到8点钟 Until 8 a.m.	你确定吗我怎么觉得其实是没拍好 Are you sure? Why do I feel it was just a mistake
Seq2Seq	我也喜欢希望能做好 I also like it and hope I could do well	**通宵**上到8点钟 **Working overnight** until 8 a.m.	你确定吗我怎么觉得其实是没拍好 Are you sure? Why do I feel it was just a mistake

Table 5: Examples for incomplete utterance restoration. A_1 to B_2 are previous utterances and A_3 is the original utterance. The proposed PAC model could accurately identify pertinent words even those words appear at the beginning of the conversation like Example 2.

generative network tend to directly copy the original utterance, instead of taking risks to choose words from previous utterances. Restoration recall scores of these models can prove this claim. PAC model is also higher than other models in restoration precision. It demonstrates that the words selected by the PAC model are more accurate than that of other models.

We also report the BLEU and ROUGE score of each model. PAC model also achieves the highest scores in these metrics. However, scores of other models are comparatively high because these two metrics take each gram in the reference as equally important. It illustrates the necessity of the proposed evaluation metrics.

Human Evaluation

Results of the human evaluation are shown in Table 4. The proposed method has the highest quality score among all methods. We find the model with the higher score in human evaluation also has a higher restoration score. This tendency demonstrates that the proposed metric makes sense and is practical in this task. As for the fluency, all models except the Seq2Seq model achieve a similar score. The reason why the score of the PAC model is slightly lower than the syntactic model is that the fluency score follows a principle that the more you do, the more mistakes you make. Fluency of original utterances is high, so models with lower recall may get higher fluency.

4.4 Case study

Table 5 shows some examples of the incomplete utterance restoration among different models. At a closer look to the restored utterances, we find PAC model is capable of understanding the conversation context and correctly identify omitted concepts from multiple possible choices. In Example 1, none except PAC model can successfully restore the word "cooking" from previous utterances. In Example 2, most models can only extract the word "stay up" from previous utterances, and they fail to generate a fluent utterance. In contrast, the PAC model picks the correct words from previous utterances and generates a more fluent utterance. As for Example 3, only the PAC model can correctly restore the original utterance. This scenario also supports our analysis at the beginning of section 3.2 that vanilla models have a tendency to regenerate the original utterance.

We conclude an interesting phenomenon from this case study: only the PAC model owns the capability to extract omitted words from very early context. Other models merely choose words from the latest utterance exactly before the original one, namely the B_2 in Table 5. This great context comprehension ability is benefit from the pick process that identifies words from previous utterances.

5 Improving Existing Dialogue Systems

The main motivation for incomplete utterance restoration is to facilitate dialogue systems better

understanding conversation context and generate better responses. Thus, we conduct another experiment to evaluate if the response quality of single-turn and multi-turn dialogue systems improves after incomplete utterance restoration.

5.1 Settings

We applied the proposed PAC model to a single-turn and a multi-turn response generation model to evaluate its effectiveness on dialogue systems. Details are shown as follows:

- **MMI:** a state-of-the-art single-turn response generation model (Li et al., 2016). It adopts the maximum mutual information (MMI) as the objective function to prevent generating general responses. Note MMI model does not have access to the dialogue context. It only takes the original utterance as the input.

- **SMN:** the sequential matching network (SMN) is a state-of-the-art multi-turn reranking modeling proposed by Wu et al. (2017). The query expanded with keywords is used to retrieve top 100 response candidates. When reranking, SMN first matches a response with each utterance and distills matching information into a vector. The vectors are then accumulated through a recurrent neural network (RNN). The hidden states of the RNN is used to calculate the final matching score. In our experiments, SMN is applied to rerank candidates by taking the original utterance and conversation context as model inputs. The candidate with the highest matching score is chosen as the response.

- **MMI after Restoration:** The MMI model takes the restored utterance as the input.

- **SMN after Restoration:** The SMN model takes the restored utterance and conversation context as inputs.

Three hundred randomly selected conversations from the test set are selected and restored by all models. Five annotators are asked to evaluate whether responses to restored utterances are more relevant and appropriate than those of the original utterances. There are three choices for the annotators: better, similar or worse. If the response gets more appropriate *or* more relevant to the context, it will be judged as "better". If the response quality does not change much after the restoration, it will be regarded as "similar". Otherwise, it will be judged as "worse". Utterances that are not restored are judged as "Not Restored" (NR).

MMI	Better	Similar	Worse	NR
Syntactic	14.97	12.16	**2.81**	70.06
Seq2Seq	19.45	15.95	5.32	59.28
Pointer	21.25	16.51	5.35	56.89
PAC	**28.82**	21.95	6.12	**43.11**

Table 6: Human evaluation on response quality from the single-turn dialogue system after restoration. All numbers are in percentage (%), and NR represents *Not Restored*. Both are same for Table 7.

SMN	Better	Similar	Worse	NR
Syntactic	13.17	10.18	**6.59**	70.06
Seq2Seq	13.77	18.56	8.38	59.28
Pointer	16.17	16.76	10.18	56.89
PAC	**27.74**	17.37	11.78	**43.11**

Table 7: Human evaluation on response quality from the multi-turn dialogue system after restoration.

Note that we also train a generative multi-turn response generation model (Serban et al., 2016) on the collected dataset. But generated responses are not clear and coherent enough to understand. As a result, no human study is further conducted for such approach.

5.2 Results

Evaluation results of single-turn and multi-turn dialogue systems are shown in Table 6 and Table 7, respectively. Both systems get better responses after the restoration process, and PAC model is the most effective one. The PAC model achieves the lowest NR, which means it can restore more utterances than other models. Meanwhile, for single-turn dialogue systems, more than 50% queries get better responses when they are restored, and only 10% get worse responses. For multi-turn dialogue systems, although the SMN already has the ability to understand conversation context, the restoration still improves the response quality of nearly half of quires, and only about 20% responses get worse. The results prove that the proposed dataset and approach significantly improve the response quality.

	Example 1	Example 2	Example 3
A_1	下午去哪儿逛逛我的小单反 Where shall I go with my DSLR camera this afternoon	明天就直奔**西安**来了 Heading for **Xi'an** tomorrow	这是啤酒节啊 What a beer festival
B_1	火车站 Rail station	最近雨后会暴热 西安的定律 It''l become very hot after the rain. It''s a rule in Xi'an.	只有我一个人关注啤酒不错的吗 Am I the only one who finds the beer seems to be great?
A_2	不去那人太多了 No, there are too many people	要不要这么悲催的 How miserable it is	你也可以去的 You can go as well
B_2	哈哈**拍人头** Hah, you can **take photos of people**	肿么了你还没到 What, you haven't arrived yet?	**不喝酒** **No drinking**
A_3	(拍人头)不行不行镜头达不到 No, the lens is not good enough (for taking photos of people)	明早到(西安)呢现在在家 Will arrive (in Xi'an) tomorrow. Now I am home	(不喝酒)你可以去吃东西 (If no drinking,) You can go and eat something
MMI	你达不到？ Are you not good enough?	你在哪上班呢? Where do you work?	*没东西吃 *There is nothing to eat
MMI+PAC	*那就别拍了 *Then try to take photos of something else	*你在西安干嘛呢? *What are you doing in Xi'an?	我不喝酒的 I do not drink
SMN	不行呗 Just cannot do it	你现在好好的啊 You look cool now	*啥东西吃? *What to eat?
SMN+PAC	*不行也得行 *Yes, you can do it	*带你高校一日游外加兴庆公园赏雪哈哈 *I will take you to enjoy the one day trip to colleges and the beauty of snow in Park of Xingqing Palace	不喝酒那喝什么? What to drink if not the beer?

Table 8: A comparison between taking original and restored utterance as the input query to a response generation model. In A_3, words in parentheses are restored by PAC model. Responses annotated with * are better.

Some examples are shown in Table 8. Example 1 and 2 are cases where the response quality improves after incomplete utterance restoration. One interesting response is the one from *SMN+PAC* in Example 2. Note that the Park of Xingqing Palace is a famous historic relic in Xi'an. The dialogue system successfully generates a much more concrete and relevant response after restoring the city name *Xi'an*. There are also cases where responses become worse even if the model correctly restores omitted words, as shown in Example 3. One possible explanation is in this case the restored part is less important than the original utterance. This motivates us to find more strategies to alleviate this issue in future studies, such as training the restoration model and response generation model in an end-to-end manner or deploying a ranking model to choose the better response.

6 Related Work

6.1 Non-sentential Utterance Resolution

In addition to a further corpus and taxonomy study, Fernández et al. (2007) design a series of linguistic features to determine the NSU class. Dragone (2015); Dragone and Lison (2016) extend these features for NSU classification. Raghu et al. (2015) propose generating NSU resolution from templates. Poulami Debnath (2018) design a set of rules to classify and restore NSUs for customer support systems. All methods above heavily rely on the syntactic structure or frequent patterns observed empirically, which may not scale well for unseen domains. Kumar and Joshi (2016) approach NSU resolution as a Seq2Seq learning problem. One type of NSU that draws extra attention is ellipsis (Merchant, 2016; Kenyon-Dean et al., 2016; McShane and Babkin, 2016).

6.2 Multi-turn Dialogue Systems

Many efficient approaches have been proposed for developing intelligent dialogue systems (He et al., 2017; Shang et al., 2018; Tian et al., 2019; Cai et al., 2019). For multi-turn dialogue systems, Serban et al. (2016) and Xing et al. (2018) adopt hierarchical neural networks to model context. Tian et al. (2017) conduct an extensive comparison among existing methods and found context information is conducive for neural networks to generate longer, more meaningful and diverse replies. Wu et al. (2017) and Zhang et al. (2018) utilize the context information via a matching network and rerank the retrieved responses. Zhou et al. (2018) investigate matching a response with its multi-turn context using dependency information based entirely on attention. Most of these studies model the dialogue context information as an extra input in the response generation or response matching process. Our method, on the other hand, tries to understand the context by restoring the incomplete utterance to a context-free and complete form. It can be complementary to above studies.

7 Conclusion

In this paper, we propose to facilitate the comprehension of conversation context by restoring incomplete utterances. A large-scale dataset Restoration-200K, which consists of multi-turn

conversations for open-domain dialogue systems, is collected and manually annotated. Based on this dataset, the proposed pick-and-combine method achieves promising results on the incomplete utterance restoration task. Experimental results also demonstrate that the annotated dataset and the proposed approach significantly improve response quality for existing dialogue systems.

References

Deng Cai, Yan Wang, Wei Bi, Zhaopeng Tu, Xiaojiang Liu, Wai Lam, and Shuming Shi. 2019. Skeleton-to-response: Dialogue generation guided by retrieval memory. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1219–1228.

Jaime G Carbonell. 1983. Discourse pragmatics and ellipsis resolution in task-oriented natural language interfaces. In *Proceedings of the 21st annual meeting on Association for Computational Linguistics*, pages 164–168. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Paolo Dragone. 2015. Non-sentential utterances in dialogue: Experiments in classification and interpretation. Master's thesis, Sapienza University of Rome.

Paolo Dragone and Pierre Lison. 2016. Classification and resolution of non-sentential utterances in dialogue. *Italian Journal of Computational Linguistics*, 2(1):45–62.

Raquel Fernández, Jonathan Ginzburg, and Shalom Lappin. 2005. Using machine learning for non-sentential utterance classification. In *6th SIGdial Workshop on Discourse and Dialogue*.

Raquel Fernández, Jonathan Ginzburg, and Shalom Lappin. 2007. Classifying non-sentential utterances in dialogue: A machine learning approach. *Computational Linguistics*, 33(3):397–427.

Johann Hauswald, Michael A Laurenzano, Yunqi Zhang, Cheng Li, Austin Rovinski, Arjun Khurana, Ronald G Dreslinski, Trevor Mudge, Vinicius Petrucci, Lingjia Tang, et al. 2015. Sirius: An open end-to-end voice and vision personal assistant and its implications for future warehouse scale computers. In *ACM SIGPLAN Notices*, volume 50, pages 223–238. ACM.

He He, Anusha Balakrishnan, Mihail Eric, and Percy Liang. 2017. Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1766–1776.

Kian Kenyon-Dean, Jackie Chi Kit Cheung, and Doina Precup. 2016. Verb phrase ellipsis resolution using discriminative and margin-infused algorithms. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1734–1743.

Vineet Kumar and Sachindra Joshi. 2016. Non-sentential question resolution using sequence to sequence learning. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2022–2031.

Vineet Kumar and Sachindra Joshi. 2017. Incomplete follow-up question resolution using retrieval based sequence to sequence learning. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 705–714. ACM.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Marjorie McShane and Petr Babkin. 2016. Detection and resolution of verb phrase ellipsis. *LiLT (Linguistic Issues in Language Technology)*, 13.

Jason Merchant. 2016. Ellipsis: A survey of analytical approaches. *Handbook of ellipsis. Oxford: Oxford University Press. Ms. University of Chicago*.

Kingma Diederik P and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Harshawardhan M. Wabgaonkar Poulami Debnath, Shubhashis Sengupta. 2018.

Dinesh Raghu, Sathish Indurthi, Jitendra Ajmera, and Sachindra Joshi. 2015. A statistical approach for non-sentential utterance resolution for interactive qa system. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 335–343.

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Mingyue Shang, Zhenxin Fu, Nanyun Peng, Yansong Feng, Dongyan Zhao, and Rui Yan. 2018. Learning to converse with noisy data: Generation with calibration. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4338–4344. International Joint Conferences on Artificial Intelligence Organization.

Zhiliang Tian, Wei Bi, Xiaopeng Li, and Nevin L Zhang. 2019. Learning to abstract for memory-augmented conversational response generation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 3816–3825.

Zhiliang Tian, Rui Yan, Lili Mou, Yiping Song, Yansong Feng, and Dongyan Zhao. 2017. How to make context more useful? an empirical study on context-aware neural conversational models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 231–236.

Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 496–505.

Chen Xing, Yu Wu, Wei Wu, Yalou Huang, and Ming Zhou. 2018. Hierarchical recurrent attention network for response generation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao, and Gongshen Liu. 2018. Modeling multi-turn conversation with deep utterance aggregation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3740–3752.

Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu. 2018. Multi-turn response selection for chatbots with deep attention matching network. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1118–1127.