Improving Knowledge Production Efficiency With Question Answering on Conversation

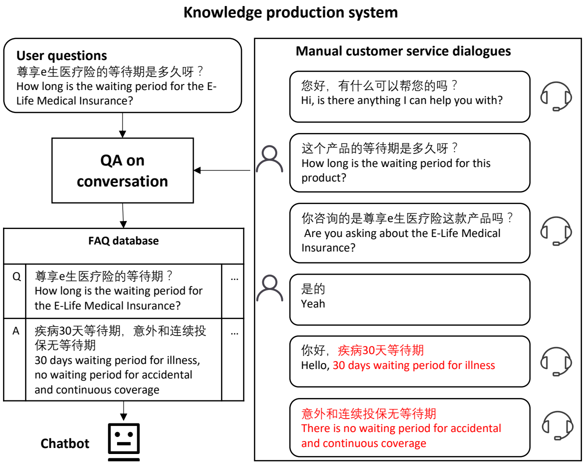
Changlin Yang, Siye Liu, Sen Hu, Wangshu Zhang
Teng Xu, Jing Zheng
Ant Group
`{changlin.ycl,siye.lsy,hs272483, wangshu.zws}@antgroup.com`
`{harvey.xt,jing.zheng}@antgroup.com`

Abstract

Through an online customer service application, we have collected many conversations between customer service agents and customers. Building a knowledge production system can help reduce the labor cost of maintaining the FAQ database for the customer service chatbot, whose core module is question answering (QA) on these conversations. However, most existing researches focus on document-based QA tasks, and there is a lack of researches on conversation-based QA and related datasets, especially in Chinese language. The challenges of conversation-based QA include: 1) answers may be scattered among multiple dialogue turns; 2) understanding complex dialogue contexts is more complicated than documents. To address these challenges, we propose a multi-span extraction model on this task and introduce continual pre-training and multi-task learning schemes to further improve model performance. To validate our approach, we construct two Chinese datasets using dialogues as the knowledge source, namely *ant-qaconv* and *kd-qaconv*, respectively. Experimental results demonstrate that the proposed model outperforms the baseline on both datasets. The online application also verifies the effectiveness of our method. The dataset *kd-qaconv*[1] will be released publicly for research purposes.

1 Introduction

With the rapid advance of Natural Language Processing (NLP), customer service chatbots have been widely applied in industries, as they can significantly reduce the cost of human customer service. Retrieval-based question answering model often plays an essential role in a chatbot system. However, building and maintaining a high-quality Frequently Asked Question (FAQ) database is labor-intensive, which relies on human experts to produce the answers. In *Alipay*'s online customer service system, there are many dialogues between customers and service agents collected daily, which contain customer questions and corresponding answers. To utilize these data, we design a knowledge production system, as depicted in Figure 1, to improve the efficiency of building the FAQ database. Our system extracts appropriate answers from the retrieved conversations for user questions that the chatbot cannot answer due to lacking relevant QA pairs in the FAQ database. After updating the FAQ database with produced QA pairs, the chatbot can answer the user questions correctly. The core of our system is the *QA on conversations* module, which extracts the answer from the candidate dialogues for a given question.


Knowledge production system
User questions
QA on conversation
FAQ database
Q
A
...
Chatbot
Manual customer service dialogues

尊享e生医疗险的等待期是多久呀？
How long is the waiting period for the E-Life Medical Insurance?

尊享e生医疗险的等待期？
How long is the waiting period for the E-Life Medical Insurance?

疾病30天等待期，意外和连续投保无等待期
30 days waiting period for illness, no waiting period for accidental and continuous coverage

您好，有什么可以帮您的吗？
Hi, is there anything I can help you with?

这个产品的等待期是多久呀？
How long is the waiting period for this product?

你咨询的是尊享e生医疗险这款产品吗？
Are you asking about the E-Life Medical Insurance?

是的
Yeah

你好，疾病30天等待期
Hello, 30 days waiting period for illness

意外和连续投保无等待期
There is no waiting period for accidental and continuous coverage

Figure 1: System for QA on conversation

Current QA researches mainly focuses on document-based QA tasks, such as SQuAD (Rajpurkar et al., 2016), or conversational QA tasks (Reddy et al., 2019), which need to answer sequential dialogue-like questions based on the understanding of the given document, instead of QA on conversation. There are only a few relevant datasets, such as FriendsQA (Yang and Choi, 2019), Molweni (Li et al., 2020), QAConv (Wu et al., 2022), whose dialogues are in English and col-

[1] https://github.com/yclzju/kd-qaconv

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 5: Industry Track, pages 225–234
July 10-12, 2023 ©2023 Association for Computational Linguistics

lected from emails, TV shows, with the focus mainly on multi-party dialogues and speaker information. In contrast, our main objective is to extract knowledge from two-party dialogues, which are more common in the customer service industry.

Therefore, in this work, we construct *ant-qaconv*, a QA dataset consisting of human customer service dialogues. And to better validate our proposed method, we build another dataset, *kd-qaconv*, based on the public dataset *kdconv* (Zhou et al., 2020). The main challenges for our task, as reflected in the datasets, include: 1) knowledge information can be distributed in multiple turns rather than in a single sentence in a document, which means the answer can comprise multiple spans of text, 2) it's difficult to model the hierarchical structure of a complex dialogue. Taking Figure 1 as an example, to answer the question "How long is the waiting period for the E-Life Medical Insurance?", the QA model must understand the context, including clarifications and co-references, then extract the answer from multiple turns of the dialogue. To address these challenges, based on a span-based machine reading comprehension model, we introduce a tag-based module to handle the multi-span challenge and propose a key utterance selection auxiliary task and continual pre-training to improve dialogue modeling. Experimental results show improved accuracy over baseline models from the proposed approach. Our main contributions can be summarized as follows:

- We design a knowledge production system based on the QA on conversation module, which improves the efficiency of maintaining the FAQ database for a chatbot.
- We introduce a construction pipeline and release the resulting dataset *kd-qaconv*, which, to the best of our knowledge, is the first public Chinese dataset for QA on conversation.
- We apply a multi-span extraction model and further improve its performance through continual pre-training and multi-task learning, and validate the approach's effectiveness through extensive experiments on two datasets.

2 Related work

2.1 Pre-trained Language Models

Pre-trained language models (PLMs) such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and AlBERT (Lan et al., 2019) have achieved state-of-the-art performance on many NLP tasks, including question answering (QA) tasks. These models are based on a self-attention mechanism and Transformer architecture (Vaswani et al., 2017) and are pre-trained on large corpora, which enables them to encode texts into contextualized representations.

However, most of these models are pre-trained on general domain textual corpora, such as Wikipedia, News, or Books, which can be notably different from the writing style and domain-specific language used in customer service conversations. As a result, many researchers (Gururangan et al.) have found that training PLMs on domain-related dialog corpora can help improve the model's performance on dialogue-related and domain-specific downstream tasks.

2.2 Question answering

Question answering (QA) (Hirschman and Gaizauskas, 2001) is one of the most widely researched NLP tasks, which aims at providing correct answers to questions based on the given knowledge source.

Considering that dialogue is one of the primary forms of interaction and significantly different from the document, some researchers propose using dialogue as a knowledge source, named QA on conversation (Wu et al., 2022; Yang and Choi, 2019; Li et al., 2020). There are several unique challenges to QA on conversation: 1) information is scattered across multiple dialogue turns; 2) co-reference resolution is more difficult for understanding dialogues than documents. To alleviate such difficulties, Li and Zhao (2021) design self-supervised tasks on speaker prediction and key-utterance detection to capture salient information in long dialogues. Li and Choi (2020) propose several dialogue pre-training tasks, including utterance order prediction and mask language modeling, to learn both token and utterance embeddings to understand dialogue contexts better.

3 Dataset

The data collection pipeline (shown in Figure 2) includes three stages as follows:

3.1 Dataset Pre-processing

For *ant-qaconv*, we sample 4193 dialogues from the online customer service of our company and remove the personally identifiable information and

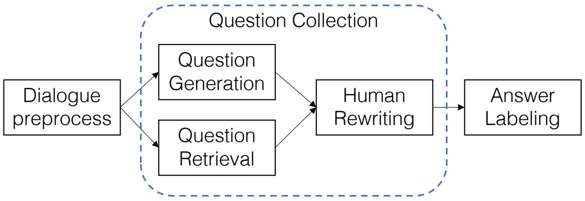


Figure 2: The data collection pipeline

non-text elements such as emojis and pictures.

For *kd-qaconv*, we choose *kdconv*, a public Chinese knowledge-driven conversation dataset about film, music, and travel, as the original dialogue data source.

3.2 Question Collection

We use three strategies to ensure the efficiency of the collection and quality of the candidate questions for each dialogue.

Question generation: Synthetic dataset construction has been proven effective in building robust and complex datasets (Feng et al., 2021). For *ant-qaconv*, we train the question generator (QG) with BART(Lewis et al., 2020)[2] on Dureader (He et al., 2018) and select one candidate answer extracted by an internal extractive dialogue summary model for each dialogue to generate candidate questions. Since each conversation in *kdconv* is annotated with multiple knowledge graphs (KG) triples, we train a question generator on KgCLUE[3] and randomly select two KG triples for each dialogue as the input of the QG model to generate candidate questions. Furthermore, we use a machine reading model trained from document datasets to filter out those generated questions with the predicted answer being used in QG since we suspect these questions may be too easy.

Question retrieval: For *ant-qaconv*, we also use BM25 (Robertson et al., 1995) to retrieve the most similar question from internal FAQ database as candidate question.

Human rewriting: For each dialogue, we ask the annotators to rewrite or remove the candidate questions with syntactic or semantic errors and try to write a new question that is different from the candidate question to ensure diversity of the questions.

3.3 Answer Labeling

For each sample, we ask internal annotators to read the questions and the dialogues and then label the answers, which can be nonexistent, a single text span, or multiple non-contiguous text spans in the dialogues.

3.4 Unanswerable questions

Previous work (Rajpurkar et al., 2018) show that unanswerable questions can force the model to decide whether a dialogue entails that a span of text can answer the question. To increase the number of unanswerable questions, we randomly sampled questions, and the annotators verified whether a text span was able to answer the selected question.

3.5 Dataset statistics

	ant-qaconv	*kd-qaconv*
dialogues	3895	4500
avg turns	18	19
avg dialogue's length	416	396
questions	6550	9384
- no-answer	855	769
- single-span	5262	7592
- multi-span	433	1023
avg answer's length	42	13
std answer's length	58	15

Table 1: Dataset statistics of *kd-qaconv* and *ant-qaconv*

Both datasets have been divided into training, development, and testing sets in an 8:1:1 ratio. The data sample of the emphkd-qaconv dataset can be found in Section efsec:appendix, while detailed data statistics are presented in Table eftab:dataset. It is worth noting that the conversations in the emphant-qaconv dataset are not strictly structured as a question-answer format, and therefore, we consider one utterance as one turn for statistics of conversation turns. Compared to existing datasets, our proposed datasets offer a wider range of answer types, with larger mean and standard deviation of answer length, making them more challenging.

4 Method

4.1 Task Formulation

We define dataset as $D = (C_m, q_m, a_m)_{m=1}^{M}$, the dialogue including k turns is represented as $C_m = (S_{m,1}, U_{m,1}), (S_{m,2}, U_{m,2}), ..., (S_{m,k}, U_{m,k})$, where $S_{m,i}$ and $U_{m,i}$ represents the ith speaker and

[2] https://huggingface.co/fnlp/bart-base-chinese

[3] https://github.com/CLUEbenchmark/KgCLUE

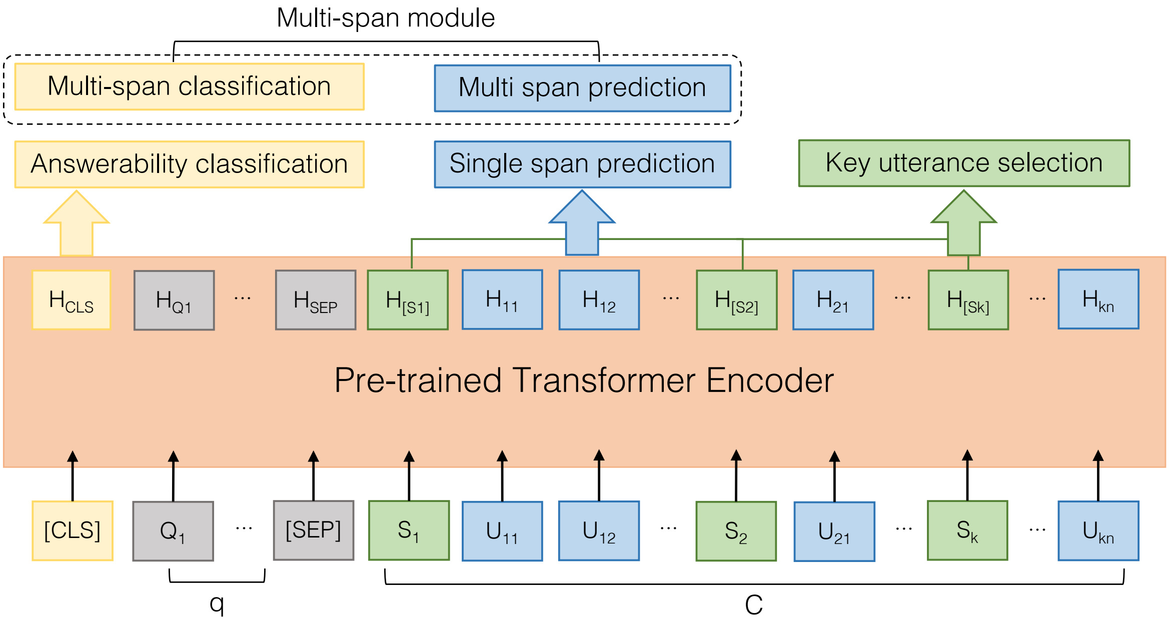


Figure 3: Overview of our model. The multi-span classification and answerability classification are answer-types classification modules. Tag-based multi-span prediction module and span-based single-span prediction module are answer prediction modules. The key utterance selection is an auxiliary task.

utterance of the mth dialogue respectively. For a given dialogue C and question q, the model needs to predict the corresponding answer a, which could be a no-answer, single-span or multi-span answer. The task can be formulated as $p(a|C, q)$.

4.2 Model

To better model the multiple types of answers, based on the span-based model, we apply the additional tag-based multi-span module and answer types classification modules. As shown in Figure 3, the model consists of the following components:

Transformer Encoder: The pre-trained transformer encoder aims to model the contextual feature representations of the question and dialogue. The input sequence of the encoder contains the question q and the flattened dialogue C, represented as X = [[CLS]; q; [SEP]; C], where X is the input token sequence, and C is the concatenation of each utterance U_k and corresponding speaker-id S_k of the kth utterance.

Answer types classification: The encoder representation $H_{[CLS]}$ of token [CLS] is treated as the dialogue level feature, which is used to predict the answerability and the multi-span classification by:

$$p^a = sigmoid(MLP^a(H_{[CLS]})) \quad (1)$$

$$p^m = sigmoid(MLP^m(H_{[CLS]})) \quad (2)$$

where MLP are multi-layer feed-forward networks, p^a is the prediction score of answerability, which means whether the question can be answered. p^m is the probability of whether the answer contains multiple spans. The loss function can be defined as follow, where y^a, y^m are the ground truths of answerability and multi-span classification, respectively:

$$L_a = -y^a \cdot log(p^a) - (1 - y^a) \cdot log(1 - p^a) \quad (3)$$

$$L_m = -y^m \cdot log(p^m) - (1 - y^m) \cdot log(1 - p^m) \quad (4)$$

Answer Prediction: The answer prediction task is to predict the answer text from dialogue context, including the single-span prediction and multi-span prediction.

$H_U = (H_{11}, ..., H_{ij}, ..., H_{kn})$ is a sequence of contextualized representations for all utterance tokens. The single-span prediction task is to compute the probability of each token being the start or the end of an answer span. Formally, feed-forward networks are used to calculate a score for each token, then a softmax function computes the start and end probability distributions along all tokens in this sequence:

$$p^s = softmax(MLP^s(H_U)) \quad (5)$$

$$p^e = softmax(MLP^e(H_U)) \quad (6)$$

The loss of single-span prediction can be defined as follows, where y_s and y_e mean the labeled start and end position of the answer respectively:

$$L_{span} = -\frac{1}{2}(log(p^s_{y_s}) + log(p^e_{y_e})) \tag{7}$$

The multi-span prediction task extracts a variable number of spans from the input dialogue utterances. We followed the method proposed by Segal et al. (2020), casting the task as a sequence tagging problem with IO tag schema, predicting for each token whether it should be part of the span or not. The probability of the model assigns to token ij having labeled tag T_{ij} is:

$$p^{ij}(T_{ij}) = softmax(MLP^t(H_{ij})) \tag{8}$$

and the loss function can be defined as follows, where N is the total count of tokens, n_i is the count of tokens in ith utterance's :

$$L_{tag} = -\frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} log(p^{ij}(T_{ij})) \tag{9}$$

4.3 Key Utterance Selection

Since each answer span is a subsequence of utterance, modeling the key utterance selection(KUS) can help locate the answer spans. Formally, $H_s = (H_{[s_1]}, ...H_{[s_i]}..., H_{[s_k]})$ is the sequence of representations of each utterance, which is the corresponding contextualized representations of the speaker-id tokens, the probability that the i-th utterance contains the answer can be calculated as follows:

$$p^{[s_i]} = sigmoid(MLP^K(H_{[s_i]})) \tag{10}$$

The loss can be defined as:

$$\begin{aligned} L_{KUS} = -\frac{1}{k}\sum_{i=1}^{k}(y_{[s_i]} \cdot log(p^{[s_i]}) \\ +(1 - y_{[s_i]}) \cdot (1 - log(p^{[s_i]})) \end{aligned} \tag{11}$$

$y_{[s_i]}$ represents the label for utterance $[s_i]$, $y_{[s_i]} = 1$ if it contains answer spans.

We adopt a multi-task learning scheme, which has been proven to be an effective way to improve model performance in NLP-related works (Chen et al., 2021). λ_a, λ_m, λ_{span}, λ_{tag}, λ_{KUS} are hyperparameters to control the weights of each task, and the model loss is defined as:

$$\begin{aligned} L_{model} = \lambda_{span} \cdot L_{span} + \lambda_{tag} \cdot L_{tag} \\ +\lambda_a \cdot L_a + \lambda_m \cdot L_m + \lambda_{KUS} \cdot L_{KUS} \end{aligned} \tag{12}$$

4.4 Continual Pre-training

The public pre-trained transformer models such as BERT (Devlin et al., 2019), RoBERTa(Liu et al., 2019) have demonstrated state-of-the-art(SOTA) performance in various NLP tasks. However, they are primarily trained on general domain textual corpus such as Wikipedia, News or Books, notably different from discourse structure, writing style and domain in customer service conversations. Therefore, we introduce the dialogue continual pre-training approach to help better model the dialogue structure.

For *ant-qaconv*, we collect about one million unlabeled customer service dialogues from the online customer service chat log. For *kd-qaconv*, we download and process several publicly released Chinese dialogue data, including *Duconv* (Wu et al., 2019), *Douban* (Wu et al., 2017), *Ecommerce* (Zhang et al., 2018), *DuRecDial* (Liu et al., 2020) and *LCCC* (Wang et al., 2020). Considering the hierarchical structure of the dialogue, we design the following token and utterance level pre-training tasks.

Masked language model(MLM): MLM is an essential task to achieve better contextualized representations. For BERT (Devlin et al., 2019), 15% of tokens are picked randomly. 80% of these tokens are replaced with "[MASK]", 10% are replaced with another random token, and 10% of the tokens are kept unchanged. For RoBERTa (Liu et al., 2019), we mask the whole word instead of the token (Cui et al., 2021). Then we compute the cross-entropy loss to predict the original token.

Response selection: Response selection is a classic dialogue pre-training task, which can enhance the contextual understanding of dialogue (He et al., 2022). The positive example (with label $l = 1$) is obtained by concatenating C with its corresponding response r in the original conversation. For negative samples, we randomly sample a response r^- from other dialogue. We feed the concatenated sequence of C and r into the transformer encoder and a binary classification head on the token [CLS]:

$$p(l = 1|C, r) = sigmoid(H_{[CLS]}) \tag{13}$$

to classify whether r is the proper response for context C. The cross-entropy loss is defined as:

$$\begin{aligned} L_{RS} = -log(p(l = 1|C, r) \\ -log(p(l = 0|C, r^-) \end{aligned} \tag{14}$$

5 Experiment

5.1 Experimental Setting

Following the standard evaluation metrics in the QA community, we choose word-level F1 and Exact Match(EM) accuracy as metrics to measure the overlap of the prediction and the ground truth answer(Rajpurkar et al., 2016). As the answer is long and descriptive in *ant-qaconv*, EM is unsuitable, so we choose F1 as our primary metric. To test our method, we choose bert-base-chinese[4] and chinese-roberta-wwm-ext[5] as our PLMs, which are widely used in Chinese NLP tasks.

Due to the 512 positional embedding limit of RoBERTa and BERT, truncating inputs by removing overflowing tokens can result in loss of contextual information and samples. To address this issue, we utilize a sliding window mechanism to construct training samples. During prediction, we select the prediction with the answer position in the middle. In our experiments, the sliding window size is set to 128.

For training, we use the Adam optimizercitekingma2014adam with default parameters and learning rates of 1e-5, and a batch size of 16 for 15 epochs. We select the best model based on F1 score on the development set and evaluate on the test set.

The hyper-parameters setting is shown in Table 2. If we choose not to add some sub-tasks, we just set the weights to 0.

symbols	tasks	weights
λ_m	answerability classification	0.3
λ_m	multi-span classification	0.3
λ_{span}	span model	0.6
λ_{tag}	tag model	0.6
λ_{KUS}	key utterance selection	0.3

Table 2: The weights of the loss for different sub-tasks

5.2 Experimental Results

Table 3 shows our experimental results on *kd-qaconv* and *ant-qaconv*. As shown in Table 4, we further analyze the model's F1 performance for different answer types in the *kd-qaconv* dataset. The tag-based models, such as *bert-tagger* and *roberta-tagger*, have been observed to perform well in multi-span cases, but not as effectively in single-span cases, which make up the majority of datasets. This may be due to the fact that the tag model needs to predict whether each token is part of the answer, which can be challenging to optimize, especially when the mean and variance of the length of answers in the datasets are large. In Table 4, the span-based models, *bert-span* and *roberta-span*, were chosen as our base models owing to their superior overall performance. We then sequentially added proposed modules to further improve the model's performance:

Model	ant-qaconv		kd-qaconv	
	F1	EM	F1	EM
bert-tagger	46.01	15.26	62.77	42.86
bert-span	66.58	34.34	79.35	65.53
+multi-span	67.04	36.55	82.33	69.98
+pre-training	69.19	**37.75**	84.17	72.05
+KUS task	**69.74**	37.55	**85.11**	**73.71**
roberta-tagger	52.09	18.88	68.96	50.31
roberta-span	67.19	36.75	82.54	69.15
+multi-span	68.55	38.15	85.76	75.26
+pre-training	**71.25**	**39.16**	86.31	76.29
+KUS task	69.75	37.15	**86.58**	**76.50**

Table 3: Result on *ant-qaconv* and *kd-qaconv*, KUS refers to the key utterance selection task.

Model	All	Single span	Multi span
bert-tagger	62.77	59.99	**79.74**
bert-span	79.35	86.33	55.03
+multi-span	82.33	85.75	74.72
+pre-training	84.17	88.33	76.46
+KUS task	**85.11**	**89.79**	74.89

Table 4: The models' F1 performances in single-span and multi-span cases in *kd-qaconv*.

First, The tag-based multi-span module can help handle the multi-span answers, thus improving F1 by 0.5% in *ant-qaconv*, 3% in *kd-qaconv*. The latter improves more because there are more multi-span cases in the *kd-qaconv* dataset.

Second, continual pre-training on *ant-qaconv* leads to improvements of 2.15% and 3.1% over BERT and RoBERT, respectively. The gain for *kd-qaconv* is 1.84% for BERT and 0.55% for RoBERTa. We suspect the improvements come from improved dialogue representation and domain adaptation. Pre-training improves *ant-qaconv* even

[4] https://huggingface.co/bert-base-chinese

[5] https://huggingface.co/hfl/chinese-roberta-wwm-ext

more because there is a more significant gap between internal customer service dialog and general domain text corpus used by the original PLMs. So the domain knowledge and complicated dialogue structure introduced by continual pre-training can be of incredible help for downstream QA tasks.

Third, the key utterance selection(KUS) task improves the model accuracy in most places except for RoBERTa on *ant-qaconv*, probably because it can help the model to identify which turn contains information to answer a given question.

In all, our method is effective on both datasets and with both BERT and RoBERTa as PLMs.

5.3 Application in Knowledge Production

We have deployed the proposed model in real-world knowledge production for the FAQ database used by *Alipay*'s online customer service chatbot in the following workflow: 1) A cluster module and a classifier process the chatbot log data to identify incorrectly answered user questions; 2) A retrieval model based on BM25 and Simcse (Gao et al., 2021) retrieves the most relevant dialogues from human customer service data; 3) The proposed QA model extracts the answers for user questions from the retrieved dialogues; 4) human operators decide whether to adopt the QA pairs into the FAQ database and they can also refine them.

Online Evaluation: We choose the adoption rate as our end-to-end accuracy. Based on the statistics of three months' data after the system deployment, the overall adoption rate is about 65%. We sample and analyze the QA pairs that are not adopted, only 8% of which are caused by inaccurate and incomplete extraction of the extraction model. The rest are caused by: retrieval mistakes, some queries or answers that are unclear or not suitable as the content of the FAQ database, etc. In the future, we will jointly optimize the knowledge production pipeline for better performance.

And we also choose knowledge production efficiency as our metric. The average time cost of producing a QA pair is reduced from 10 minutes to 2 minutes. When knowledge operators produce QA pairs directly, they must summarize answers from chat logs and related documents. However, with the assistance of our deployed system, they are only required to review and refine the recommended answers.

6 Conclusion

We design a knowledge production system including QA on conversations to help mine answers from human customer service dialogues. Based on the span-based extraction model, we add a multi-span extraction module trained with multi-task learning and continual pre-training schemes to extract incontiguous answers from conversational contexts. Experimental results show our approach outperforms the baseline models on both *ant-qaconv* and *kd-qaconv* datasets, the latter of which will be publicly released. Finally, the proposed method has been deployed to support the *Alipay*'s customer service chatbot system, which significantly saves the time cost of human operators' producing new QA pairs.

7 Ethical Considerations

We present the following ethical considerations for data authorization, privacy, and deployments.

- We have obtained explicit permissions from the customer to collect and utilize customer service dialog data.

- We use desensitization tools to remove sensitive information in customer service conversations. Only a few annotators can access this data, and they will check again to ensure that there is no user personal information in the dataset. At the same time, the dataset *ant-qaconv* is only used for internal research.

- The QA pairs produced by the knowledge production system will be checked by human operators to make sure private message is removed.

8 Acknowledgments

Ant Group is China's leading technology and financial services company that provides a range of digital financial solutions such as online payment and wealth management.

The authors would like to thank Changlin Yang, Siye Liu, Sen Hu, Wangshu Zhang, Teng Xu, Jing Zhen, and other members of Ant Group for helpful discussions and comments. We would also like to thank reviewers for their valuable comments.

References

Shijie Chen, Yu Zhang, and Qiang Yang. 2021. Multi-task learning in natural language processing: An overview. *arXiv preprint arXiv:2109.09138*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

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.

Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for nlp. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: Adapt language models to domains and tasks.

Wanwei He, Yinpei Dai, Yinhe Zheng, Yuchuan Wu, Zheng Cao, Dermot Liu, Peng Jiang, Min Yang, Fei Huang, Luo Si, et al. 2022. Galaxy: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10749–10757.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, et al. 2018. Dureader: a chinese machine reading comprehension dataset from real-world applications. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 37–46.

Lynette Hirschman and Robert Gaizauskas. 2001. Natural language question answering: the view from here. *natural language engineering*, 7(4):275–300.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Changmao Li and Jinho D Choi. 2020. Transformers to learn hierarchical contexts in multiparty dialogue for span-based question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5709–5714.

Jiaqi Li, Ming Liu, Min-Yen Kan, Zihao Zheng, Zekun Wang, Wenqiang Lei, Ting Liu, and Bing Qin. 2020. Molweni: A challenge multiparty dialogues-based machine reading comprehension dataset with discourse structure. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2642–2652.

Yiyang Li and Hai Zhao. 2021. Self-and pseudo-self-supervised prediction of speaker and key-utterance for multi-party dialogue reading comprehension. *arXiv preprint arXiv:2109.03772*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Towards conversational recommendation over multi-type dialogs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1036–1049, Online. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

Elad Segal, Avia Efrat, Mor Shoham, Amir Globerson, and Jonathan Berant. 2020. A simple and effective model for answering multi-span questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3074–3080.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yida Wang, Pei Ke, Yinhe Zheng, Kaili Huang, Yong Jiang, Xiaoyan Zhu, and Minlie Huang. 2020. A large-scale chinese short-text conversation dataset. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 91–103. Springer.

Chien-Sheng Wu, Andrea Madotto, Wenhao Liu, Pascale Fung, and Caiming Xiong. 2022. Qaconv: Question answering on informative conversations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5389–5411.

Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019. Proactive human-machine conversation with explicit conversation goal. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3794–3804.

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.

Zhengzhe Yang and Jinho D Choi. 2019. Friendsqa: Open-domain question answering on tv show transcripts. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 188–197.

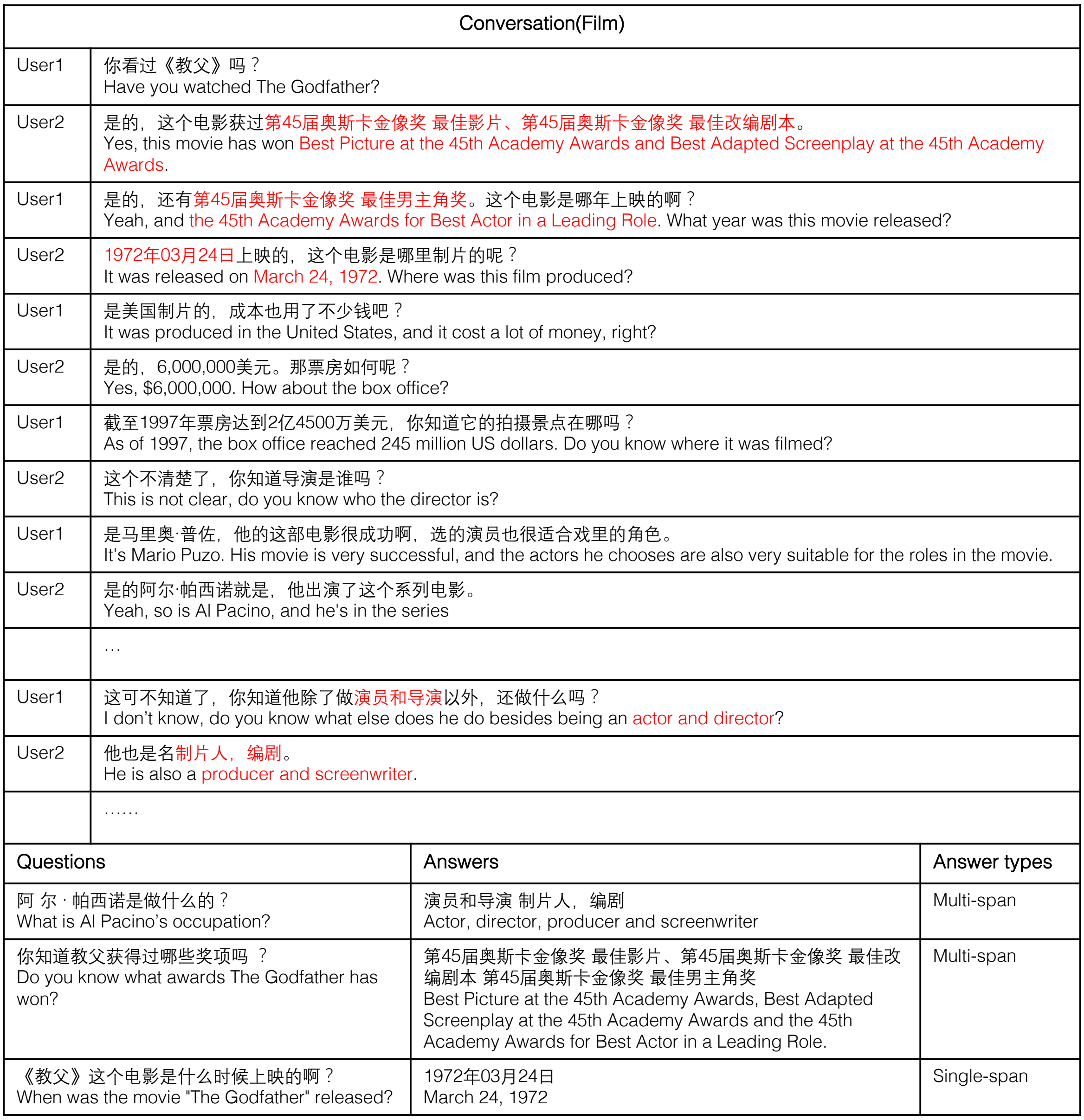
Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao, and Gongshen Liu. 2018. Modeling multi-turn conversation with deep utterance aggregation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3740–3752.

Hao Zhou, Chujie Zheng, Kaili Huang, Minlie Huang, and Xiaoyan Zhu. 2020. KdConv: A Chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7098–7108, Online. Association for Computational Linguistics.

A Appendix

A.1 Data samples

As shown in Figure 4, we sample a conversation and corresponding questions and answers from *kd-qaconv*.



Conversation(Film)	
User1	你看过《教父》吗？ Have you watched The Godfather?
User2	是的，这个电影获过第45届奥斯卡金像奖 最佳影片、第45届奥斯卡金像奖 最佳改编剧本。 Yes, this movie has won Best Picture at the 45th Academy Awards and Best Adapted Screenplay at the 45th Academy Awards.
User1	是的，还有第45届奥斯卡金像奖 最佳男主角奖。这个电影是哪年上映的啊？ Yeah, and the 45th Academy Awards for Best Actor in a Leading Role. What year was this movie released?
User2	1972年03月24日上映的，这个电影是哪里制片的呢？ It was released on March 24, 1972. Where was this film produced?
User1	是美国制片的，成本也用了不少钱吧？ It was produced in the United States, and it cost a lot of money, right?
User2	是的，6,000,000美元。那票房如何呢？ Yes, $6,000,000. How about the box office?
User1	截至1997年票房达到2亿4500万美元，你知道它的拍摄景点在哪吗？ As of 1997, the box office reached 245 million US dollars. Do you know where it was filmed?
User2	这个不清楚了，你知道导演是谁吗？ This is not clear, do you know who the director is?
User1	是马里奥·普佐，他的这部电影很成功啊，选的演员也很适合戏里的角色。 It's Mario Puzo. His movie is very successful, and the actors he chooses are also very suitable for the roles in the movie.
User2	是的阿尔·帕西诺就是，他出演了这个系列电影。 Yeah, so is Al Pacino, and he's in the series
	…
User1	这可不知道了，你知道他除了做演员和导演以外，还做什么吗？ I don't know, do you know what else does he do besides being an actor and director?
User2	他也是名制片人，编剧。 He is also a producer and screenwriter.
	……

Questions	Answers	Answer types
阿 尔 · 帕西诺是做什么的？ What is Al Pacino's occupation?	演员和导演 制片人，编剧 Actor, director, producer and screenwriter	Multi-span
你知道教父获得过哪些奖项吗 ？ Do you know what awards The Godfather has won?	第45届奥斯卡金像奖 最佳影片、第45届奥斯卡金像奖 最佳改编剧本 第45届奥斯卡金像奖 最佳男主角奖 Best Picture at the 45th Academy Awards, Best Adapted Screenplay at the 45th Academy Awards and the 45th Academy Awards for Best Actor in a Leading Role.	Multi-span
《教父》这个电影是什么时候上映的啊？ When was the movie "The Godfather" released?	1972年03月24日 March 24, 1972	Single-span

Figure 4: An example in *kd-qaconv*. The conversation is from *kdconv*(Zhou et al., 2020), and the questions and answers are constructed with our data collection pipeline. *kd-qaconv* is a Chinese QA on conversation dataset, and we also translate the content to English for better understanding.