Knowledge Based High-Frequency Question Answering in *AliMe Chat*

Shuangyong Song, Chao Wang, Haiqing Chen

Alibaba Group, Beijing 100102, China.
{shuangyong.ssy; chaowang.wc; haiqing.chenhq}@alibaba-inc.com [1]

Abstract. In our online chatbot serving, *AliMe Chat*, we design a knowledge graph based approach for solving high-frequency chitchat question answering. For meeting the demand of high Question per Second (QPS) of online system, we design several solutions to escape from questioning a large knowledge graph, details of those solutions are given in this paper, and the experimental results show the effectiveness and efficiency of them.

Keywords: Knowledge Graph, E-commerce Chatbot, Lucene Index, Text Matching, Multiple Answers Generation, Index of Subgraph.

1 Introduction

AliMe Chat, presented by Alibaba in 2015, has provided services for billions of users and now on average with ten million of users access per day [4]. *AliMe* service can be roughly classified into assistance service, customer service and chatting service, and the main idea of this paper is to improve ability of *AliMe Chat* with knowledge graph.

A seq2seq based re-ranking and generation method has been proposed in [1] to chat with *AliMe* users with general topics, such as greetings, jokes and other kinds of chitchats. However, fact-based and knowledge-based chatting ability of *AliMe* is still weak, and for improving those kinds of ability of *AliMe* and meanwhile increasing the diversity of chatting answers, we design a question-answering framework.

Since online servicing has a very high demand of QPS, our framework is just oriented to high-frequent questions or entities in historical user question logs. We design several methods: 1) for high-frequent questions, we try to find which of them can be answered with knowledge graph and those 'question-answer' pairs are indexed by Lucene for online matching and re-ranking; 2) for high-frequent entities, we extract subgraphs from complete knowledge graph, and differing from some related work which do this step in real time [3], we prepared those subgraphs offline for reducing online processing. We classify questions with those entities to 3 kinds: questions with an unambiguous entity, questions with an ambiguous entity and questions with multiple entities. For different kinds of questions, we design different answer generation methods.

In the following parts of this article, we will illustrate the details of the proposed framework, and report the experimental results.

2 Proposed Framework

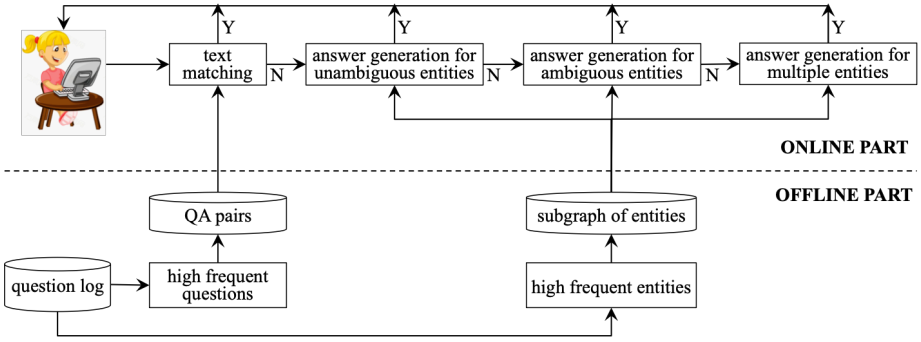

Fig. 1. The proposed framework.

Figure 1 shows the proposed framework, we will introduce it in detail with two parts: question-answering with high frequent questions and question-answering with high frequent entities.

2.1 Question-answering with high frequent questions

Text clustering is utilized to cluster users' question log and representative questions in top ranking clusters are extracted as high frequent questions. On the clustering step, we utilize a self-adapting clustering method proposed in [5] and set a strict threshold to ensure that questions in a cluster are very similar to each other. On the representative question extraction step, we consider cluster-level keywords, question length and distance to cluster center as three factors, and a question with more keywords, average question length and nearest distance to cluster center has more chance to be chosen as the final representative one.

A classic knowledge graph based question-answering technique [6] is for obtaining answers of each representative questions, and all questions with knowledge graph based answers are collected into a 'question-answer' index with Lucene and in the online part, we first use Lucene to roughly recall top K candidates and then use a deep learning based text similarity model [2] to exactly rank those candidates to get the final answer.

2.2 Question-answering with high frequent entities

Entities with high frequency are extracted from user question log, and then we categorize those entities to unambiguous entities and ambiguous entities. For unambiguous entities, we can answer questions such as "where was Joe Hisaishi born" easily with classic knowledge graph based question-answering technique [6]. And for questions with ambiguous entities, such as "you know Carlos, right?", we can answer this question with "you mean the Brazilian football player?" or "you mean the Brazilian football player or Carlos the Jackal?".

Especially, for a user question that contains more than one entity, such as a question "Who is older, Louis Koo or Andy Lau?", the method proposed in [7] is referred in our work.

3 Experiments

3.1 Dataset and Parameter Settings

Datasets:

Question log: we collect anonymous online user question log from Nov. 1, 2018 to Dec. 31, 2018. This dataset contains 125.9 million user questions and with merging duplicate ones we can obtain 44.9 million diverse user questions.

High frequent questions: occurrences of 1.26 million questions are greater than or equal to 5, which are chosen as high frequent questions (HFQs).

QA pairs: we input each HFQ into knowledge-based QA system, and if we can get an answer, we take this 'HFQ-answer' as a QA pair. We totally obtained 53,187 QA pairs.

High frequent entities: 25,682 high frequent entities (HFEs), more than 10.

Subgraph of entities: we extract all subgraphs of HFEs from Wikipedia.

Text matching training data: for creating enough dataset for training the text matching model, we implement following strategies: we randomly select 10,000 user questions from chatbot log, and top 15 candidates for each of them can be obtained with Lucene index of all question log. Then 8 service experts labeled those candidates with right/wrong, and some examples are shown in Table 1. Serious data unbalance shows in above labeled data, since just 14.3% candidates are labeled as right ones (positive samples). For balancing the data, we randomly extract about 20% candidates, which are labeled as wrong, of whole dataset as negative samples.

Parameter settings:

When choosing top K candidates from Lucene index, we empirical set $K = 20$, which is a number not too small to recall the real answer and not too huge to be quickly processed in the text matching step.

For the text-matching threshold, we check each decimal in (0,1) with an interval of 0.1, with respect to F1-value final answer obtaining, and a threshold of 0.85 can help obtain the best F1-value.

3.2 Experimental Results

The main purpose of the proposed framework is to increase the coverage of *AliMe Chat*, and reduce the 'no-answer' situations. With the real online testing, the coverage of *AliMe Chat* in the whole *Alime Assist* has been increased from 4.18% to 4.87%, which realizes a 16.5% increase.

In Fig. 2, we show several examples of online results of the proposed approach. In left sub-figure, the first user question is a frequent asked question and it can be answered with knowledge graph, so Lucene has indexed it. The second question contains a entity of 'East Hope' which has no ambiguity in knowledge graph and we can answer it with 'East Hope Group is a company' or 'East Hope Group is in electrolytic aluminum industry' etc. In right sub-figure, the first user question contains an ambiguous name 'James', which is also a 'half' person name. We can give user some choices of this ambiguous half name, and if then user choose one of the choices and ask some related question, we can continue to answer it.

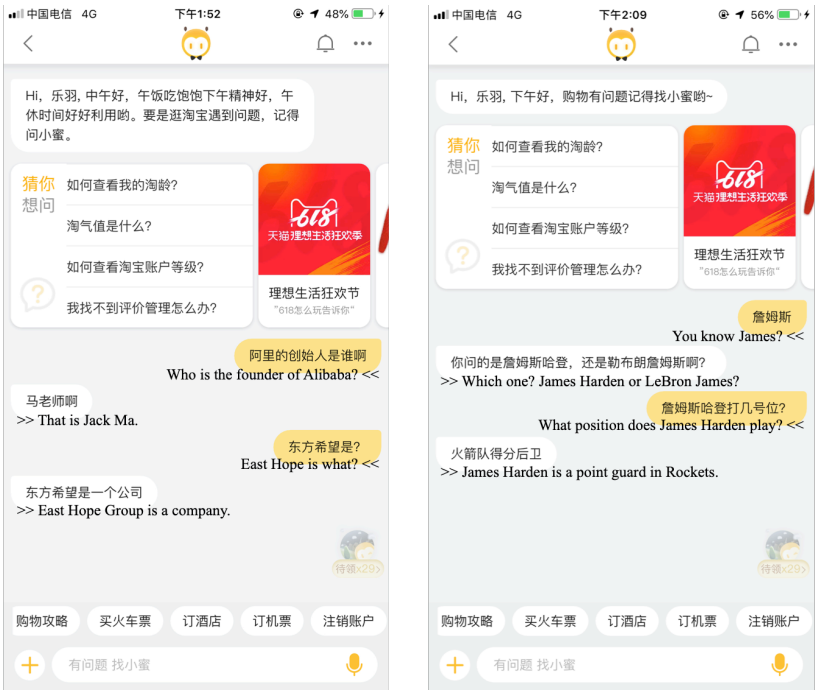

Fig. 2. Online *AliMe Chat* severing with proposed framework.

4 Future Works

This paper is only a preliminary work. Knowledge based multi-turn conversation in e-commerce chatbot will be a key point in our future work, and the utilization of knowledge based named entity disambiguation models, especially that on abbreviation disambiguation, are predictable to be a helpful way of getting better responses.

References

1. Qiu, M., Li, F., Wang, S., Gao, X., Chen, Y., Zhao, W., Chen, H., Huang, J. and Chu, W., *AliMe Chat*: A Sequence to Sequence and Rerank based Chatbot Engine. In ACL'17.
2. Yu, J., Qiu, M., Jiang, J., Huang, J., Song, S., Chu, W., Chen, H. Modelling Domain Relationships for Transfer Learning on Retrieval-based Question Answering Systems in E-commerce. WSDM 2018: 682-690.
3. Zafar, H., Napolitano, G., and Lehmann, J., Formal query generation for question answering over knowledge bases. In ESWC'18.
4. Li, F. L., Qiu, M., Chen, H., Wang, X., Gao, X., Huang, J., Ren, J., Zhao, Z., Zhao, W., Wang, L., Jin, G., Chu, W. AliMe assist: an intelligent assistant for creating an innovative e-commerce experience. In CIKM'17.
5. Song, S., Meng, Y., & Zheng, Z. Summarizing Microblogging Users with Existing Well-defined Hashtags. International Journal of Asian Language Processing 23(2):111-125.
6. Rinaldi, F., Dowdall, J., Hess, M., Mollá, D., Schwitter, R., Kaljurand, K. Knowledge-based question answering. In KES'03.
7. Dubey, M., Banerjee, D., Chaudhuri, D., Lehmann, J. Earl: Joint entity and relation linking for question answering over knowledge graphs. In ISWC 2018, pp. 108-126.