Beyond Static: Related Questions Retrieval Through Conversations in Community Question Answering

Xiao Ao[1], Jie Zou[1]*, Yibiao Wei[1], Peng Wang[1], Weikang Guo[2]*

[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China
[2]School of Management Science and Engineering, Southwestern University of Finance and Economics
202422080439@std.uestc.edu.cn, jie.zou@uestc.edu.cn, 202411081540@std.uestc.edu.cn, p.wang6@hotmail.com, guowk@swufe.edu.cn

Abstract

In community question answering (cQA) platforms like Stack Overflow, related question retrieval is recognized as a fundamental task that allows users to retrieve related questions to answer user queries automatically. Although many traditional approaches have been proposed for investigating this research field, they mostly rely on static approaches and neglect the interaction property. We argue that the conversational way can well distinguish the fine-grained representations of questions and has great potential to improve the performance of question retrieval. In this paper, we propose a related question retrieval model through conversations, called TeCQR, to locate related questions in cQA. Specifically, we build conversations by utilizing tag-enhanced clarifying questions (CQs). In addition, we design a noise tolerance model that evaluates the semantic similarity between questions and tags, enabling the model to effectively handle noisy feedback. Moreover, the tag-enhanced two-stage offline training is proposed to fully exploit the mutual relationships among user queries, questions, and tags to learn their fine-grained representations. Based on the learned representations and contextual conversations, TeCQR incorporates conversational feedback by learning to ask tag-enhanced clarifying questions to retrieve related questions more effectively. Experimental results demonstrate that our model significantly outperforms state-of-the-art baselines.

Code — https://github.com/AIT55/TeCQR

Introduction

Community question answering (cQA) platforms have become vital resources for knowledge sharing and acquisition (Chen et al. 2018; Ye, Xing, and Kapre 2017). On platforms such as Stack Overflow (SO), users often prefer to retrieve answers from resolved questions rather than post new queries and wait for replies. Therefore, effective question retrieval plays a critical role in enhancing the utility of cQA platforms.

For question retrieval in cQA, a wide range of methods have been proposed (Li, Du, and Chen 2020; Rücklé, Swarnkar, and Gurevych 2019; Zhang et al. 2014; Gao et al. 2023; Hazra et al. 2024; Liu et al. 2024). Most of these approaches (Lei et al. 2016; Zhang, Sun, and Wang 2018; Wang, Zhang, and Jiang 2019, 2020; Huang et al. 2023; Qin et al. 2022; Shirani et al. 2019; Kumari et al. 2021; Guo et al. 2022; Zhao et al. 2017; Ha et al. 2021; Zhao and Huang 2022; Liu et al. 2023) embed both the query and candidate questions as vectors, then measure their similarity in explicit or latent semantic space. However, queries in question retrieval are often short and ambiguous (Figure 1(a)), making accurate retrieval particularly challenging.

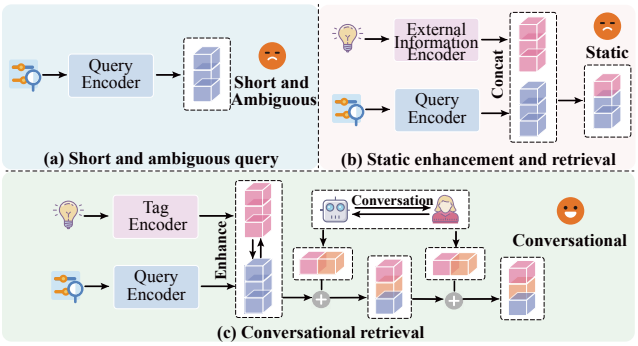


Figure 1: (a) Short and ambiguous query hinder intent understanding. (b) Static enhancement and one-off interaction may introduce unconfirmed information and fail with initially unclear user intent. (c) We iteratively incorporate user-confirmed information through tag-based clarifying questions to elicit more accurate user intent for retrieval.

To alleviate the above challenge, several studies have attempted to enhance query representations by incorporating external information, such as tags associated with the query (Hazra et al. 2024; Li et al. 2024), knowledge graphs (Liu et al. 2023), answerer authority (Zhao et al. 2017), answer content (Liang et al. 2019; Zhang and Wu 2018), and other contextual signals (Nie et al. 2016; Lu et al. 2015; Ghasemi and Shakery 2024; Wang et al. 2017, 2025). Although these methods have shown effectiveness in enriching query information, they typically rely on the static information enhancement and one-off query interaction to perform static retrieval (Figure 1(b)), yet still suffer from two limitations. First, external information that has not been confirmed by the user may be unreliable, or even misaligned with the

*Jie Zou and Weikang Guo are corresponding authors.

user's actual intent (e.g., when a user intends to understand the usage of *pop()* function in a dictionary but does not explicitly specify *dict*, the system may incorrectly associate *list* as external information, leading to irrelevant retrieval outcomes.). Such mismatches can lead the retrieval results to deviate from the user's actual intent. Second, users often struggle to articulate their intent clearly when they first submit a query, especially when they are unfamiliar with relevant technical terms (Xu et al. 2017; Rahman et al. 2018). Inspired by research on clarifying questions (Aliannejadi et al. 2021; Zamani et al. 2020; Xu et al. 2019; Aliannejadi et al. 2019; Zou and Kanoulas 2019; Ma et al. 2024; Zou et al. 2023) in related information retrieval domains, conversational retrieval with clarifying questions offers a natural solution to these issues. By iteratively introducing user-confirmed information by clarifying questions, this approach helps guide users toward expressing a more complete and accurate intent for related question retrieval.

Building upon the idea of clarifying questions, and aiming to address both the limitations of external information and the ambiguity of users' initial queries, we propose a novel **T**ag-**e**nhanced **C**onversational **Q**uestion **R**etrieval model, called **TeCQR**(Figure 1(c)), to retrieve related questions effectively through conversations. To begin with, we utilize a pre-trained language model to initialize separate embeddings for queries, questions, and tags. Then, we construct a *tag-enhanced two-stage offline training* framework to enhance the representations for more effective conversational retrieval, which consists of *query–question training via conversation simulation* and *tag–question training via contrastive learning*. Additionally, we introduce a *noise tolerance* model to mitigate the impact of noisy user feedback. Building on this, the *tag-enhanced conversational retrieval* phase selects an optimal sequence of tags to form clarifying questions and iteratively augments the query via user-confirmed tag information. After the conversations built upon clarifying questions, we utilize a soft-matching method to iteratively integrate user feedback to locate related questions effectively.

Our main contributions are summarized as follows:

- We formalize the paradigm of Conversational Question Retrieval (CQR) and propose a novel CQR model, named TeCQR, which enables users to clarify their intent through tag-enhanced conversation.
- We design a tag-enhanced two-stage offline training to enhance the representations of queries, questions, and tags, making them more effective for conversational retrieval.
- We introduce a noise tolerance modeling to effectively handle noisy user feedback.
- Extensive experimental results demonstrate that TeCQR significantly outperforms existing state-of-the-art baselines.

Related Work

Question retrieval is the task of retrieving related questions in response to a user-submitted query on cQA platforms. Early methods like BM25 (Robertson and Walker 1994) relied on exact term matching, lacking semantic understanding. Later work (Yang et al. 2013; Xu et al. 2017) introduced classical machine learning to incorporate semantic features.

Despite these advancements, classical machine learning models still struggled with capturing nuanced semantics, prompting a shift toward pre-trained embeddings and deep learning approaches for question retrieval (Shirani et al. 2019; Zhang, Sun, and Wang 2018; Kumari et al. 2021; Li, Du, and Chen 2020; Liu et al. 2023; Gao et al. 2023; Ha et al. 2021; Zhao and Huang 2022; Wang, Zhang, and Jiang 2019, 2020). BiLSTM-based models constructed semantic interaction modules to model the relationships between queries and candidate questions (Huang et al. 2023; Qin et al. 2022; Pei et al. 2021; Guo et al. 2022). Similarly, RCNN (Rücklé, Swarnkar, and Gurevych 2019; Lei et al. 2016) utilized gated convolutions to encode questions, distilling key information from potentially noisy text. HCA (Zahedi, Rahgozar, and Zoroofi 2020) further enriched representations by combining sentence and word-level cues. However, these deep models still suffer from short and ambiguous queries that lack sufficient contextual cues for accurate retrieval.

To alleviate this limitation, prior work leverage external information on cQA platforms. Query expansion was explored to handle short and ambiguous queries (Nie et al. 2016; Lu et al. 2015; Ghasemi and Shakery 2024), but it often introduced irrelevant or non-technical terms, increasing retrieval complexity. To address this, tags were widely applied to refine query representations. For example, Hazra et al. (2024) learned tag representations with node2vec on a tag co-occurrence graph, Liu et al. (2024) derived distributed tag embeddings by applying DeepWalk (Perozzi, Al-Rfou, and Skiena 2014) to a tag graph, and Li et al. (2024) embed questions by concatenating the tag sequence with the question text. However, these methods were developed for duplicate question retrieval, where questions in the database are already labeled with tags. This setting differs from the real-world scenario we focus on, where users submit tagless queries to cQA platforms. Additionally, Zhang et al. (2022) conducted an initial user study with human volunteers, demonstrating the potential of asking CQs to improve retrieval accuracy. Beyond tags, other studies have leveraged external signals, including knowledge graphs (Liu et al. 2023; Zou et al. 2026), the authority of answer providers (Zhao et al. 2017; Zhang et al. 2014), and answer content (Liang et al. 2019; Zhang and Wu 2018).

However, the above methods treat external information and questions as separate components, relying on static information enhancement and one-off query interaction. In contrast, we innovatively introduce a tag-enhanced conversational fusion mechanism that dynamically incorporates external information during the retrieval process. By guiding user interaction in an incremental manner, our approach enables a more accurate understanding of query intent and enhances retrieval performance.

Problem Formalization

Given an initial query Q submitted by a user, the goal of question retrieval is to identify a set of related ques-

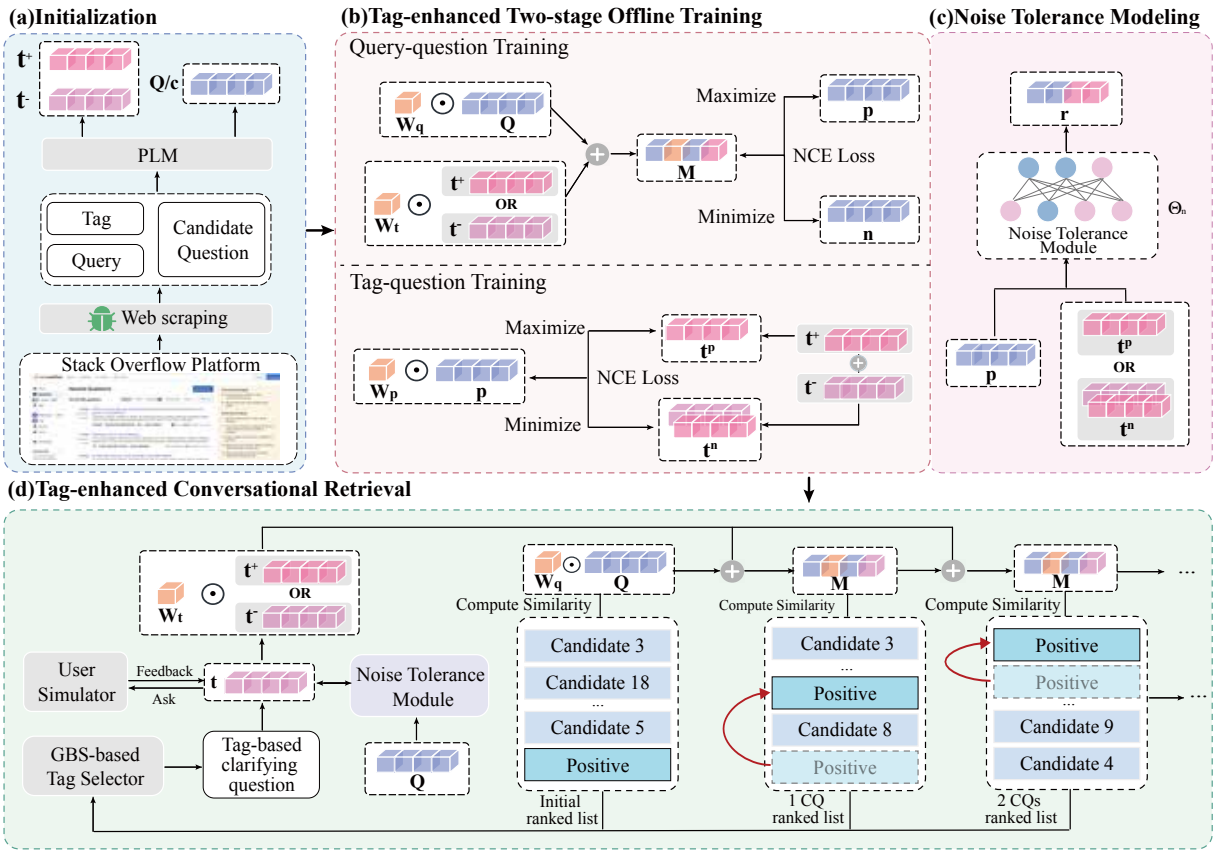


Figure 2: Overview of TeCQR.

tions $P = \{p_1, p_2, \ldots, p_k\}$ from a candidate pool $C = \{c_1, c_2, \ldots, c_n\}$, where $P \subseteq C$. The ground-truth set P is annotated by real users on the Stack Overflow platform, while the remaining candidates $N = C \setminus P = \{n_1, n_2, \ldots, n_j\}$ are treated as negative examples. The typical retrieval objective can be formally expressed as:

$$Q \rightarrow \{\text{ranked}(c_1, c_2, \ldots, c_n)\} \rightarrow P. \quad (1)$$

In our work, the retrieval process is enhanced by introducing clarifying questions to interact with the user. Specifically, the system asks a clarifying question q^t based on a specific tag t during the retrieval session. The user responds with feedback f, forming a clarifying question & feedback pair, which is denoted as (q^t, f). Each retrieval session contains multiple rounds of conversation. Thus, a sequence of clarifying questions is asked to the user, and a sequence of associated feedback is collected for these questions. Suppose l rounds of clarifying questions have been asked, the sequence of clarifying question & feedback pairs can be represented as: $\{(q_1^t, f_1), (q_2^t, f_2), \ldots, (q_l^t, f_l)\}$. The sequence of actions in conversational question retrieval can be represented as follows:

$$Q \rightarrow \{\text{ranked}(c_1, c_2, \ldots, c_n)\}, \{(q_i^t, f_i)\}_{i=1}^{l} \rightarrow P. \quad (2)$$

Methodology

In this section, we introduce our proposed TeCQR model. An overview of TeCQR is shown in Figure 2. The TeCQR model consists of four modules: (a) initialization using pre-trained language models; (b) tag-enhanced two-stage offline training, including query-question training via conversation simulation and tag-question training via contrastive learning to enhance these embeddings; (c) noise tolerance model, which effectively handle noisy user feedback; (d) tag-enhanced conversational retrieval, which iteratively refines user intent through tag-based clarifying questions to improve retrieval performance. The dataset construction process is described in *Section of Dataset*.

Initialization

To place query, candidate question, and tag in a unified semantic space, we share a single encoder, the pre-trained language model `All-MiniLM` (Wang et al. 2020), to encode them into representations $\mathbf{Q}$, $\mathbf{c}$, and $\mathbf{t}$, respectively, as illustrated in Figure 2(a). Since queries and candidate questions share the same format, both are encoded using All-MiniLM:

$$\mathbf{Q} = \mathbf{F}_{\text{All-MiniLM}}(i_Q), \quad \mathbf{c} = \mathbf{F}_{\text{All-MiniLM}}(i_c), \quad (3)$$

where i_Q and i_c are the corresponding textual inputs of query or candidate questions.

Modeling negative feedback is particularly challenging, as relevant items often share similar characteristics, while the reasons for irrelevance can vary significantly. To address this, we learn a separate embedding for negative feedback, which serves as a strong supervisory signal and plays a crucial role in guiding the retrieval process. Specifically, we encode the tag text i_t as:

$$\mathbf{t}^+ = \mathbf{F}_{\text{All-MiniLM}}(i_t), \quad \mathbf{t}^- = -\mathbf{t}^+, \tag{4}$$

where positive tag representation $\mathbf{t}^+$ corresponds to a tag that is semantically relevant to the query (i.e., receives positive feedback), while the negative tag representation $\mathbf{t}^-$ is constructed by negating $\mathbf{t}^+$, aiming to capture the semantics of negative feedback.

Tag-enhanced Two-stage Offline Training

Query-question Training via Conversation Simulation. In query–question training, we aim to capture the relationship between a query and its related questions. To achieve this, we first enhance the query representation by simulating the process of asking clarifying questions and receiving user feedback. Specifically, at each round of the conversation, we pose a tag-based clarifying question and receive user feedback on the corresponding tag. To incorporate this feedback into the query representation, we construct a mixture query representation $\mathbf{m}$ as follows:

$$\mathbf{m} = W_Q \cdot \mathbf{Q} + W_t \cdot \mathbf{t}^{\pm}, \tag{5}$$

where $\mathbf{Q}$ denotes the original query embedding, and $\mathbf{t}^{\pm}$ represents either the positive tag embedding $\mathbf{t}^+$ or the negative tag embedding $\mathbf{t}^-$, depending on the user feedback. In conversational retrieval, the query and tag contribute unequally to identifying the target question. We introduce learnable weights W_Q and W_t to adaptively balance their contributions by capturing their mutual relationship.

After enhancing the query representation with user feedback from clarifying questions, we use the refined query for retrieval. Specifically, we adopt a Noise Contrastive Estimation (NCE) loss (Gutmann and Hyvärinen 2010) as follows:

$$\begin{aligned}\mathcal{L}_{\text{QQ}} = -\frac{1}{|M|}\sum_{\mathbf{m}\in M}\Bigg[&\log\sigma(\mathbf{m}^\top\mathbf{p}) \\ &+ \frac{1}{|N|}\sum_{\mathbf{n}\in N}\log\Big(1-\sigma(\mathbf{m}^\top\mathbf{n})\Big)\Bigg],\end{aligned} \tag{6}$$

where $\mathbf{p}$ denotes the embedding of the positive (target) question, and $\mathbf{n} \in N$ denotes the representation of a negative question sample. During training, each query is paired with its corresponding positive question(s) as labeled by the Stack Overflow platform. If multiple positive questions exist for a given query, one is randomly selected in each training round. Negative samples are randomly drawn from the entire question pool. $\sigma(x)$ is the sigmoid function. $|M|$ and $|N|$ denote the total number of mixture queries and the number of negative samples per query, respectively.

Tag-question Training via Contrastive Learning. Since each tag captures a specific aspect of a question, a combination of relevant tags can serve as a rough semantic approximation of the question. Therefore, in the embedding space, a question representation should be closer to its associated tags and farther from unrelated ones. To enforce this, we introduce the tag-question training, which models the mutual relationship between questions and tags.

To adjust the relative contribution of questions and tags during alignment, we also employ an adaptive parameter on the question representation. Specifically, the adjusted question $\mathbf{p}'$ representation is computed as:

$$\mathbf{p}' = W_p \cdot \mathbf{p}, \tag{7}$$

where $\mathbf{p}$ denotes the original question embedding, and W_p is a learnable parameter that controls the contribution of the question in the alignment process. We further employ a NCE loss (Gutmann and Hyvärinen 2010) to optimize this alignment, defined as follows:

$$\begin{aligned}\mathcal{L}_{\text{TQ}} = -\frac{1}{|P'|}\sum_{\mathbf{p}'\in P'}\Bigg[&\log\sigma\left(\mathbf{p}'^\top\mathbf{t}^p\right) \\ &+ \frac{1}{|T^n|}\sum_{\mathbf{t}^n\in T^n}\log\Big(1-\sigma\left(\mathbf{p}'^\top\mathbf{t}^n\right)\Big)\Bigg],\end{aligned} \tag{8}$$

where $\mathbf{t}^p$ and $\mathbf{t}^n$ denote the representations of the positive and negative tags, respectively. The positive tag is labeled by the Stack Overflow platform, while negative tags are randomly sampled from the entire tag set. $|P'|$ and $|T^n|$ represent the total number of questions and the number of negative tags per question, respectively.

During the offline training, we use the Alternating Least Square (ALS) (Takács and Tikk 2012) technique to train the model, i.e., repeatedly optimize one of $\mathcal{L}_{\text{QQ}}$ and $\mathcal{L}_{\text{TQ}}$.

Noise Tolerance Modeling

Prior work (Zou et al. 2022; Zou and Kanoulas 2020; Zou, Li, and Kanoulas 2018; Zhang et al. 2018; Zou, Chen, and Kanoulas 2020) on clarifying questions make a strong assumption: users always know the correct answer to a clarifying question. In the context of cQA, this implies that users are assumed to be 100% certain about a query's relevance to a tag. However, in practice, users may provide noisy answers due to limited domain knowledge or simple operational errors. To address this issue, after the two-stage offline training steps, we introduce noise tolerance modeling to mitigate the impact of noisy feedback during the clarification process. This modeling aims to evaluate the semantic similarity between a question p and an associated tag t, using fixed representations obtained from earlier training. We optimize the model using binary cross-entropy loss, defined as follows:

$$\begin{aligned}\mathcal{L}_{\text{NR}} = -\frac{1}{|P|}\sum_{\mathbf{p}\in P}\Bigg[&y\cdot\log\left(\Theta_n([\mathbf{p};\mathbf{t}^p])\right) \\ &+ \frac{1}{|T^n|}\sum_{\mathbf{t}^n\in T^n}(1-y)\cdot\log\left(1-(\Theta_n([\mathbf{p};\mathbf{t}^n]))\right)\Bigg],\end{aligned} \tag{9}$$

where $\mathbf{p}$ denotes the question embedding, and $\mathbf{t}^p$ and $\mathbf{t}^n$ denote the representations of the positive and negative tags, respectively. Θ_n represents the parameters of the noise tolerance module, which consists of two linear layers and maps the concatenated representation $[\mathbf{p};\mathbf{t}]$ to a relevance score r. $|P|$ and $|T^n|$ denote the total number of questions and the number of negative tags per question, respectively.

Tag-enhanced Conversational Retrieval

Question Ranking with TeCQR. After tag-enhanced two-stage offline training, we obtained enhanced representations of queries, questions, and tags. These representations are then fixed and utilized to perform conversational question retrieval. In this section, we introduce how to retrieve questions based on the trained embeddings and user feedback.

After l rounds of conversation (where $l > 0$), we obtain l rounds of clarifying questions and the corresponding user feedback, denoted as $S_{(q^t,f)} = \{(q_1^t, f_1), (q_2^t, f_2), \ldots, (q_l^t, f_l)\}$. The probability of relevance for each candidate question in the l-th iteration is ranked according to:

$$P(\mathbf{c} \mid C, \mathbf{Q}, S_{(q^t,f)}) = \frac{\exp\left[\mathbf{c} \cdot \left(W_Q \cdot \mathbf{Q} + \sum_{(q^t,f)\in S_{(q^t,f)}} W_t \cdot \mathbf{e}_{(q^t,f)}\right)\right]}{\sum_{\mathbf{c}'\in C} \exp\left[\mathbf{c}' \cdot \left(W_Q \cdot \mathbf{Q} + \sum_{(q^t,f)\in S_{(q^t,f)}} W_t \cdot \mathbf{e}'_{(q^t,f)}\right)\right]}, \tag{10}$$

where $\mathbf{c}$ represents a candidate question from the candidate list C, and $\mathbf{Q}$ is the representation of initial query. W_Q and W_t are personalized weighting parameters. The term $\mathbf{e}_{(\mathbf{q^t},\mathbf{f})}$ represents the embedding of a clarifying question q^t with feedback f, which is defined as:

$$\mathbf{e}_{q^t f} = \begin{cases} \text{ask another}, & \Theta_n([\mathbf{Q}; \mathbf{t}^f]) \le \alpha \\ \mathbf{t}^f, & \Theta_n([\mathbf{Q}; \mathbf{t}^f]) > \alpha \end{cases} \quad \text{where } f \in \{+, -\}, \tag{11}$$

where $\mathbf{t}^+$ and $\mathbf{t}^-$ denote the positive and negative tag representations. To guide the decision-making process, we introduce a confidence threshold α. If the confidence score $\Theta_n([\mathbf{Q}; \mathbf{t}^f]) \le \alpha$, we regard such a signal as ambiguous and select a suboptimal tag for asking another clarifying question. If $\Theta_n([\mathbf{Q}; \mathbf{t}^f]) > \alpha$, we assume the model's prediction aligns with the user's intent, and the feedback ($\mathbf{t}^+$ or $\mathbf{t}^-$) is accepted and integrated to the representation $\mathbf{e}_{q^t f}$. The same mechanism applies to both positive and negative feedback.

It is worth noting that, there may be no conversation between the system and the user (e.g., when the first iteration (i.e., $l = 0$), the system has not yet asked any clarifying questions). In this case, $\mathbf{e}_{q^t f}$ is set to zero. In specific, our model can still retrieve a question q_u from the candidate question set C according to:

$$P(\mathbf{c} \mid C, \mathbf{Q}) = \frac{\exp(\mathbf{c} \cdot \mathbf{Q})}{\sum_{\mathbf{c}'\in C} \exp(\mathbf{c}' \cdot \mathbf{Q})}. \tag{12}$$

Learning to Ask. In previous sections, we described how to utilize queries and clarifying questions with user feedback to retrieve related questions. Here, we explain how to select the best-suited tag to generate a clarifying question to ask the user. The goal of this selection process is to identify high-reward tags that can maximize the information gain and minimize the number of question times. Inspired by Zou and Kanoulas (2020), and Zou et al. (2022), we employ the Generalized Binary Search (GBS) (Nowak 2008) strategy as our tag selection strategy. GBS is a greedy approach that learns to select a sequence of best-suited tags for clarifying questions. The selection process is formally defined as:

$$t_l = \arg\min_t \left| \sum_{c\in C} \left(2\mathbb{1}\{t^c = 1\} - 1\right) \pi_l(c) \right|, \tag{13}$$

where t_l is the tag selected in the l-th round, and t^c indicates whether the question c is related to the tag t. If the question c is related to the tag t, then $\mathbb{1}\{t^c = 1\} = 1$; otherwise, if c is unrelated to t, then $\mathbb{1}\{t^c = 1\} = 0$. $\pi_l(c)$ represents a contribution score, defined as:

$$\pi_l(c) = \frac{1}{\text{index}_c + 1}, \tag{14}$$

where index_c is the rank of question c in the candidate questions list during the l-th round. Once a specific tag is selected as a clarifying question, the user provides feedback based on this question. The feedback is given as "yes" or "no". Details of the user simulator are provided in the *User Simulator*.

Experiments

Dataset. Due to the lack of conversational question retrieval datasets, we construct a new dataset named *StackOverflow-Tag*, to advance research in conversational question retrieval. We begin by collecting queries from related questions on the SO platform. For each query, we employ the BM25 algorithm (Robertson and Walker 1994) to retrieve 20 candidate questions from the corpus. We then use the official SO API to collect associated tags for each question, and discard any questions without tags. The resulting *StackOverflow-Tag* dataset contains 54,786 questions and 2,278 unique tags. We divide the data into a training set with 13,772 queries and a test set with 1,482 queries, providing a realistic dataset for evaluating conversational question retrieval methods.

Evaluation Metrics. We use Recall@k (k=1, 3, 5), and NDCG@k (k=3, 5, 10), Mean Average Precision, Mean Reciprocal Rank as metrics to evaluate the retrieval performance, following prior work such as (Rücklé, Swarnkar, and Gurevych 2019; Qin et al. 2022; Hazra et al. 2024).

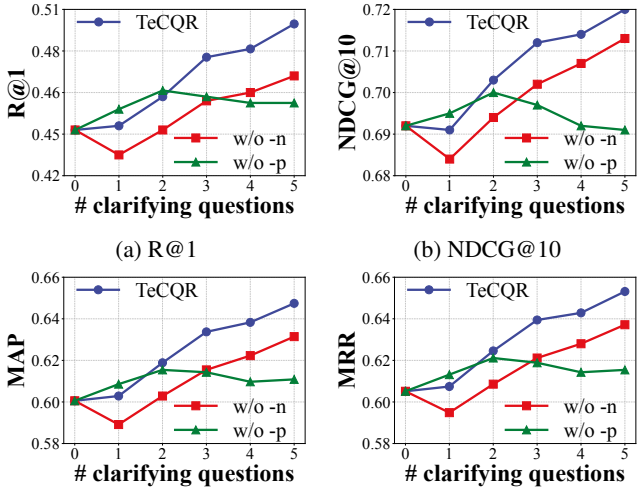
Implementation Details. All experiments are conducted on an NVIDIA RTX A6000 GPU. We use All-MiniLM as the encoder backbone and train with SGD (initial learning rate 0.1, decayed to 0). The embedding size is set to 384, matching the pre-trained model. The threshold α for the noise tolerance module is set to 0.5. Baseline results are reported using their optimal hyperparameters. Additional parameter sensitivity studies on the number of clarifying questions, batch size, and negative samples, as well as detailed software and hardware specifications, are provided in the *Appendix*.

Comparison Methods. We compare TeCQR with several baselines, including **BM25** (Robertson and Walker 1994), **GRU** (Cho et al. 2014), **RCNN** (Rücklé, Swarnkar, and Gurevych 2019), **RMSO** (Qin et al. 2022), **Sentence-BERT** (Reimers and Gurevych 2019; Ha et al. 2021), **TS-QR** (Liu et al. 2024), **All-MiniLM** (Wang et al. 2020), **Query-Tag**, and **TEcotag** (Hazra et al. 2024). Detailed descriptions of all compared methods are available in the *Appendix*.

User Simulator. Following prior work (Zou et al. 2022; Zou and Kanoulas 2020; Zou, Li, and Kanoulas 2018; Zhang et al. 2018), we assume users can judge whether a query is related to a given tag. Based on this assumption, we design it as follows: if the tag in the clarifying question is related to the questions, the simulator responds with "yes"; otherwise, it responds with "no". To better reflect real-world noise, in section *Effects of Noise Tolerance Modeling*, we conducted

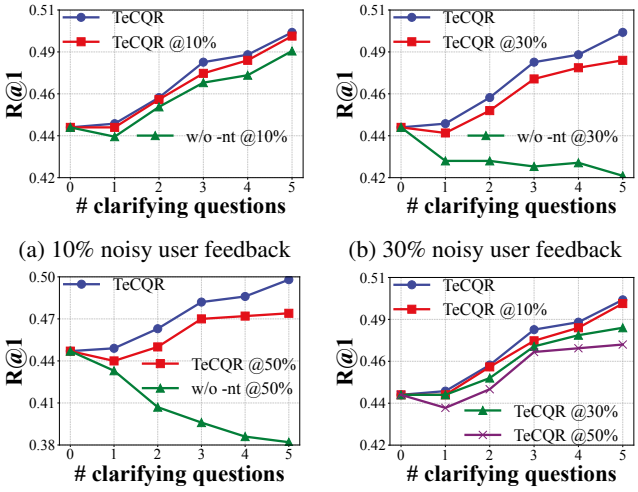
Method	R@1	R@3	R@5	NDCG@3	NDCG@5	NDCG@10	MAP	MRR
BM25	0.258	0.318	0.387	0.280	0.308	0.352	0.321	0.324
GRU	0.365	0.435	0.504	0.401	0.430	0.464	0.440	0.441
RCNN	0.386	0.465	0.534	0.429	0.457	0.490	0.466	0.468
RMSO	0.412	0.502	0.565	0.462	0.497	0.522	0.505	0.507
Sentence-BERT	0.228	0.405	0.521	0.332	0.379	0.443	0.378	0.380
TS-QR	0.382	0.463	0.538	0.422	0.455	0.494	0.466	0.469
All-MiniLM	0.443	0.673	0.798	0.579	0.631	0.687	0.599	0.604
Query-Tag	0.432	0.661	0.789	0.563	0.617	0.670	0.584	0.588
TEcotag	0.448	0.675	0.801	0.563	0.628	0.690	0.605	0.608
TeCQR (0 CQ)	0.447	0.671	0.792	0.580	0.632	0.692	0.603	0.607
TeCQR$_{\text{random}}$ (5 CQs)	0.446	0.675	0.793	0.580	0.629	0.687	0.600	0.604
TeCQR (5 CQs)	**0.498***	**0.717***	**0.833***	**0.629***	**0.677***	**0.720***	**0.644***	**0.649***

Table 1: TeCQR Performance of Question Retrieval on the *StackOverflow-Tag* dataset. The symbol "*" refers to a significant improvement compared to the TEcotag baseline at the $p < 0.05$ level using the two-tailed pairwise t-test.



(a) R@1 (b) NDCG@10 (c) MAP (d) MRR

Figure 3: Effects of positive and negative user feedback.



(a) 10% noisy user feedback (b) 30% noisy user feedback (c) 50% noisy user feedback (d) various noisy user feedback

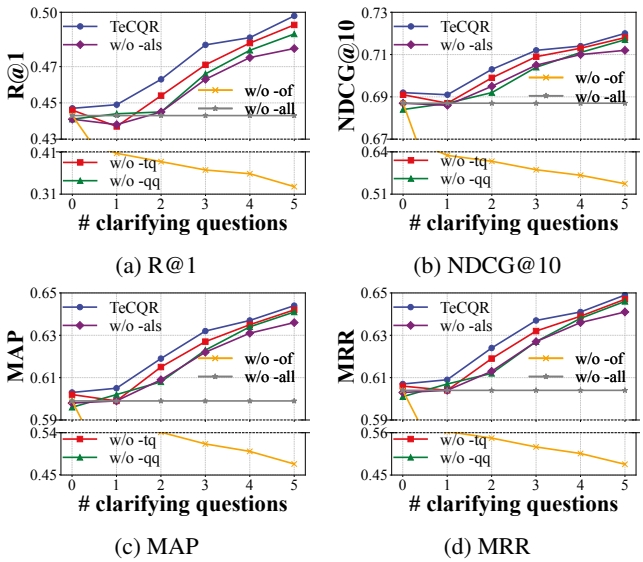
Figure 4: Effects of noisy user feedback.

experiments under the assumption that the user simulator has a certain probability of making errors, and we provide a detailed analysis of the impact of noisy user feedback and our noise tolerance.

Overall Performance Comparison

As shown in Table 1, BM25 performs the worst due to its lack of semantic modeling. GRU and RCNN improve results through deeper representations, while RMSO further enhances performance by fusing query and candidate embeddings. Although Sentence-BERT benefits from pre-training, it yields limited gains and even underperforms BM25 on R@1. TS-QR improves results by incorporating tag information via DeepWalk. All-MiniLM, trained on large-scale semantic matching data, achieves the best baseline results on the *StackOverflow-Tag* dataset. Naive tag integration methods, such as Query-Tag (concatenation) and TEcotag (node2vec), show limited or even negative gains, highlighting the challenge of leveraging tag information.

Our TeCQR with 5 CQs significantly outperforms all baselines, achieving MAP and MRR improvements of 6.4% and 6.7% over TEcotag. R@1 increases by 10%, NDCG@10 by 4%, with further gains in R@3 (+6.2%), R@5 (+3.9%), NDCG@3 (+11.7%), and NDCG@5 (+7.8%), demonstrating strong top-k performance. Even without clarification, TeCQR (0 CQs) shows slight gains due to tag-enhanced two-stage offline training. Building on this, our core contribution lies in introducing conversation into question retrieval, where tag-enhanced two-stage offline training helps the system better handle short and ambiguous queries. A variant with random tag selection TeCQR$_{\text{random}}$ shows degraded performance as CQs increase, highlighting the importance of effective tag selection via GBS.



(a) R@1 (b) NDCG@10

(c) MAP (d) MRR

Figure 5: Ablation Study.

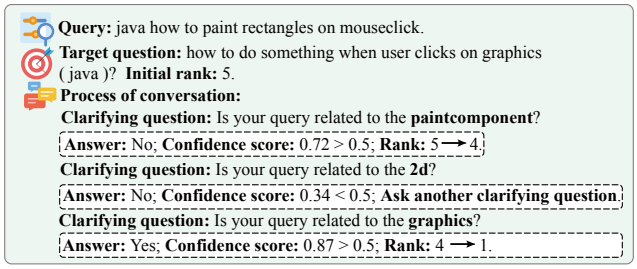


Figure 6: Case Study.

Effects of Positive and Negative User Feedback

After receiving tag-based CQs, users give positive or negative feedback to indicate tag relevance. We evaluate two variants: (1) **TeCQR w/o -n**, which incorporates tags with positive feedback; and (2) **TeCQR w/o -p**, which incorporates tags with negative feedback. As shown in Figure 3, TeCQR w/o -n shows an initial drop but steady long-term gains, while TeCQR w/o -p yields slight early improvement with diminishing returns. By combining both feedback and adaptively weighting tag contributions, TeCQR achieves the best overall performance, demonstrating the effectiveness of modeling both positive and negative feedback.

Effects of Noise Tolerance Modeling

Following prior work (Zou et al. 2022; Zou and Kanoulas 2020; Zou, Li, and Kanoulas 2018; Zhang et al. 2018), we assume users can judge query–tag relevance, but real users may provide noisy feedback. To address it, we introduce noisy feedback with error rates from 10% to 50%, in steps of 20%. We evaluate both the TeCQR and a variant without the noise tolerance module **TeCQR w/o -nt** under different noise condition. Results are shown in Figure 4, where k in **TeCQR @k%** and **TeCQR w/o -nt @k%** represents the error rate. As shown in Figure 4, performance degrades as noise increases, which is expected. Nevertheless, TeCQR consistently benefits from CQs across all noise levels, while removing noise tolerance leads to significant drops. When noise exceeds 30%, CQs even harm performance without noise tolerance, demonstrating both the vulnerability to noisy feedback and the effectiveness of our noise tolerance module.

Contribution of Key Components

We conduct an ablation study to assess the contribution of each component in TeCQR (Figure 5). We compare: (1) **TeCQR w/o -tq**, removing tag-question training; (2) **TeCQR w/o -qq**, removing query-question training; (3) **TeCQR w/o -of**, directly applying All-MiniLM in the conversational retrieval phase without our tag-enhanced two-stage offline training; (4) **TeCQR w/o -als**, replacing ALS strategy with a standard joint loss; and (5) **TeCQR w/o -all**, removing all proposed modules and using All-MiniLM for static retrieval. All components contribute to performance. TeCQR w/o -of performs worst, confirming the value of two-stage offline training; removing either training objective (w/o -tq or w/o -qq) degrades performance, demonstrating their complementary benefits; TeCQR w/o -als is competitive but underperforms the ALS strategy; and TeCQR w/o -all remains a strong static baseline. Overall, TeCQR achieves the best results, validating our overall design.

Case Study

We present a case from the *StackOverflow-Tag* dataset to illustrate TeCQR's conversational retrieval process. Initially, the target question was ranked 5. After the first clarifying question and a negative user response, its rank improved to 4. In the next round, the model's prediction from the noise tolerance model deviated from the user feedback. As a result, another clarifying question was asked, which received positive feedback. Ultimately, the rank of the target question improved to 1, demonstrating TeCQR's effectiveness in refining retrieval through conversation.

Conclusion

In this paper, we propose TeCQR, a novel model for conversational question retrieval in cQA platforms. It tackles the challenges of short, ambiguous queries and unreliable external information by introducing a novel multi-turn conversation into the question retrieval, allowing users to gradually clarify their intent. The TeCQR model mainly consists of three core modules: (1) a tag-enhanced two-stage offline training framework designed to improve the representations for more effective conversational retrieval, including query–question training via conversation simulation and tag–question training via contrastive learning; (2) a noise tolerance module that mitigates the impact of noisy user feedback; and (3) a tag-enhanced conversational retrieval module, which selects optimal tags to generate CQs and iteratively refines the query representation based on user feedback, enabling accurate related question retrieval.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (62402093), the Sichuan Science and Technology Program (2025ZNSFSC0479), and the Fundamental Research Funds for the Central Universities (JBK202511020). This work was also supported in part by the National Natural Science Foundation of China under grants U20B2063 and 62220106008, and the Sichuan Science and Technology Program under Grant 2024NSFTD0034.

References

Aliannejadi, M.; Kiseleva, J.; Chuklin, A.; Dalton, J.; and Burtsev, M. 2021. Building and Evaluating Open-Domain Dialogue Corpora with Clarifying Questions. In *EMNLP 2021-2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, 4473–4484.

Aliannejadi, M.; Zamani, H.; Crestani, F.; and Croft, W. B. 2019. Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, 475–484.

Chen, C.; Chen, X.; Sun, J.; Xing, Z.; and Li, G. 2018. Data-Driven Proactive Policy Assurance of Post Quality in Community qa Sites. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW).

Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Gao, Z.; Xia, X.; Lo, D.; Grundy, J.; Zhang, X.; and Xing, Z. 2023. I know what you are searching for: Code snippet recommendation from stack overflow posts. *ACM Transactions on Software Engineering and Methodology*, 32(3): 1–42.

Ghasemi, S.; and Shakery, A. 2024. Harnessing the power of metadata for enhanced question retrieval in community question answering. *IEEE Access*, 12: 65768–65779.

Guo, A.; Li, X.; Pang, N.; and Zhao, X. 2022. Adversarial Cross-domain Community Question Retrieval. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 21(3).

Gutmann, M.; and Hyvärinen, A. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 297–304.

Ha, T.-T.; Nguyen, V.-N.; Nguyen, K.-H.; Nguyen, K.-A.; and Than, Q.-K. 2021. Utilizing SBERT For Finding Similar Questions in Community Question Answering. In *2021 13th International Conference on Knowledge and Systems Engineering (KSE)*, 1–6.

Hazra, R.; Saha, D.; Sahoo, A.; Banerjee, S.; and Mukherjee, A. 2024. Duplicate Question Retrieval and Confirmation Time Prediction in Software Communities. In *Proceedings of the 2023 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 203–212.

Huang, S.; Wu, Y.; Lu, J.; and Deng, C. 2023. Related Questions Detection Model in Stack Overflow based on Semantic Matching. In *SEKE*, 339–344.

Kumari, R.; Mishra, R.; Malviya, S.; and Tiwary, U. S. 2021. Detection of semantically equivalent question pairs. In *Intelligent Human Computer Interaction: 12th International Conference, IHCI 2020, Daegu, South Korea, November 24–26, 2020, Proceedings, Part I 12*, 12–23.

Lei, T.; Joshi, H.; Barzilay, R.; Jaakkola, T.; Tymoshenko, K.; Moschitti, A.; and Màrquez, L. 2016. Semi-supervised Question Retrieval with Gated Convolutions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1279–1289.

Li, B.; Du, X.; and Chen, M. 2020. Cross-language question retrieval with multi-layer representation and layer-wise adversary. *Information Sciences*, 527: 241–252.

Li, H.; Li, J.; Jin, H.; Chen, Z.; and Zou, W. 2024. Combining Multi-granularity Text Semantics with Graph Relational Semantics for Question Retrieval in CQA. In *International Conference on Intelligent Computing*, 53–64.

Liang, D.; Zhang, F.; Zhang, W.; Zhang, Q.; Fu, J.; Peng, M.; Gui, T.; and Huang, X. 2019. Adaptive Multi-Attention Network Incorporating Answer Information for Duplicate Question Detection. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 95–104.

Liu, M.; Yu, S.; Peng, X.; Du, X.; Yang, T.; Xu, H.; and Zhang, G. 2023. Knowledge graph based explainable question retrieval for programming tasks. In *2023 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, 123–135.

Liu, Y.; Tang, W.; Liu, Z.; Tang, A.; and Zhang, L. 2024. Similar question retrieval with incorporation of multi-dimensional quality analysis for community question answering. *Neural Computing and Applications*, 36: 3663–3679.

Lu, M.; Sun, X.; Wang, S.; Lo, D.; and Duan, Y. 2015. Query expansion via wordnet for effective code search. In *2015 IEEE 22nd International Conference on Software Analysis, Evolution, and Reengineering (SANER)*, 545–549.

Ma, H.; Zou, J.; Aliannejadi, M.; Kanoulas, E.; Bin, Y.; and Yang, Y. 2024. Ask or Recommend: An Empirical Study on Conversational Product Search. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 3927–3931.

Nie, L.; Jiang, H.; Ren, Z.; Sun, Z.; and Li, X. 2016. Query expansion based on crowd knowledge for code search. *IEEE Transactions on Services Computing*, 9(5): 771–783.

Nowak, R. 2008. Generalized binary search. In *2008 46th annual Allerton conference on communication, control, and computing*, 568–574.

Pei, J.; Wu, Y.; Qin, Z.; Cong, Y.; and Guan, J. 2021. Attention-based model for predicting question relatedness on Stack Overflow. In *2021 IEEE/ACM 18th International Conference on Mining Software Repositories (MSR)*, 97–107.

Perozzi, B.; Al-Rfou, R.; and Skiena, S. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 701–710.

Qin, Z.; Wu, Y.; Pei, J.; Lu, J.; Huang, S.; and Liu, L. 2022. Related Questions Retrieval Model in Stack Overflow based on Semantic Matching. In *2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC)*, 321–326.

Rahman, M. M.; Barson, J.; Paul, S.; Kayani, J.; Lois, F. A.; Quezada, S. F.; Parnin, C.; Stolee, K. T.; and Ray, B. 2018. Evaluating how developers use general-purpose web-search for code retrieval. In *Proceedings of the 15th International Conference on Mining Software Repositories*, 465–475.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language*

Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 3982–3992.

Robertson, S. E.; and Walker, S. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR'94: Proceedings of the 37th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*, 232–241.

Rücklé, A.; Swarnkar, K.; and Gurevych, I. 2019. Improved Cross-Lingual Question Retrieval for Community Question Answering. In *The World Wide Web Conference*, 3179–3186.

Shirani, A.; Xu, B.; Lo, D.; Solorio, T.; and Alipour, A. 2019. Question relatedness on stack overflow: the task, dataset, and corpus-inspired models. *arXiv preprint arXiv:1905.01966*.

Takács, G.; and Tikk, D. 2012. Alternating least squares for personalized ranking. In *Proceedings of the sixth ACM conference on Recommender systems*, 83–90.

Wang, B.; Yang, Y.; Xu, X.; Hanjalic, A.; and Shen, H. T. 2017. Adversarial Cross-Modal Retrieval. In *Proceedings of the 25th ACM International Conference on Multimedia*, 154–162.

Wang, L.; Zhang, L.; and Jiang, J. 2019. Detecting Duplicate Questions in Stack Overflow via Deep Learning Approaches. In *2019 26th Asia-Pacific Software Engineering Conference (APSEC)*, 506–513.

Wang, L.; Zhang, L.; and Jiang, J. 2020. Duplicate Question Detection With Deep Learning in Stack Overflow. *IEEE Access*, 8: 25964–25975.

Wang, W.; Wei, F.; Dong, L.; Bao, H.; Yang, N.; and Zhou, M. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. volume 33, 5776–5788.

Wang, Z.; Gao, Z.; Yang, Y.; Wang, G.; Jiao, C.; and Shen, H. T. 2025. Geometric Matching for Cross-Modal Retrieval. *IEEE Transactions on Neural Networks and Learning Systems*, 36: 5509–5521.

Xu, B.; Xing, Z.; Xia, X.; and Lo, D. 2017. AnswerBot: automated generation of answer summary to developersź technical questions. In *Proceedings of the 32nd IEEE/ACM International Conference on Automated Software Engineering*, 706–716.

Xu, J.; Wang, Y.; Tang, D.; Duan, N.; Yang, P.; Zeng, Q.; Zhou, M.; and Sun, X. 2019. Asking clarification questions in knowledge-based question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 1618–1629.

Yang, L.; Qiu, M.; Gottipati, S.; Zhu, F.; Jiang, J.; Sun, H.; and Chen, Z. 2013. CQArank: jointly model topics and expertise in community question answering. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, 99–108.

Ye, D.; Xing, Z.; and Kapre, N. 2017. The structure and dynamics of knowledge network in domain-specific QA sites: a case study of stack overflow. *Empirical Softw. Engg.*, 22(1): 375–406.

Zahedi, M. S.; Rahgozar, M.; and Zoroofi, R. A. 2020. HCA: Hierarchical compare aggregate model for question retrieval in community question answering. *Information Processing & Management*, 57(6): 102318.

Zamani, H.; Dumais, S.; Craswell, N.; Bennett, P.; and Lueck, G. 2020. Generating clarifying questions for information retrieval. In *Proceedings of the web conference*, 418–428.

Zhang, K.; Wu, W.; Wu, H.; Li, Z.; and Zhou, M. 2014. Question Retrieval with High Quality Answers in Community Question Answering. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, 371–380.

Zhang, M.; and Wu, Y. 2018. An unsupervised model with attention autoencoders for question retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 4978–4986.

Zhang, N.; Huang, Q.; Xia, X.; Zou, Y.; Lo, D.; and Xing, Z. 2022. Chatbot4QR: Interactive Query Refinement for Technical Question Retrieval. *IEEE Transactions on Software Engineering*, 48(4): 1185–1211.

Zhang, X.; Sun, X.; and Wang, H. 2018. Duplicate question identification by integrating framenet with neural networks. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*.

Zhang, Y.; Chen, X.; Ai, Q.; Yang, L.; and Croft, W. B. 2018. Towards Conversational Search and Recommendation: System Ask, User Respond. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 177–186.

Zhao, X.; and Huang, J. X. 2022. Bert-QAnet: BERT-encoded hierarchical question-answer cross-attention network for duplicate question detection. *Neurocomputing*, 509: 68–74.

Zhao, Z.; Lu, H.; Zheng, V. W.; Cai, D.; He, X.; and Zhuang, Y. 2017. Community-based question answering via asymmetric multi-faceted ranking network learning. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, 3532–3538.

Zou, J.; Chen, Y.; and Kanoulas, E. 2020. Towards Question-based Recommender Systems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 881–890.

Zou, J.; Huang, J.; Ren, Z.; and Kanoulas, E. 2022. Learning to ask: Conversational product search via representation learning. *ACM Transactions on Information Systems*, 41(2): 1–27.

Zou, J.; and Kanoulas, E. 2019. Learning to Ask: Question-based Sequential Bayesian Product Search. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 369–378.

Zou, J.; and Kanoulas, E. 2020. Towards question-based high-recall information retrieval: Locating the last few relevant documents for technology-assisted reviews. *ACM Transactions on Information Systems (TOIS)*, 38(3): 1–35.

Zou, J.; Li, D.; and Kanoulas, E. 2018. Technology Assisted Reviews: Finding the Last Few Relevant Documents by Asking Yes/No Questions to Reviewers. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 949–952.

Zou, J.; Lin, C.; Guo, W.; Wang, Z.; Wei, J.; Yang, Y.; and Shen, H. T. 2026. Multi-type context-aware conversational recommender systems via mixture-of-experts. *Information Fusion*, 126: 103638.

Zou, J.; Sun, A.; Long, C.; Aliannejadi, M.; and Kanoulas, E. 2023. Asking Clarifying Questions: To benefit or to disturb users in Web search? *Inf. Process. Manage.*, 60.