Question Tagging via Graph-guided Ranking*

XIAO ZHANG, Shandong University, China
MENG LIU, Shandong Jianzhu University, China
JIANHUA YIN, Shandong University, China
ZHAOCHUN REN, Shandong University, China
LIQIANG NIE, Shandong University, China

With the increasing prevalence of portable devices and the popularity of community Question Answering (cQA) sites, users can seamlessly post and answer many questions. To effectively organize the information for precise recommendation and easy searching, these platforms require users to select topics for their raised questions. However, due to the limited experience, certain users fail to select appropriate topics for their questions. Thereby, automatic question tagging becomes an urgent and vital problem for the cQA sites, yet it is non-trivial due to the following challenges. On the one hand, vast and meaningful topics are available yet not utilized in the cQA sites, how to model and tag them to relevant questions is a highly challenging problem. On the other hand, related topics in the cQA sites may be organized into a directed acyclic graph. In light of this, how to exploit relations among topics to enhance their representations is critical. To settle these challenges, we devise a graph-guided topic ranking model to tag questions in the cQA sites appropriately. In particular, we first design a topic information fusion module to learn the topic representation by jointly considering the name and description of the topic. Afterwards, regarding the special structure of topics, we propose an information propagation module to enhance the topic representation. As the comprehension of questions plays a vital role in question tagging, we design a multi-level context modeling based question encoder to obtain the enhanced question representation. Moreover, we introduce an interaction module to extract topic-aware question information, and capture the interactive information between questions and topics. Finally, we utilize the interactive information to estimate the ranking scores for topics. Extensive experiments on three Chinese cQA datasets have demonstrated that our proposed model outperforms several state-of-the-art competitors.

CCS Concepts: • **Information systems** → **Specialized information retrieval**; **Structured text search**; **Question answering**.

Additional Key Words and Phrases: Graph-guided Topic Ranking, Community Question Answering, Question Tagging

ACM Reference Format:
Xiao Zhang, Meng Liu, Jianhua Yin, Zhaochun Ren, and Liqiang Nie. 2020. Question Tagging via Graph-guided Ranking. *ACM Transactions on Information Systems* 1, 1, Article 111 (August 2020), 23 pages. https://doi.org/10.1145/xxxxxxx.xxxxxxx

*Both Meng Liu (mengliu.sdu@gmail.com) and Jianhua Yin (jhyin@sdu.edu.cn) are the corresponding authors.

Authors' addresses: Xiao Zhang, xiao.zhang@mail.sdu.edu.cn, Shandong University, Qingdao, Shandong, China; Meng Liu, mengliu.sdu@gmail.com, Shandong Jianzhu University, Jinan, Shandong, China; Jianhua Yin, jhyin@sdu.edu.cn, Shandong University, Qingdao, Shandong, China; Zhaochun Ren, Zhaochun.ren@sdu.edu.cn, Shandong University, Qingdao, Shandong, China; Liqiang Nie, nieliqiang@gmail.com, Shandong University, Qingdao, Shandong, China.

1046-8188/2020/8-ART111 $15.00
https://doi.org/10.1145/xxxxxxx.xxxxxxx

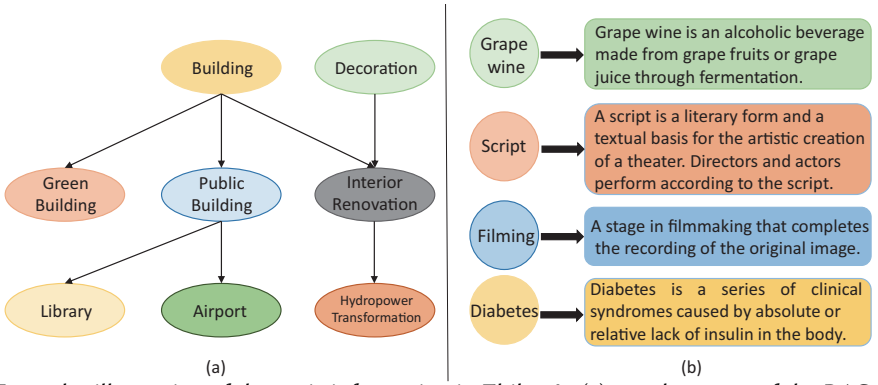

Fig. 1. Exemplar illustration of the topic information in Zhihu. In (a), we show part of the DAG-structure; And in (b), we list the descriptions of some topics.

1 INTRODUCTION

With the development of web 2.0, community Question Answering (cQA) sites, such as Quora[1] and Zhihu[2], have become more and more prevalent. Numerous questions and answers are uploaded daily by users. Nevertheless, with the booming data, it becomes increasingly difficult for users to locate their desired information in the cQA sites. Fortunately, questions are commonly tagged with at least one topic in the cQA sites, which largely benefits several topic-based tasks, such as question organizing, searching, and browsing. Besides, platforms can recommend questions to users based on the topics they follow. However, due to the lack of experience, users sometimes fail to appropriately tag their questions. In light of this, designing an intelligent topic ranking model to help users tag questions is of great practical importance, especially in empowering the user experience and boosting the efficiency of content distribution.

Building a topic ranking system to aid question tagging for the cQA sites is non-trivial, due to the following reasons: 1) Considering Zhihu as an example, according to the editorial guideline of topics in this site, all topics are organized into a Directed Acyclic Graph (DAG). As shown in Fig. 1(a), topics are linked by directed edges, namely topics in the DAG are not independent. To be more specific, for each topic (e.g., "Public Building"), topics from the higher layer linked to it are parent topics (e.g., "Building"), otherwise the children topics (e.g., "Airport"). Therefore, we need to consider the inherent topic relations defined by a given structure when learning their representations. 2) As any senior user can create a new topic at any time, there are numerous topics where plentiful topics are barely tagged to any question. Whereas, these topics are meaningful and should tag questions like the previous topics, which are trained and well represented. Heretofore, how to tag questions with these topics is largely untapped. And 3) the uploaded questions of cQA sites are mostly complex with one complete sentence or several sentences. Thereby, the key for topic ranking is to well comprehend the complex question information and capture the relations between questions and topics.

To tackle the topic ranking issue, a straightforward approach is to treat it as a text classification problem, i.e., regarding each topic as a class label. Over the past few years, a considerable amount of work has been dedicated to addressing the issue of text classification [13, 27, 43, 44]. Among these approaches, a common strategy is to first extract the text representation, and then predict the probability of all classes through a simple function mapping. Although they have achieved

[1] www.quora.com.
[2] www.zhihu.com.

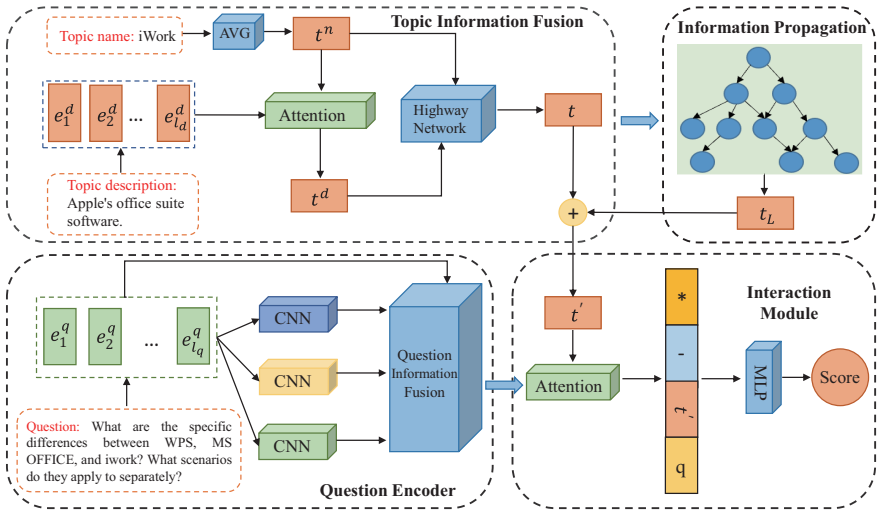

Fig. 2. The pipeline of our proposed HERE model. Firstly, the topic information fusion module generates the topic representation $\mathbf{t}$ by jointly considering the topic name and description. $\mathbf{t}_L$ is obtained through the information propagation between topics. The question encoder module utilizes multiple different convolutional neural networks and position encoding to obtain question representations. And then the interaction module builds the relationship between the question and the topic. Finally, it employs MLP for the interactive information to predict the matching score.

promising performance, they cannot be directly applied to the cQA sites. The main reason is that they require intensive manual labeling, i.e., thousands of instances for each topic, while new topics are constantly appearing all the time in the cQA sites. In other words, there are amount of topics without training data, i.e., unseen topics. More importantly, they ignore the inherent DAG-structure of topics when learning their representations. Recently, some zero-shot text classification methods have been developed to tackle the new class issue, and made some progress [24, 26]. Particularly, Pushp et al. [24] presented three simple models to capture the relations between texts and classes, while considering all classes as independent, which is not applicable in our case. Unlike the above approach, Rios et al. [26] introduced the Graph Convolutional Neural Network (GCNN) [7, 10, 15] to propagate information among topics. Thereby, the representations of new topics could be enhanced via other topics. Concretely, representations of new topics are obtained by averaging representations of their associated topics, namely associated topics of the new topic are treated equally. But these associated topics may contribute differently to the current topic in the cQA sites. As shown in Fig. 1 (a), topics like "Building" and "Decoration" are linked to "Interior Renovation". The former is the relevant place where the child topic "Interior Renovation" takes place, while the latter is more semantically related to "Interior Renovation". Although parent topics can enrich child topic information, they play the different role in new topic modeling. Therefore, simply averaging their information may obscure useful clues within crucial topics. In addition, most existing methods [24] encode question information via averaging word embeddings or adopting LSTM [6]. They may not be able to accurately capture meaningful semantic information from complex questions, further deteriorating the accuracy of topic tagging.

To better address the challenges mentioned above, in this paper, we present a grapH-guidEd Ranking modEl (HERE), as shown in Fig. 2. More importantly, it is capable of tagging questions with unseen topics. Since topic descriptions contain external knowledge to supplement the semantic

of topics, as illustrated in Fig. 1(b), we integrate this information into our model. Concretely, we first design a topic information fusion module to extract the critical information from the topic description, and then combine it with the topic name to represent each topic. Afterwards, to strengthen the representation of each topic, especially the unseen ones, we establish a DAG-based information propagation module to transfer information from the connected parent topics to the current ones. Meanwhile, we introduce a question encoder to enhance the comprehension of questions, utilizing multiple-scale convolutional networks to capture multi-level contextual information. Subsequently, we present an interaction module to capture the interactive information between the enhanced topic representation and the question representation. Finally, we adopt a Multi-Layer Perception (MLP) network to process the interactive information for predicting ranking scores.

The main contributions of this work are three-fold:

- We present a novel topic ranking model, named HERE, to address the question tagging problem in the cQA sites. It can obtain multi-level context-aware question representations and tag questions with unseen topics.
- To better represent topics, especially the unseen ones, we design a DAG-based information propagation module. It utilizes the multi-dimensional attention mechanism to extract meaningful information from parent topics. Moreover, we utilize the topic description to strengthen topic representations.
- To validate our model, we construct two datasets based on Zhihu, which is a popular Chinese cQA site. Extensive experiments on two datasets constructed by ourselves and one public dataset have verified the superiority of our proposed model. As a side contribution, we have released the codes and datasets to facilitate other researchers[3].

The rest of the paper is organized as follows. Section 2 reviews the related work. Section 3 details the question tagging problem and our proposed HERE model. We present the experimental results in Section 4, followed by the conclusion and future work in Section 5.

2 RELATED WORK

As far as we know, there is no work about utilizing the novel topic to tag questions in the cQA sites. Therefore, in this section, we mainly introduce the work related to our method. Particularly, we first present the work about text classification from the traditional setting and zero-shot learning, respectively. Hereafter, we briefly review the development of the attention mechanism.

2.1 Text Classification

2.1.1 Traditional Setting. To tackle the issue of question tagging, a simple and direct approach is to treat it as the text classification task, whereby each topic is a class label. In recent years, great efforts have been made on text classification, especially a few deep learning based approaches have been proposed. They commonly utilize the deep neural network to learn the text representation, and then a fully connected layer with the softmax function is adopted to predict the labels. According to the strategies of text feature extraction, they can be divided into two categories: the Recurrent Neural Network (RNN) based models and the Convolutional Neural Network (CNN) based ones.

The RNN-based models consider the text as sequence information, and utilize RNN to model the text representation. Specifically, Liu et al. [20] proposed three RNN-based architectures with parameters sharing to extract text representations, which are trained by multiple related tasks. Considering the structure of long text, document for example, it is composed of sentences, which are made up of words. Yang et al. [44] proposed the Hierarchical Attention Network (HAN), which

[3]https://anonymousrank.wixsite.com/here.

combines the Gate Recurrent Unit (GRU) and two levels of attention mechanisms to model the document representation. However, none of the above methods considers the relationship between labels, which is essential in multi-label text classification. In light of this, Yang et al. [42] treated this task as a sequence generation problem and modeled the correlation between labels through the LSTM structure. Hereafter, in order to strengthen the ability of the recurrent neural network to capture the long-term dependency in text, Zhao et al. [47] added dense connections between recurrent units and modeled inherent hierarchical structures in text.

Differently, the CNN-based models focus on modeling the relationship between the current word and its context by adjusting the filter size of CNN. In particular, Text-CNN [13] utilizes multiple convolutional filters to extract multiple feature maps, and then applies the max-pooling operation to obtain the sentence representation. Different from Text-CNN, which focuses on mining the relationship between words, Char-CNN [46] employs CNN to model the word characters to learn the text representation. As these models cannot learn variable N-gram features flexibly, Wang et al. [36] utilized densely connected CNN to obtain multi-scale N-gram features. Subsequently, an attention mechanism is adopted to adaptively select effective features from these features for text classification. Traditional convolutions always employ the same set filters regardless of different inputs, which lacks flexibility. In light of this, Choi et al. [4] designed the filter-generating networks, which generates filters dynamically conditioned on the inputs.

Although both CNN and RNN based methods have achieved promising performance, they cannot tag text with new classes. Thereby, they cannot be applied to the cQA sites, where new topics appear constantly.

2.1.2 Zero-shot Learning. Recently, zero-shot learning has made significant progress in the computer vision community [17, 23, 39], which aims to recognize unseen classes without training instances. Motivated by this, some zero-shot text classification methods [24, 26] have been introduced to tackle the new class issue. Unlike previous classification methods, zero-shot ones focus on modeling the semantics of classes and building the correlation between the text and each class. Therefore, they can easily generalize the unseen classes. For instance, Pushp et al. [24] proposed three simple neural networks to capture the relations between the text and class labels. Hereafter, based on the relationship among classes, Rios et al. [26] utilized two-layer GCNN to transfer information among classes for enriching their representations, especially the unseen ones. Despite these models achieve significant improvement in performance, they cannot be directly adopted to tackle the question tagging problem for the cQA sites. The reasons are as follows: 1) Pushp et al. [24] considers all textual classes are independent, while the topics are organized into a DAG in the cQA sites, as illustrated in Fig. 1(a). And 2) Rios et al. [26] argues that the topics linked to the current one have equal contributions to the topic representation modeling. However, the linked topics play different roles in modeling the semantic of the current topic. As shown in Fig. 1(a), the topic "Decoration" and " Building" may enrich the information of the topic "Interior Renovation" from different aspects. Considering the above issues, in this paper, we propose a novel question tagging model by designing a more flexible information propagation module.

2.2 Attention Mechanism

The attention mechanism has achieved promising performance in various tasks, such as image classification [34, 40], image/video retrieval [19, 25], and object detection [38, 45]. Inspired by its powerful ability, the attention mechanism has been applied to text understanding tasks, such as text classification [3, 21, 35, 44], question answering [12, 32, 41], and natural language inference [30, 31, 37]. To be more specific, Ma et al. [21] proposed an interactive attention mechanism to capture the relationship between the context and the target for classification. Yang et al. [41] devised a

hierarchical attention mechanism and a multi-head co-attention mechanism. The former integrates factual knowledge to better represent the question and answer, and the latter captures correlations between the question and answer. Tan et al. [30] designed four attention functions to match words in the sentence pair, and then aggregated matching information to predict the final results.

Recently, to capture the fine-grained information, the multi-dimensional attention mechanism [28] is proposed, which outputs a weight vector instead of a single scalar to represent the relation between two samples. The weight vector contains more associated information, which is able to operate on each of the dimensions. Besides, the self-attention mechanism is introduced in many researches [11, 33], which considers information from important positions to encode the sentences independently. Particularly, in the Natural Language Processing (NLP) domain, some studies applied the self-attention mechanism to generate the sentence-level embedding [18], while others utilize it to capture contextual information and model the long-term dependency of sentences [33].

In this paper, we utilize the attention mechanism to extract meaningful information from topic descriptions and questions. In addition, a multi-dimensional attention mechanism is designed to maintain useful information during the information propagation among topics.

3 OUR PROPOSED MODEL

The framework of our approach is illustrated in Fig. 2, comprising the following components: 1) the topic information fusion module fuses the name and description information of topics to generate initial topic representations; 2) the information propagation module enriches the topic representations via the DAG-structure; 3) the question encoder module represents questions via the multi-level contextual information modeling; and 4) the interaction module explores the interactions between the question and the topic, as well as estimates the ranking scores of topics. In what follows, we will first give the formulation of the problem and then introduce each module in detail.

3.1 Problem Formulation

Let $\mathcal{D} = \{(Q^i, T^i, y^i)\}_{i=1}^{M}$ denotes M training instances and $\mathcal{S}$ denotes the seen topic set, where Q^i refers to the i-th question, T^i represents a topic from $\mathcal{S}$, and y^i is an indicator label of the question-topic pair (Q^i, T^i). Specifically, if (Q^i, T^i) is a matched pair, y^i is 1; otherwise is 0. Besides, we have the unseen topic set $\mathcal{U}$, which satisfies the condition that $\mathcal{S} \cap \mathcal{U} = \emptyset$. In this work, given the instance set $\mathcal{D}$, we aim to learn a ranking model that enhances the representations of topics and estimates the relevance score of each question-topic pair. Moreover, in this paper, we consider two testing settings: 1) predicting the top-k relevant topics in $\mathcal{U}$ for a given testing question; and 2) predicting the top-k relevant topics in both $\mathcal{U}$ and $\mathcal{S}$ for a given testing question. In the following subsections, we will omit the upper superscript for a better understanding.

3.2 Topic Information Fusion

A topic T usually contains two types of information: name and description, as shown in Fig. 1(b). In this paper, we respectively utilize $\mathcal{W}^n = \{w_1^n, w_2^n, \ldots, w_{l_n}^n\}$ and $\mathcal{W}^d = \{w_1^d, w_2^d, \ldots, w_{l_d}^d\}$ to represent the topic name and description, where w_i^* and l_* ($*$ refers to n or d) separately indicate the word and the length of the corresponding sequence. As Fig. 1(b) illustrates, the name and description depict the topic from different levels. More concretely, the name reflects the coarse information, while the description characterizes the fine-grained one.

To better represent the topic, we jointly fuse the coarse-fine-grained information via a fusion model. As illustrated in Fig. 2, we present a topic information fusion module. Particularly, we first embed each word w_i^* to a feature vector $\mathbf{e}_i^* \in \mathbb{R}^k$, where k is the dimension of the word embedding. In this way, we can obtain the embedding matrix $\mathbf{E}^n \in \mathbb{R}^{k \times l_n}$ and $\mathbf{E}^d \in \mathbb{R}^{k \times l_d}$ for the topic name

and description, respectively. Afterwards, we apply average pooling to the name embedding matrix $\mathbf{E}^n$, and obtain the name embedding $\mathbf{t}^n \in \mathbb{R}^k$. However, since the topic description is very long and contains redundant information, directly utilizing the average pooling to obtain its representation may bring in noise. To avoid such issue, it is crucial to build a description processing model to adaptively select keywords from the description.

Inspired by this, we design an attentive description modeling scheme, which employs the topic name to filter out useless information from the topic description. Concretely, given the name representation $\mathbf{t}^n \in \mathbb{R}^k$ and the description embedding matrix $\mathbf{E}^d \in \mathbb{R}^{k \times l_d}$, we first capture the interactive information between the topic name and each word in the topic description. Hereafter, we compute the attention score and derive the description embedding $\mathbf{t}^d \in \mathbb{R}^k$. The specific operations are as follows,

$$\begin{cases} \mathbf{h}_j^d = \sigma(\mathbf{W}_1[\mathbf{t}^n \oplus \mathbf{e}_j^d] + \mathbf{b}_1), \\ d_j = \mathbf{W}_2 \mathbf{h}_j^d + b_2, \\ \alpha_j = \frac{\exp(d_j)}{\sum_{i=1}^{l_d} \exp(d_i)}, \\ \mathbf{t}^d = \sum_{j=1}^{l_d} \alpha_j \mathbf{e}_j^d, \end{cases} \tag{1}$$

where $\mathbf{e}_j^d \in \mathbb{R}^k, j \in \{1, 2, \ldots, l_d\}$ is the embedding vector of the j-th word in the description, d_j is the corresponding attention score for the j-th word, and the softmax function is applied to it to obtian the normalized score α_j. In addition, $\mathbf{W}_1 \in \mathbb{R}^{c \times 2k}$, $\mathbf{W}_2 \in \mathbb{R}^{1 \times c}$, $\mathbf{b}_1 \in \mathbb{R}^c$, and $b_2 \in \mathbb{R}$ are trainable parameters, $\oplus$ is the concatenation operation, and σ is the ReLU activation function.

Subsequently, we fuse $\mathbf{t}^d$ with the representation of the topic name $\mathbf{t}^n$. Although the topic name and description are complementary, their contributions to the topic representation may be different. To adequately exploit the helpful information from them, a highway network [29] is applied to fuse the description and topic name information, as well as output the initial topic representation. To be specific, we respectively learn two gate vectors to filter the topic name and description information as follows,

$$\begin{cases} \mathbf{G}_n = \sigma(\mathbf{W}_3 \mathbf{t}^n + \mathbf{b}_3), \\ \mathbf{G}_d = \sigma(\mathbf{W}_4 \mathbf{t}^n + \mathbf{b}_4), \end{cases} \tag{2}$$

where $\mathbf{W}_3 \in \mathbb{R}^{k \times k}$, $\mathbf{W}_4 \in \mathbb{R}^{k \times k}$, $\mathbf{b}_3 \in \mathbb{R}^k$, and $\mathbf{b}_4 \in \mathbb{R}^k$ are trainable parameters, and σ is the Sigmoid activation function. Thereafter, we could obtain the new topic representation $\mathbf{t} \in \mathbb{R}^k = \mathbf{G}_n * \mathbf{t}^n + \mathbf{G}_d * \mathbf{t}^d$, where $*$ refers to the element-wise multiplication operation.

3.3 Information Propagation

As shown in Fig. 1(a), topics in the cQA sites are organized into the DAG structure, where a child topic may have more than one parent topic, which is complementary to the current topic. Thereby, it is essential to take advantage of their parent topics to enhance the semantic understanding of the current topics, especially the unseen ones.

A straightforward approach is to apply the average pooling for parent topic representations to obtain a single representation and then combine it with the current topic. However, parent topics may contribute differently to the current topic. For example, in Fig. 1(a), the parent topic "Decoration" and "Building" enrich the semantic of the child topic "Interior Renovation" in different aspects. Motivated by this, we introduce the DAG-based information propagation module, as shown in Fig. 3, which utilizes the multi-dimensional attention to subtly filter the information of parent topics. In addition, we add a self-loop edge to each topic to preserve the original information.

Supposing that the current topic T has l_f parent topics and the j-th of them is represented by $\mathbf{t}^{f_j}$ ($j \in \{0, 1, 2, \ldots, l_f\}$), where $\mathbf{t}^{f_0}$ is equal to the current topic representation $\mathbf{t}$, corresponding to the

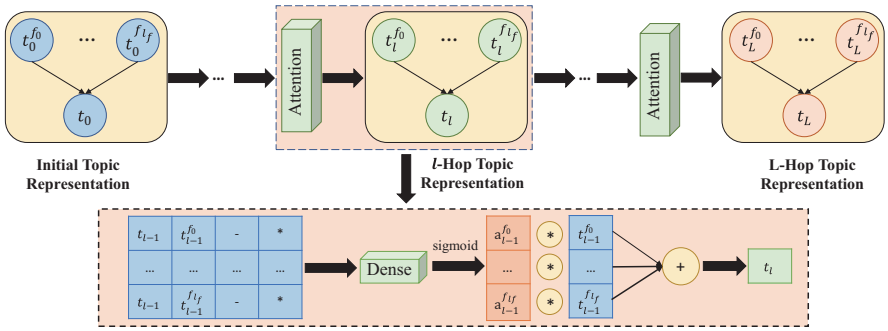

Fig. 3. Illustration of the information propagation module. It aims to map initial topic representations to new representations with the encoded DAG knowledge. For a better understanding, we show the $(l-1)$-th propagation process.

self-loop edge. To better characterize the relation between each parent topic and the current topic, we perform element-wise subtraction and multiplication on them to mine the relative information. More specifically, we adopt the concatenation operation for the current topic embedding, the parent topic embedding, and the element-wise relative information, to obtain the fused information $\mathbf{h}_{l-1}^{f_j} \in \mathbb{R}^{4k}$as follows,

$$\mathbf{h}_{l-1}^{f_j} = [\mathbf{t}_{l-1} \oplus \mathbf{t}_{l-1}^{f_j} \oplus (\mathbf{t}_{l-1} - \mathbf{t}_{l-1}^{f_j}) \oplus (\mathbf{t}_{l-1} * \mathbf{t}_{l-1}^{f_j})], \tag{3}$$

where $(l-1) \in \{1, 2, \ldots, L\}$ refers to the $(l-1)$-th propagation, L is the number of propagation, $\oplus$ is the concatenation operation,and $*$ represents the multiplication operation at the element level.

Inspired by the good performance of the multi-dimensional attention, we utilize the multi-dimensional attention vector to filter the useful information from the corresponding parent topics to obtain the updated topic representation $\mathbf{t}_l \in \mathbb{R}^k$. Concretely, we apply a nonlinear function for the fused information to obtain an attention vector for each parent-child pair. The attention vector stores the association between the current topic and the parent topic, and contains more information than the scalar. Formally, we summarize the above process as follows,

$$\begin{cases} \mathbf{a}_{l-1}^{j} = \sigma(\mathbf{W}_{l-1}\mathbf{h}_{l-1}^{f_j} + \mathbf{b}_{l-1}), \\ \mathbf{t}_l = \sum_{j=0}^{l_f} \mathbf{a}_{l-1}^{j} * \mathbf{t}_{l-1}^{f_j}, \end{cases} \tag{4}$$

where $\mathbf{W}_{l-1} \in \mathbb{R}^{k \times 4k}$ and $\mathbf{b}_{l-1} \in \mathbb{R}^k$ are trainable parameters, $\mathbf{a}_{l-1}^{j} \in \mathbb{R}^k$ denotes the dimension-wise attention vector of the j-th parent topic, σ is the sigmoid activation function, and $*$ refers to the element-wise multiplication. From the equation above, we find that the current topic T could obtain indirect information from its ancestor topics when the number of propagation is greater than one. As shown in Fig. 1(a), for the first propagation, the topic "Hydropower Transformation" gets the information from its parent topic "Interior Renovation", and the topic "Interior Renovation" obtains the information from its parent topics "Building" and "Decoration". For the second propagation, the topic "Hydropower Transformation" obtains the information from the topic "Interior Renovation" with the parent topic information.

As the DAG-structure is created by users in the cQA sites, it may contain some noise topics. To avoid the large offset of the topic information, a short-cut mechanism is adopted as follows,

$$\mathbf{t}^{'} = \mathbf{t} + \mathbf{t}_L, \tag{5}$$

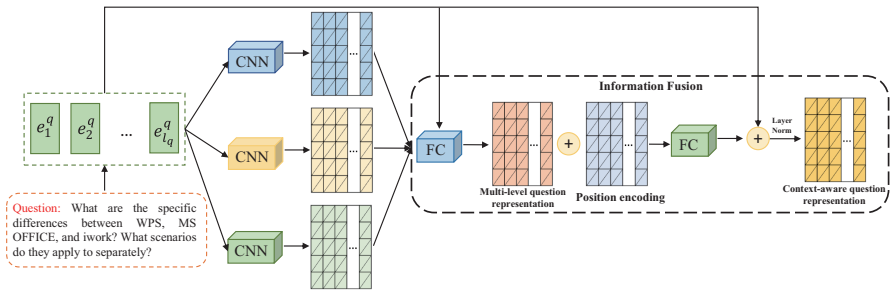

Fig. 4. Illustration of the question encoder module. The CNN modules in different colors represent that they have filters with different sizes. The multi-level question representation is obtained by integrating multiple question representations with different contextual information. Hereafter, the multi-level representation is merged with the position encoding to obtain the context-aware question representation.

where $\mathbf{t}_L$ is the topic representation output by the L-th information propagation, and $\mathbf{t}^{'} \in \mathbb{R}^k$ is the final topic representation.

3.4 Question Encoder

Having obtained the enhanced topic representations, we should encode the question to estimate the similarity score of each question-topic pair. The detail of this module is shown in Fig. 4. Given the question Q with l_q words, which is represented by $\mathcal{W}^q = \{w_1^q, w_2^q, \ldots, w_{l_q}^q\}$, we leverage the same embedding approach as the topic encoder to get the embedding vector $\mathbf{e}_i^q$ for each word w_i^q. And then we obtain the embedding matrix $\mathbf{E}^q$ for the given question. Considering that sometimes phrases are more meaningful than individual words, we employ multiple CNNs with filters of different sizes for questions to capture different contextual information, which is also beneficial for the following interaction modeling between questions and topics, formulated as,

$$\mathbf{E}^{q^j} = \theta_j(\mathbf{g}_j, \mathbf{E}^q), j \in \{1, 2, \ldots, m\}, \tag{6}$$

where θ_j indicates the j-th convolutional operation, which also contains bias and the ReLU activation function, as well as $\mathbf{g}_j \in \mathbb{R}^{k \times o_j \times k}$ refers to k filters of size $o_j \times k$, which aggregate o_j word vectors. To obtain results with the same size, in this work, we adopt zero padding to $\mathbf{E}^q$.

Having obtained multiple question representations, i.e., $\mathcal{E}^Q = \{\mathbf{E}^q, \mathbf{E}^{q^1}, \mathbf{E}^{q^2}, ..., \mathbf{E}^{q^m}\}$, we design a question information fusion module to output the multi-level question representation. To be specific, we first fuse them as follows,

$$\tilde{\mathbf{e}}_i^q = \sigma(\mathbf{W}_5[\mathbf{e}_i^q \oplus \mathbf{e}_i^{q^1} \oplus \mathbf{e}_i^{q^2} \oplus \ldots \oplus \mathbf{e}_i^{q^m}] + \mathbf{b}_5), \tag{7}$$

where $\mathbf{W}_5 \in \mathbb{R}^{k \times (m+1)k}$, $\mathbf{b}_5 \in \mathbb{R}^k$, σ is the ReLU activation function, and $\tilde{\mathbf{e}}_i^q$ is the i-th words representation after fusion. Though we consider different local contextual information, the position information is ignored. It indicates the time sequence of words explicitly, which is helpful to capture the semantic of multiple words. Therefore, we utilize position encoding as [33] did as follows,

$$\begin{cases} \mathbf{p}_{2j}^{pos} = \sin(pos/10000^{2j/k}), \\ \mathbf{p}_{2j+1}^{pos} = \cos(pos/10000^{2j/k}), \end{cases} \tag{8}$$

where pos is the value of position, and j is the dimension of position encoding. Afterwards, we add the position encoding for question representation to model the temporal information as follows,

$$\tilde{\mathbf{e}}_i^{q^{'}} = \sigma(\mathbf{W}_6[\tilde{\mathbf{e}}_i^q + \mathbf{p}^i] + \mathbf{b}_6), \tag{9}$$

where $\mathbf{p}^i$ is the position embedding for the i-th word in the question, $\mathbf{W}_6 \in \mathbb{R}^{k\times k}$, $\mathbf{b_6} \in \mathbb{R}^k$, and σ is the ReLU activation function.

To avoid the original question information loss, we utilize the short-cut mechanism to integrate the new and old question representation. Meanwhile, we also apply LayerNorm [2] to obtain the final representation $\mathbf{e}_i^{q'}$ of the word in questions, which contains the multi-level contextual information and position information. The above processes are summarized as follows,

$$\mathbf{e}_i^{q'} = LayerNorm(\tilde{\mathbf{e}}_i^{q'} + \mathbf{e}_i^q). \tag{10}$$

3.5 Interaction Module

As questions in the cQA sites are usually long and contain redundant information, the whole question sentence may introduce useless cues and further confuse the learning model. Therefore, building a model to adaptively choose helpful information for topic tagging is very necessary. To fill this need, we use topic information to filter out useless cues from the question. More concretely, an attention mechanism is developed, which can be formulated as follows,

$$\begin{cases} \mathbf{h}_i^r = \sigma(\mathbf{W}_7[\mathbf{t}' \oplus \mathbf{e}_i^{q'}] + \mathbf{b}_7), \\ r_i = \mathbf{W}_8\mathbf{h}_i^r + b_8, \\ \beta_i = \frac{\exp(r_i)}{\sum_{j=1}^{l_q}\exp(r_j)}, \\ \mathbf{q} = \sum_{i=1}^{l_q}\beta_i\mathbf{e}_i^{q'}, \end{cases} \tag{11}$$

where $\mathbf{W}_7 \in \mathbb{R}^{v\times 2k}$, $\mathbf{W}_8 \in \mathbb{R}^{1\times v}$, $\mathbf{b}_7 \in \mathbb{R}^v$, and $b_8 \in \mathbb{R}$ are learnable parameters, σ is the ReLU activation function, and r_i is the preliminary relevance score of the i-th word. And we utilize the softmax function to normalize the score and perform a weighted sum of the words in the question to obtain topic-aware question representation $\mathbf{q}$.

To further capture the interactive information between the question and topic, we adopt the information fusion approach as follows,

$$\mathbf{q}_{inter} = [\mathbf{q} \oplus \mathbf{t}' \oplus (\mathbf{q} - \mathbf{t}') \oplus (\mathbf{q} * \mathbf{t}')], \tag{12}$$

where $\mathbf{q}_{inter} \in \mathbb{R}^{4k}$ is the question-topic representation. Afterwards, we leverage a multi-layer perception network to the interactive information $\mathbf{q}_{inter}$, to predict the overall relevance score of the given question-topic pair (Q, T). Formally, it can be summarized as follows,

$$\begin{cases} \mathbf{h}_q = \sigma(\mathbf{W}_9\mathbf{q}_{inter} + \mathbf{b}_9), \\ s = \mathbf{W}_{10}\mathbf{h}_q + b_{10}, \end{cases} \tag{13}$$

where $\mathbf{W}_9 \in \mathbb{R}^{u\times 4k}$, $\mathbf{W}_{10} \in \mathbb{R}^{1\times u}$, $\mathbf{b}_9 \in \mathbb{R}^u$, and $b_{10} \in \mathbb{R}$ are trainable parameters, σ is the ReLU activation function, and s is the matching score of the question-topic pair (Q, T).

3.6 Loss Function

The loss function of our model is the sum of the binary cross-entropy of all instances, computed as follows,

$$\mathcal{L}_{loss} = -\sum_{i}^{M}[y^i log(s^i) + (1 - y^i)log(1 - s^i)], \tag{14}$$

where M is the number of training instances, and s^i is the score of the question-topic pair (Q^i, T^i) computed by Eqn. (13).

Table 1. Statistics of three datasets.

Dataset	Dataset I	Dataset II	Dataset III
# Questions	683,179	1,077,158	2,999,952
# DAG Edges	2,862	4,485	2,655
# Seen Topics	1,579	2,325	1,415
# Unseen Topics	582	926	584
Topics with description	1,161	1,965	-
Topics Per Question	2.39	2.56	2.34
Parents Per Topic	1.32	1.38	1.33
Avg. Length of Questions	14.19	13.23	12.91

4 EXPERIMENT

In this section, we sequentially detail the datasets, the evaluation metrics, and the implementation. Afterwards, we report the experimental results on three datasets to answer the following questions:

- **RQ1**: Can our proposed model achieve superior performance on the task of question tagging?
- **RQ2**: Does the description information benefit the question tagging performance?
- **RQ3**: Is the information propagation module of our proposed model helpful in boosting the ranking accuracy?
- **RQ4**: Does our proposed model perform better on the generalized zero-shot setting?

Thereafter, we perform the visualization analysis for the question-topic attention in the interaction module and the parent-child topic attention in the information propagation module. Finally, we conduct the qualitative analysis for our model and certain baselines.

4.1 Datasets

For the cQA sites, we did not find a suitable and public English dataset including the relationships between the topics. Therefore, in this paper, we evaluated our proposed HERE model over three Chinese question tagging datasets, i.e., Dataset-I, Dataset-II, and Dataset-III. Thereinto, Dataset-III is a publicly accessible benchmark dataset, while Dataset-I and Dataset-II are constructed by ourselves. All these datasets are built based on Zhihu, which is one of the most representative cQA sites. On Zhihu, numerous kinds of questions are created, answered, edited, and organized by users daily. The three datasets are detailed as follows.

Dataset-I: This dataset totally contains 683,179 questions labeled with 2,161 topics related to the "society" theme, and these topics are organized into a DAG with 2,862 edges. Thereinto, 1,161 topics contain descriptions, accounting for more than half of the total topics. In general, the internal topics with broader concepts appear earlier and are tagged with more questions. The leaf topics, tagged with few questions, are more likely to be new topics rather than the internal topics. Thereby, in this work, we randomly selected half of the leaf topics from the DAG as newly created topics, i.e., unseen topics. Finally, we formed a seen set of 1,579 topics and an unseen one of 582 topics.

Dataset-II: This dataset is collected in the same way as Dataset I, but the topics of its questions are related to the "life" and "sports" themes. There are 1,077,158 questions related to 3,251 topics in this dataset. Among all topics, 1,965 topics have descriptions, which is 60% of the total. Similarly, 926 topics are randomly selected from the leaf topics as the unseen set in experiments, and the remaining 2,325 topics are seen topics.

Dataset-III: The last dataset is released by the Zhihu machine learning challenge 2017[4], which aims to infer topics for untagged questions based on the bonding relationship between questions and topics. This dataset contains nearly 3 million questions tagged by 1,999 topics, and 2,655 directed edges connect these topics. As mentioned above, we selected 584 leaf topics as unseen topics, and the rest were treated as seen topics. Note that due to user privacy and data security, this competition does not provide the original textual information of questions and topics.

The statistics of the aforementioned three datasets are summarized in Table 1. We found that among three datasets, Dataset-II has the most topics, and its DAG is the densest. Note that to ensure the universality of our model, the topics with broader concepts, tagged with many questions, are selected as the root topics to create Dataset-I and Dataset-II. Moreover, we respectively split the datasets into 80%, 10%, and 10% as the training, validation, and testing set. Specifically, in the validation and testing set, 5% of the information is for the zero-shot setting, and the rest is for the generalized zero-shot setting.

4.2 Evaluation Metrics

To thoroughly measure our model and the baselines, we employed the weighted-Precision@K (P@K), Recall@K (R@K), and F_1 as evaluation metrics to measure the model performance from different angles. As each question has at most 5 topics, we hence set K to 5.

- P@5: Different from the traditional Precision@5 that is set as the fraction of relevant topics among five returned topics, we utilized weighted precision to encourage the relevant topic to be ranked higher. This indicator is defined by the Zhihu competition[5]. In addition, for this evaluation metric, the higher the value, the better the model performs. Formally, it is computed as follows,

$$P@5 = \sum_{pos \in \{1,2,3,4,5\}} \frac{Precision@pos}{log(pos + 1)}. \tag{15}$$

- R@5: It represents the proportion of the retrieved question-related topics to the ground truth topics. A high recall score means that the model returns most of relevant topics.
- F_1: It is the harmonic average of the precision and recall, formulated as,

$$F_1 = \frac{P@5 * R@5}{P@5 + R@5}. \tag{16}$$

4.3 Implementation Details

We performed the standard Chinese word segmentation with the help of jieba[6]. In this work, we respectively set the maximum length of questions, names, and descriptions as 30, 5, and 50. If the real length is less than the threshold, we padded it with zero; otherwise, we truncated the text. Moreover, we set the topic names as their descriptions for topics with no topic descriptions[7]. For our experiments, we adopted the pre-trained word2vec model [22] to generate the 256-dimensional (i.e., k=256) word embedding for Dataset-I and Dataset-II. As to Dataset-III, we utilized the word vectors provided by the Zhihu competition.

During the training process, we selected Adam [14] as our optimizer, and the learning rate is set to 0.001. For Dataset-I and Dataset-II, the mini-batch size is set to 500, while it is set to 1,000 for Dataset-III. For the question encoder module, we set up three convolutional neural networks

[4]https://biendata.com/competition/zhihu/.
[5]https://biendata.com/competition/zhihu/evaluation/.
[6]https://pypi.org/project/jieba/.
[7]For topics with no topic descriptions, we can set the topic names as their descriptions or pad it with the zero vector.

and the corresponding parameters o_j are 2, 3, and 4, respectively. Besides, the hyperparameters c, v, and u are all set to 256. We processed the original training set to generate question-topic pairs, and set the sample ratio of positive and negative to 1:1. During the testing, we considered two types of experimental settings: the zero-shot setting and the generalized zero-shot setting. To be more specific, for each testing question, the former predicts the top-5 relevant topics from the unseen set, while the latter returns the top-5 topics from both the seen and unseen set. Note that we mainly focused on the first setting to explore the effectiveness of our model. Moreover, our model is implemented in the MXNet framework with a NVIDIA GeForce GTX TITAN Xp GPU.

4.4 Experimental Results

In this part, we first introduced the baseline models and then presented our comparison results on three datasets. To evaluate the key component of our proposed approach, we conducted ablation studies. Hereafter, we compared our model with baselines on the generalized zero-shot experiment setting.

4.4.1 Baselines. There is no work that exploring the ability to tag questions with the novel topics in the cQA sites. Therefore, we selected methods related to the zero-shot text classification task as our baselines. As introduced in section 2.1, it is the most similar task to our work. Concretely, to demonstrate the effectiveness of our proposed HERE model, we compared it with the following state-of-the-art baselines.

- **W2VM** [22]: This is an unsupervised method for the text classification. It first adopts the average pooling to obtain representations for the text and label, respectively. And then the inner-product between these two vectors is set as their similarity score for classification.
- **Arch-I** [24]: Different from the W2VM that adopts the inner-product to calculate scores, it concatenates the embeddings of the sentence and category, and then utilizes the fully connected layer to output the similarity score.
- **Arch-II** [24]: This model uses the Long Short-Term Memory Network (LSTM) to encode the sentence, and then it concatenates its last hidden state with the embedding of the category. Afterwards, it passes the concatenated vector into a fully connected layer for classification.
- **Arch-III** [24]: Different from Arch-I and Arch-II, it concatenates the category label with each word of the sentence, and then sets them as the input of the LSTM. Hereafter, it passes the last hidden state into a fully connected layer for classification.
- **Dazer** [16]: It first utilizes the convolution neural network to model the interactive information between the category label and document. Thereafter, a category-specific gating mechanism is designed to filter the information obtained by the previous step. Moreover, to get category-independent information, adversarial learning is devised. Finally, the probabilities are predicted through a fully connected layer.
- **ZAGCNN** [26]: This model first builds an attention mechanism utilizing the label to find the most informative ngrams from the document. Hereafter, a two-layer GCNN is adopted to propagate information from associated labels to the current label. Finally, the dot-product of the label vector and the document vector is utilized to generate predictions. It is worth noting that ZAGCNN utilizes all negative examples during the training phase.

In this paper, to compare with our proposed model HERE, sentences/documents are treated as questions, and the corresponding category labels are regarded as topics.

4.4.2 Overall Comparison (RQ1). We conducted an empirical study to investigate whether our proposed model can achieve better tagging performance. For our model and baselines, we performed

Table 2. Performance comparison between our HERE and several state-of-the-art baselines over three datasets on the zero-shot setting. The best results are highlighted in bold. (p-value*: p-value over F_1)

Model	Dataset I				Dataset II				Dataset III			
	P@5	R@5	F_1	p-value*	P@5	R@5	F_1	p-value*	P@5	R@5	F_1	p-value*
W2VM	0.419	0.344	0.189	1.639e-15	0.343	0.275	0.153	2.062e-18	0.561	0.429	0.243	1.502e-15
Arch-I	0.608	0.512	0.278	1.448e-11	0.557	0.467	0.254	2.751e-13	0.546	0.447	0.246	8.232e-14
Arch-II	0.480	0.430	0.227	1.549e-12	0.500	0.425	0.230	8.749e-13	0.468	0.390	0.212	1.345e-14
Arch-III	0.717	0.584	0.322	2.564e-10	0.701	0.558	0.311	6.388e-11	0.682	0.525	0.297	1.735e-10
Dazer	0.745	0.599	0.332	6.099e-09	0.687	0.546	0.304	1.970e-11	0.729	0.551	0.314	4.770e-09
ZAGCNN	0.659	0.544	0.298	8.269e-12	0.642	0.526	0.289	2.215e-09	0.603	0.476	0.266	5.929e-09
HERE	**0.852**	**0.679**	**0.378**	-	**0.829**	**0.658**	**0.367**	-	**0.759**	**0.585**	**0.330**	-

them 5 times separately and calculated the average of these results. The results of all methods on three datasets are presented in Table 2, where several observations stand out:

- W2VM performs worse than the other baselines. The reason may be that: 1) It adopts the average pooling to extract question representations, which introduces the noise information. And 2) it overlooks the supervision information from seen topics, hence fails to well capture the discriminative information for differentiating topics.
- Interaction modeling methods, including Arch-III, Dazer, and ZAGCNN, surpass the Arch-I and Arch-II models. This verifies the necessity of interaction modeling between the topic and the question. Moreover, Arch-III and Dazer, especially Dazer, largely outperform ZAGCNN. It reveals that directly averaging all associated topics into one feature to enhance the current topic representation is inappropriate.
- Our proposed HERE model achieves the best performance, substantially surpassing all the baselines. Particularly, HERE presents consistent improvements over information propagation model ZAGCNN, reflecting the importance of employing the multi-dimensional attention mechanism and capturing interactions between topics on enhancing the topic representations. Meanwhile, our proposed model exceeds Dazer, because the latter ignores the topic relations hidden in the DAG-structure. Furthermore, we considered the topic description as the complementary information, which further strengthens the comprehension of topics. Moreover, we employed multiple convolution networks as the question encoder to obtain different contextual information and utilized the position encoding to emphasize the temporal information.

In addition, we also conducted the significance test over F_1 between our model and each of the baselines. We can see that all the p-values are substantially smaller than 0.01, indicating that the advantage of our model is statistically significant.

4.4.3 Justification of the Topic description (RQ2). To verify the effectiveness of topic descriptions, we conducted analytic experiments on two datasets: Dataset-I and Dataset-II. This is because Dataset-III does not release the valid description information. To be more specific, we compared our HERE method with the following variants:

- HERE w/o des: During the topic information fusion, we merely utilized the topic name to represent topics.
- HERE w/o HN: We eliminated the highway network from the topic information fusion.

Table 3 shows the results of these two variants on two datasets. From this table, we have the following observations:

Table 3. Component-wise validation of our proposed HERE model by disabling one component each time over three datasets. The best results are highlighted in bold.

Model	Dataset I			Dataset II			Dataset III		
	P@5	R@5	F_1	P@5	R@5	F_1	P@5	R@5	F_1
HERE w/o des	0.809	0.652	0.361	0.780	0.625	0.347	-	-	-
HERE w/o HN	0.831	0.672	0.372	0.820	0.653	0.364	-	-	-
HERE w/o dag	0.782	0.636	0.351	0.775	0.618	0.344	0.723	0.548	0.312
HERE w/o short-cut	0.805	0.659	0.362	0.750	0.615	0.338	0.710	0.563	0.314
HERE	**0.852**	**0.679**	**0.378**	**0.829**	**0.658**	**0.367**	**0.759**	**0.585**	**0.330**

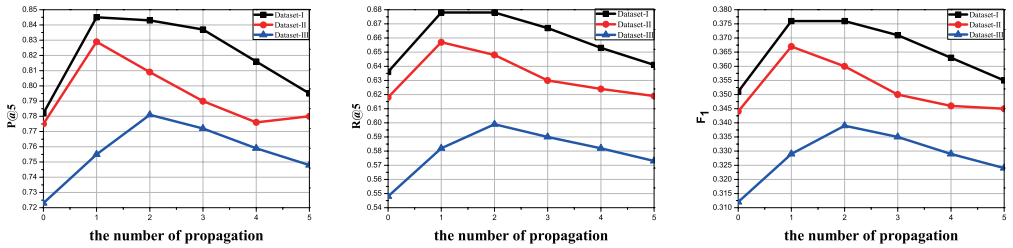

Fig. 5. Influence study regarding the propagation times on the ranking performance of the HERE model .

- Our proposed HERE model outperforms HERE w/o des by a large margin on Dataset-II and achieves considerable improvements on Dataset-I. It demonstrates that simply considering the name information cannot well characterize the content of topics. The words related to the topic name in the topic description can better represent the topic.
- The performance of HERE w/o HN has dropped by 2.1% on Dataset-I and 0.9% on Dataset-II in terms of P@5. It indicates that the name and description of topics contribute differently to the topic modeling. Specifically, the topic name plays a greater role in representing the topic since the description contains some unnecessary information.

4.4.4 Effectiveness of the DAG-structure (RQ3). Apart from achieving the superior performance, the key advantage of HERE over other methods is that its information propagation module is able to strengthen the topic representation. To verify it, we carried out experiments over three datasets. The variants of our model are as follows,

- HERE w/o dag: We eliminated the information propagation module from our learning model. Namely, we utilized the output of the topic information fusion module as the final topic features.
- HERE w/o short-cut: We eliminated the short-cut operation. That is, we only utilized the output of the information propagation as our topic representations.

From the illustration in Table 3, we gained the following insights:

- By jointly analyzing the performance of HERE w/o dag on three datasets, it can be seen that removing the DAG-based information propagation module degrades the ranking results. To be more specific, HERE w/o dag has dropped by 7.0% on Dataset-I, 5.4% on Dataset-II, and 3.6% on Dataset-III in terms of P@5. This verifies the effectiveness of the DAG-based topic information propagation.
- HERE surpasses HERE w/o short-cut, indicating that incorporating pre-propagation topic representations is beneficial to strengthen the final topic representation. This is because the

Table 4. Performance comparison between our proposed HERE model and several state-of-the-art baselines over three datasets on the generalized zero-shot setting. UR represents the recall value on unseen topics. The best results are highlighted in bold.

Model	Dataset I					Dataset II					Dataset III				
	P@5	R@5	F_1	R@S	R@U	P@5	R@5	F_1	R@S	R@U	P@5	R@5	F_1	R@S	R@U
W2VM	0.530	0.209	0.150	0.226	0.334	0.531	0.196	0.143	0.215	0.273	0.598	0.241	0.172	0.259	0.429
Arch-I	1.102	0.458	0.324	0.514	0.499	1.180	0.458	0.330	0.515	0.460	0.857	0.368	0.258	0.445	0.447
Arch-II	1.012	0.426	0.300	0.487	0.402	1.164	0.448	0.323	0.502	0.430	0.802	0.348	0.243	0.424	0.386
Arch-III	1.295	0.526	0.374	0.580	0.553	1.390	0.532	0.385	0.584	0.561	1.071	0.449	0.316	0.521	0.518
Dazer	1.278	0.522	0.371	0.583	0.581	1.363	0.522	0.378	0.582	0.545	1.085	0.453	0.320	0.527	0.545
ZAGCNN	**1.493**	**0.592**	**0.424**	**0.653**	0.498	**1.596**	**0.597**	**0.434**	**0.663**	0.531	**1.169**	**0.483**	**0.342**	**0.579**	0.480
HERE	1.371	0.557	0.396	0.608	**0.666**	1.453	0.559	0.404	0.609	**0.655**	1.096	0.460	0.324	0.534	**0.583**

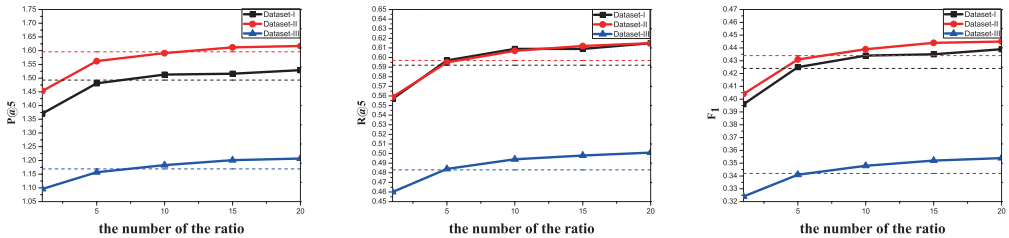

Fig. 6. Influence study regarding the sampling ratio on our proposed model HERE. The dash lines with the same color represent the ZAGCNN experimental results of the corresponding datasets.

DAG-structure created by users inevitably contains noise. In addition, the operation increases the discriminability of the propagated representation of topics with the same parent topics.

Moreover, we explored the influence regarding the propagation times L. The comparison results versus the L are illustrated in Fig. 5. We found that the performance consistently drops under different evaluation metrics when the propagation is conducted more than one or two times, especially P@5 drops significantly. This may be due to the fact that much more noise is introduced when considering the ancestor topics far off from the current topic, leading topics lack discrimination. Though the twice propagation outperforms the one propagation on Dataset-III, the one propagation on Dataset-I and Dataset-II is the best. Moreover, the one propagation is more efficient and has fewer parameters, we hence set $L = 1$ for efficiency ranking.

4.4.5 Comparison on Generalized Zero-shot Setting (RQ4). To further demonstrate the effectiveness of our proposed HERE model, we conducted experiments under the generalized zero-shot setting. In other words, the topics of the testing question contain seen topics and unseen ones. This experimental setting is more in line with the real-world scenario. Moreover, we added two indicators, i.e., R@S and R@U. The former represents the recall when the top 5 topics are selected among seen topics, and the latter refers to the recall over top 5 unseen topics. The results of all methods on three datasets are summarized in Table 4. And several observations stand out:

- Compared with the results reported in Table 2, all approaches achieve better performance because the test set contains well-trained seen topics. Meanwhile, the performance gap between baselines and our model is narrowed. This is because the number of seen topics is larger than that of unseen topics.
- Similarly, the unsupervised model W2VM performs worse than others. The interaction-based schemes, i.e., Arch-III, Dazer, and ZAGCNN, achieve better performance than the

Table 5. Visualization of the question-topic attention. The word attention is presented with different colors, and the darker red states the higher value.

	Question	Top-1 Topic
Q1	*What kind of mechanical keyboard do programmer use?*	*Mechanical keyboard*
Q2	*How to get a job offer in a Singapore kindergarten?*	*Kindergarten*
Q3	*Are there any fashion magazines suitable for boys around 20?*	*Men's clothing collocation*
Q4	*How to get closer to others through social software?*	*Social skill*
Q5	*Explaining to everyone how to choose pearls?*	*Jewelry*

Arch-I and Arch-II. In addition, interaction-based schemes outperform other methods on the evaluation metric R@U, indicating that fine-grained interactive information can improve the generalization ability of the model.

- The performance of ZAGCNN exceeds our proposed model HERE. The reason is that it utilizes all the negative samples, while the number of negative samples utilized by our model is the same as that of the positive ones. However, our model outperforms ZAGCNN in terms of R@U, even though the ratio between the negative sample and the positive one is 1:1. This indicates that ZAGCNN is more inclined to tagging questions with seen topics and unseen topics are not well represented. Differently, our model can obtain better semantic representation of unseen topics and has a strong generalization ability. This is important for unseen topics in tagging questions.

To further compare our proposed model with ZAGCNN, we added several experiments with different numbers of negative samples. In these experiments, the ratios between the negative sample and the positive one are 5:1, 10:1, 15:1, and 20:1, respectively. The corresponding experimental results are shown in Fig. 6. We utilized the dash lines of the same color to represent the ZAGCNN results of the corresponding datasets. From Fig. 6, we can find that our model achieves significant improvements with the increase of the sampling ratio on three datasets. When the sampling ratio is 5:1 and 10:1, the performance of our model rises obviously. When the sampling ratio is large enough, the performance is basically stable. Specifically, for Dataset-I, our model outperforms ZAGCNN on the indicator F_1 when the ratio is 5:1. Meanwhile, the experimental results of the other two datasets are very closed to ZAGCNN. When the ratio is 10:1, our model achieves better results than ZAGCNN on Dataset-II and Dataset-III. Therefore, with few negative samples, our model is able to outperform ZAGCNN, which performs well only relying on a large number of negative samples.

4.5 Visualization

In this section, we respectively conducted the visualization analysis for the question-topic attention in the interaction module and the parent-child topic attention in the information propagation module. The former aims to explore whether our interaction module could capture meaningful word information from the question, while the latter shows how the child topics obtain the effective information from their parent topics.

4.5.1 Visualization of the question-topic attention. As analyzed before, we fed the strengthened topic representation $\mathbf{t}^{'}$ and the representation of each word in the question $\mathbf{e}_i^{q^{'}}$ into an attention layer to model their relationship and obtain the topic-related question representation. To gain the deep insights into this attention mechanism, we randomly selected some questions from the testing

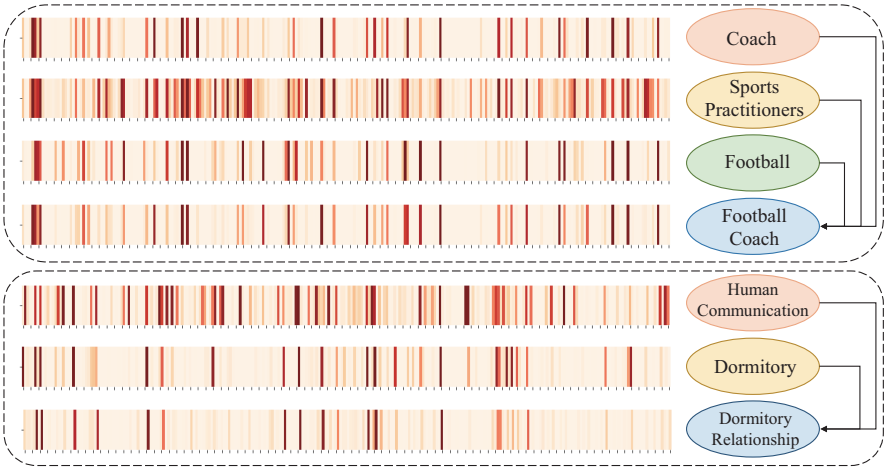

Fig. 7. Visualization of the parent-child topic attention, i.e., the $\mathbf{a}^j$ in Eqn. (4) of the 1-th propagation. The topics in the blue circle are child topics, and the other ones are parent topics. For each topic, we displayed the attention scores of all dimensions at left. The darker red indicates the higher value. We can clearly see that not all dimensional information is useful.

set to predict their topics under the zero-shot setting, and then visualized the attention values in Table 5. Specifically, in Table 5, the darker red states higher scores, while the scores of words in black are close to 0. Moreover, "Top-1 Topic" represents the first predicted topic of our model, which is also the ground truth topic for the given question.

From Table 5, several interesting observations stand out:

- The words in the question that appear in the topic have a larger attention score, such as "Mechanical keyboard" and "Kindergarten".
- The words related to the topics are marked in darker red. For instance, the topic "Social skill" pays more attention to the words "closer" and "social", while the topic "Jewelry" captures the important information "pearls" from the question. Moreover, the topic "Men's clothing collocation" captures the important related words, i.e., "fashion magazines", "boys", and "suitable".

These findings are consistent with our expectation, and further demonstrate that our proposed attention is capable of adaptively identifying the useful words according to the topic information, even though the topics are unseen. Hence, this verifies the effectiveness of our interaction module.

4.5.2 Visualization of the parent-child topic attention. In the information propagation module, a multi-dimensional attention mechanism is utilized to filter information from parent topics to enrich child topics. To explore the effectiveness of this module, we randomly selected two examples from the testing set, and then visualized attention vectors of their topics, as demonstrated in Fig. 7.

From the first example in Fig. 7, we can see that the parent topic "Sports Practitioners" attracts more attention along most dimensions than other parent topics. This may be because that the child topic "Football Coach" already contains the semantic information of the parent topics "Coach" and "Football". However, the parent topic "Sports Practitioners" contains complementary information for the topic "Football Coach". Therefore, the child topic "Football Coach" aggregates more information from its parent topic "Sports Practitioners" to enhance its representation. From the attention result shown in the bottom of Fig. 7, we found that the parent topic "Human Communication" attracts

Table 6. Qualitative examples under the zero-shot setting from Dataset-I and Dataset-II. The correct results are highlighted in bold.

Question	HERE	DAZER	ZAGCNN	Arch-III	Ground_truth
How to prevent the old person with senile dementia to dig to break blood scab?	*Healthcare* *Psychoanalysis* ***Nursing*** *Rehabilitation* ***Alzheimer's disease***	*Relative* *Pregnant women* *Healthcare* *Evidenced medicine* *Alopecia*	*Relative* *Depression* *Periodontal disease* *Stay up* *Healthcare*	*Healthcare* *Dermatitis* *Hair care* *Alopecia* *Prostatic massage*	*Alzheimer's disease* *Nursing*
What effort should a fresh graduate of the Japanese major make to enter the advertising industry?	*Japanese learning* ***Advertising copy*** *Broadcast* *Curriculum vitae* ***Advertising planning***	*Curriculum vitae* *Japanese grammar* *Investment banking* *Broadcast* *Struggle*	*Civil Service Exam* *Actionscript* *Japanese learning* *Broadcast* *Curriculum vitae*	*Japanese learning* *Curriculum vitae* *Siba Media* *Japanese grammar* *Broadcast*	*Advertising planning* *Advertising copy*
Why do children like to read books about dinosaurs?	***Child psychology*** *Psychology* *Children* *Mook* *Sex education*	*Family psychology* *Psychology* *Kindergarten* *children* *Psychology exam*	*Kindergarten* *Parenthood* *Children* *Erotic fiction* *ActionScript*	*Family psychology* *Children* *Mental health edu* *Adolescent education* *Infant feeding*	*Child psychology*
How to deal with unreasonable and powerful members of a team?	***Interpersonal conflict*** *DotA* *Sprayer* *Along* *Social skills*	*retort* *Rights protection* *DotA* *Along* *Sprayer*	*Friends* *Sparyer* *Depression* *Wechat business* *Let's talk*	*Making friends* *Social etiquette* *Socail skill* *Strangers socializing* *Along*	*Interpersonal conflict*

Table 7. Illustration of two failure examples under the zero-shot setting.

Question	HERE	DAZER	ZAGCNN	Arch-III	Ground_truth
Can cast iron pan cooking really replenish iron?	*Tableware* *Barbecue* *Bread* *Cook* *Western recipes*	*Delicacy* *Cook* *Tableware* *Bread* *Homemade food*	*Nutrition* *Cook* *Soup* *Tableware* *Bugslock*	*Delicacy* *Restaurant* *Cook* *Western recipes* *Rice noodles*	*Anemia*
What kind of company is Swisher?	*Animation design* *Font design* *Aircraft design* *Public interior* *Gold investment*	*Aircraft design* *Graphic Design* *Design ideas* *APP design* *Auto parts design*	*Hotboom* *Wechat business* *Gold investment* *Art appreciation* *Hot topics*	*Wechat business* *DotA* *Gender bias* *Gintama* *Travel strategy*	*Disinfection*

more attention than the other one do. Because it could propagate useful semantic information to the child topic "Dormitory Relationship", facilitating its comprehension of the word "Relationship". In addition, by jointly analyzing these two examples, we can see that the multi-dimensional attention vector can encode more information than a single scalar and filter the information of parent topics in a finer granularity.

4.6 Qualitative Results

Apart from the quantitative analysis, we also conducted the qualitative one to intuitively show the effectiveness of our model. In this section, examples are selected from Dataset-I and Dataset-II since Dataset-III provides encrypted text data. And three baselines with good performance are selected to compare with our proposed method. The corresponding qualitative experimental results under the zero-shot setting are summarized in Table 6.

From Table 6, we can see that experimental results of our proposed model are more accurate, while other baselines cannot tag questions exactly and even the predicted topics are very irrelevant. In particular, for the first question in Table 6, the topic "Nursing" and "Alzheimer's disease" are tagged by our model. This indicates that HERE can better capture the relationship between the words in questions and topics. Although the tagged topic "Healthcare" is not the ground truth, it is reasonable. As to the second question, our model correctly tag the ground truth topics to the question, while baselines fail. This is mainly because they pay more attention to the word "Japanese" in the question, ignoring to explore the relationship between the words "advertising industry" in

Table 8. Qualitative examples under the generalized zero-shot setting. The correct results are highlighted in bold, and the unseen topics are highlighted with the underline.

Question	HERE	DAZER	ZAGCNN	Arch-III	Ground_truth
A book planning editor wants to change his career to copywriting planning in the PR and advertising industry, recommended to join the industry?	*Change career* *Copywriting* *Career planning* ***Advertising copy*** ***Copywriting planning***	*Change career* ***Copywriting planning*** *Career planning* *Advertiser* *Editor*	*Web Editor* *Career planning* *Editor* *Magazine editor* *Change career*	*Change career* *Magazine* *Copywriting* *Publishing house* *Career planning*	*Copywriting planning* *Advertising copy*
Why is there pain associated with strabismus when I am sick (cold and fever)?	***Medicine*** ***Eye*** *Health* ***Disease*** ***Cold***	***Eye*** ***Medicine*** *Health* *Doctor* *Medical treatment*	***Medicine*** ***Eye*** *Health* *Vision care* *Ophthalmology*	***Medicine*** *Health* *Myopic eye* *Ophthalmology* *Vision care*	*Medicine* *Cold* *Eye* *Disease*
Is there any difference between the reflective lenses of the swimming goggles and the transparent ones?	***Swimming*** *Glasses* ***Swimming goggles*** *Ophthalmology* *Myopic eye*	*Glasses* *Filling a prescription* *Photography* *Vision* *Myopic eye*	***Swimming*** *Glasses* *diving* *Photographic equipment* *Photography*	*Glasses* *Filling a prescription* ***Swimming*** *Life* *Outdoor*	*Swimming* *Swimming goggles*
What should I do to protect my rights and interests when I encounter blackmail?	***Rights protection*** *Life* *Human communication* *Psychological counseling* ***Common sense of law***	*Life* *Human communication* *Health* *Lifestyle* *Communication*	*Life* *Common sense of life* *House type* ***Common sense of law*** *Human communication*	*Life* *Human communication* *Swindle* *Legal liability* *Consumer Rights*	*Common sense of law* *Rights protection*

the question and the ground truth topics. By jointly analyzing the third and forth questions, we can see that HERE can learn abstract semantics, such as "Interpersonal conflict", even though the topics are unseen. These qualitative results further show that our proposed model could well comprehend questions and model the interactive information between the questions and topics.

At the same time, we displayed some failure examples under the zero-shot setting in Table 7. For the first question, both our model and baselines tag wrong topics to the question. The reasons may be that 1) our model and baselines pay more attention to the phase "iron pan" and the word "cooking"; and 2) they lack some commonsense knowledge, such as iron deficiency may cause anemia. Similarly, to tag the last question correctly, the model needs to know the corresponding knowledge about the company Swisher. Otherwise, insufficient information will make the model guess wildly. Considering more entities information and commonsense knowledge may ameliorate this problem.

In addition, we also conducted the qualitative analysis under the generalized zero-shot setting, and the results are reported in Table 8. By jointly analyzing the results in Table 8, we can see that our proposed model HERE can tag questions more accurately with both seen and unseen topics. To be specific, for the first question, our model correctly tags the two unseen topics. The baseline methods, especially ZAGCNN and Arch-III, tag irrelevant topics, since they focus too much on the useless words (e.g., "editor"). For the second question, compared with the baselines, our model not only tag correctly the unseen topic "Cold" but the seen topic "Disease". Meanwhile, as to the second and third questions, we found that the baselines cannot tag the questions correctly even the topic appears in the question (i.e., "Cold" and "Swimming goggles"). For the last question, our proposed model captures the relationship between the question and the topic and the abstract semantic of the question under the generalized zero-shot setting, therefore correctly tagging the questions. All these results further verify the effectiveness of our proposed model.

5 CONCLUSION AND FUTURE WORK

This paper presents a graph-guided ranking model, which can tag questions with unseen topics in the cQA sites. Specifically, this method firstly considers the topic name and its description to obtain the initial topic representation. Hereafter, an information propagation module is designed to adaptively leverage the parent topic information to enhance the current topic. Simultaneously,

multiple convolutional neural networks are applied to obtain the question representation, which contains multi-level contextual information. Moreover, our proposed model can well capture the relationship between words in the question and the topic. To demonstrate the effectiveness of our method, we conduct experiments on three datasets. The experimental results demonstrate that our proposed model can achieve promising performance under various experimental settings.

In the future, we plan to make some attempts to further improve the performance of the question tagging task mainly from two directions. One is to strengthen the understanding of the question, the other is to enhance the representation of unseen topics. Particularly, first, we plan to add specific external knowledge, such as entity description. Identifying the entities in the question and linking them to the corresponding knowledge can further enhance understanding of the question. Second, we expect to explore generative adversarial networks [1, 8] to this task. Generative adversarial networks are widely used in many fields, such as image conversion [5, 9], and have achieved remarkable success. Using generative adversarial networks to generate examples of unseen topics and put them into the training may benefit the unseen topic representation.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China, No.: 61802231 and No.: 62006142; the Shandong Provincial Natural Science Foundation, No.: ZR2019QF001, the Key R&D Program of Shandong (Major scientific and technological innovation projects), No.:2020CXGC010111, as well as the special fund for distinguished Professors of Shandong Jianzhu University.

REFERENCES

[1] Martín Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein Generative Adversarial Networks. In *Proceedings of the International Conference on Machine Learning*, Vol. 70. 214–223.

[2] Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer Normalization. *CoRR* abs/1607.06450 (2016), 1–14.

[3] Jindong Chen, Yizhou Hu, Jingping Liu, Yanghua Xiao, and Haiyun Jiang. 2019. Deep Short Text Classification with Knowledge Powered Attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 6252–6259.

[4] Byung-Ju Choi, Jun-Hyung Park, and SangKeun Lee. 2019. Adaptive Convolution for Text Classification. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2475–2485.

[5] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bochao Wang, Hanjiang Lai, Jia Zhu, Zhiting Hu, and Jian Yin. 2019. Towards Multi-Pose Guided Virtual Try-On Network. In *IEEE International Conference on Computer Vision*. 9025–9034.

[6] Fuli Feng, Xiangnan He, Xiang Wang, Cheng Luo, Yiqun Liu, and Tat-Seng Chua. 2019. Temporal Relational Ranking for Stock Prediction. *ACM Trans. Inf. Syst.* 37 (2019), 27:1–27:30.

[7] Fuli Feng, Weiran Huang, Xiangnan He, Xin Xin, Qifan Wang, and Tat-Seng Chua. 2021. Should Graph Convolution Trust Neighbors? A Simple Causal Inference Method. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1–11.

[8] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems*. 2672–2680.

[9] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S. Davis. 2018. VITON: An Image-Based Virtual Try-On Network. In *IEEE Conference on Computer Vision and Pattern Recognition*. 7543–7552.

[10] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yong-Dong Zhang, and Meng Wang. 2020. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 639–648.

[11] Minghao Hu, Yuxing Peng, Zhen Huang, Xipeng Qiu, Furu Wei, and Ming Zhou. 2018. Reinforced Mnemonic Reader for Machine Reading Comprehension. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 4099–4106.

[12] Heyan Huang, Xiaochi Wei, Liqiang Nie, Xianling Mao, and Xin-Shun Xu. 2019. From Question to Text: Question-Oriented Feature Attention for Answer Selection. *ACMTransactions on Information Systems* 37 (2019), 6:1–6:33.

[13] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 1746–1751.
[14] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the International Conference on Learning Representations*. 1–15.
[15] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *5th International Conference on Learning Representations, ICLR*. 1–14.
[16] Chenliang Li, Wei Zhou, Feng Ji, Yuguang Duan, and Haiqing Chen. 2018. A Deep Relevance Model for Zero-shot Document Filtering. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. 2300–2310.
[17] Xirong Li, Shuai Liao, Weiyu Lan, Xiaoyong Du, and Gang Yang. 2015. Zero-shot Image Tagging by Hierarchical Semantic Embedding. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 879–882.
[18] Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A Structured Self-Attentive Sentence Embedding. In *Proceedings of the International Conference on Learning Representations*. 1–15.
[19] Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. 2018. Cross-modal Moment Localization in Videos. In *2018 ACM Multimedia Conference on Multimedia Conference*. 843–851.
[20] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent Neural Network for Text Classification with Multi-Task Learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 2873–2879.
[21] Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive Attention Networks for Aspect-Level Sentiment Classification. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 4068–4074.
[22] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of the International Conference on Learning Representations*. 1–12.
[23] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg Corrado, and Jeffrey Dean. 2014. Zero-Shot Learning by Convex Combination of Semantic Embeddings. In *Proceedings of the International Conference on Learning Representations*. 1–9.
[24] Pushpankar Kumar Pushp and Muktabh Mayank Srivastava. 2017. Train Once, Test Anywhere: Zero-Shot Learning for Text Classification. *Computing Research Repository* abs/1712.05972 (2017), 1–6.
[25] Leigang Qu, Meng Liu, Da Cao, Liqiang Nie, and Qi Tian. 2020. Context-Aware Multi-View Summarization Network for Image-Text Matching. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event*. 1047–1055.
[26] Anthony Rios and Ramakanth Kavuluru. 2018. Few-Shot and Zero-Shot Multi-Label Learning for Structured Label Spaces. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 3132–3142.
[27] Holger Schwenk, Loïc Barrault, Alexis Conneau, and Yann LeCun. 2017. Very Deep Convolutional Networks for Text Classification. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*. 1107–1116.
[28] Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Shirui Pan, and Chengqi Zhang. 2018. DiSAN: Directional Self-Attention Network for RNN/CNN-Free Language Understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 5446–5455.
[29] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Highway Networks. *Computing Research Repository* abs/1505.00387 (2015), 1–6.
[30] Chuanqi Tan, Furu Wei, Wenhui Wang, Weifeng Lv, and Ming Zhou. 2018. Multiway Attention Networks for Modeling Sentence Pairs. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 4411–4417.
[31] Yi Tay, Anh Tuan Luu, and Siu Cheung Hui. 2018. Compare, Compress and Propagate: Enhancing Neural Architectures with Alignment Factorization for Natural Language Inference. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 1565–1575.
[32] Nam Khanh Tran and Claudia Niederée. 2018. Multihop Attention Networks for Question Answer Matching. In *Proceedings of the International ACM SIGIR Conference on Research & Development in Information Retrieval*. 325–334.
[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the Advances in Neural Information Processing Systems*. 5998–6008.
[34] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. 2017. Residual Attention Network for Image Classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6450–6458.
[35] Pengfei Wang, Ze Yang, Shuzi Niu, Yongfeng Zhang, Lei Zhang, and ShaoZhang Niu. 2018. Modeling Dynamic Pairwise Attention for Crime Classification over Legal Articles. In *Proceedings of the International ACM SIGIR Conference on Research & Development in Information Retrieval*. 485–494.
[36] Shiyao Wang, Minlie Huang, and Zhidong Deng. 2018. Densely Connected CNN with Multi-scale Feature Attention for Text Classification.. In *Proceedings of the International Joint Conferences on Artificial Intelligence Organization*.

4468–4474.

[37] Shuohang Wang and Jing Jiang. 2016. Learning Natural Language Inference with LSTM. In *The Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1442–1451.

[38] Xudong Wang, Zhaowei Cai, Dashan Gao, and Nuno Vasconcelos. 2019. Towards Universal Object Detection by Domain Attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7289–7298.

[39] Xiaolong Wang, Yufei Ye, and Abhinav Gupta. 2018. Zero-shot Recognition via Semantic Embeddings and Knowledge Graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6857–6866.

[40] Zheng Yan, Weiwei Liu, Shiping Wen, and Yin Yang. 2019. Multi-Label Image Classification by Feature Attention Network. *IEEE Access* 7 (2019), 98005–98013.

[41] Min Yang, Lei Chen, Xiaojun Chen, Qingyao Wu, Wei Zhou, and Ying Shen. 2019. Knowledge-enhanced Hierarchical Attention for Community Question Answering with Multi-task and Adaptive Learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 5349–5355.

[42] Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. 2018. SGM: Sequence Generation Model for Multi-label Classification. In *Proceedings of the International Conference on Computational Linguistics*. 3915–3926.

[43] Yao-Yuan Yang, Yi-An Lin, Hong-Min Chu, and Hsuan-Tien Lin. 2019. Deep Learning with a Rethinking Structure for Multi-label Classification. In *Proceedings of the Asian Conference on Machine Learning*. 125–140.

[44] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical Attention Networks for Document Classification. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1480–1489.

[45] Xiaoning Zhang, Tiantian Wang, Jinqing Qi, Huchuan Lu, and Gang Wang. 2018. Progressive Attention Guided Recurrent Network for Salient Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 714–722.

[46] Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level Convolutional Networks for Text Classification. In *Proceedings of the Advances in Neural Information Processing Systems*. 649–657.

[47] Yi Zhao, Yanyan Shen, and Junjie Yao. 2019. Recurrent Neural Network for Text Classification with Hierarchical Multiscale Dense Connections. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 5450–5456.