Category-Highlighting Transformer Network for Question Retrieval

Abstract. Question retrieval, which aims to find the semantically equivalent questions from question archives for a user question, is crucial for question & answering systems. Recently, Transformer-based models have significantly advanced the progress of question retrieval and question answering. These models mainly focus on capturing the content-based semantic relations of two questions by self-attention mechanism. The question categories are very important to identify the semantic equivalence of two questions, because two questions with different categories cannot be semantically equivalent. However, Transformer-based models don't specially model the category information of questions, and thus can not well capture the category-based semantic relations of two questions. To capture both the content-based and category-based semantic relations of two questions, we study the issue of improving Transformer by highlighting and incorporating the category information. To this end, we innovatively propose the Category-Highlighting Transformer Network (CHT). Because both user questions and archived questions are not equipped with explicit categories, CHT first uses a category identification unit to automatically identify categories for a question and then to construct category-based semantic representations for the question and its embedded words. Secondly, to "deeply" capture the category-based and content-based semantic relations of two questions, we develop the category-highlighting Transformer by improving the self-attention unit with the category-based representations of words. The cascaded category highlighting Transformers are used for modelling "individual" semantics of a question and "joint" semantics of two questions. Extensive experiments have been conducted on three public datasets and the experimental results show that the category-highlighting Transformer network outperforms the state-of-the-art solutions.

Keywords: Question answering, Question retrieval, Transformer.

1 Introduction

Online question answering (QA) has been a popular platform where users can post their questions for seeking relevant answers, and also can reply the questions posted by other users. With the advances of QA systems, large-scale question and answer archives have become important information resources for selecting relevant answers to a new user query. Serving as a fundamental component of question answering systems, question retrieval aims to retrieve semantically equivalent questions from the question and answer archives. By the equivalent questions, a QA system will aggregate, rank and show their answers to users. Because of the importance of question retrieval, it has attracted much attention from the communities of both academia and industry [14, 13, 12, 4].

Meanwhile, a variety of models have been proposed for addressing the challenge of question retrieval, i.e., lexical gap— semantically equivalent questions have different words. The work [14, 6] takes the translation probability from an archived question to the user question as their semantic similarity. In work [8, 13, 4], the topic distributions of two questions are constructed and used for identifying their semantic equivalence. With the development of deep learning techniques, various neural network architectures have been proposed for deep encoding two questions [22, 15, 31]. Recently, pre-trained representation models, such as BERT [1] and ERNIE [2], have been proposed to model the semantics of questions, and have achieved significant performance gains.

These neural network architectures and pre-trained models focus on modelling the content-based semantic relations of two questions. Different from them, we propose new insights and solution to capture the semantic relations of both content-level and category-level.

Data insight: questions with different categories can not be semantically equivalent. The question categories represent the important semantics of a question, and are critical for the semantic equivalence identification of questions. This is because questions with different categories can not be semantically equivalent. For example, given two questions "how to rescue a phone that fell into the water" and "where to rescue a phone that fell into the water", they have high text similarity but are not semantically equivalent, because they belong to different question categories, i.e., "solution" and "location". To investigate universality of the insight, we make a survey over three public datasets, i.e., BankQ [19], LCQMC [20], and Quora[1]. First, we define some question categories, i.e., "time", "location", "people", "solution", "cause" and "object"; Second, we random select 500 negative examples from each dataset and manually label them with the defined categories; Third, we calculate the ratio of examples with different categories. About 40.29% of BankQ examples, 61.48% of LCQMC examples and 78.57% of Quora examples belong to different categories. The results tell us that question categories can help to distinguish the semantics of two questions.

Technical insight: Transformer can not well capture the category-based semantic relations of questions. Many Transformer-based solutions have been developed and achieved new state-of-the-art results of question retrieval. The self-attention mechanism in Transformer mainly focuses on modelling the contextual dependencies of words, while don't specially model the semantics on question categories. So Transformer-based solutions can not well capture the category-based semantic relations of two questions. To verify this insight, on three public datasets (BankQ, LCQMC and Quora), we analyze the examples mispredicted by Transformer-based solutions, i.e., BERT [1]. We find that 1) on the three datasets, 2.42%, 1.86% and 1.48% of negative examples with different categories are predicted as positive; 2) 1.21%, 2.56% and 2.59% of positive examples with the same category are predicted as negative.

Solution: category-highlighting Transformer network. According to the above insights, we propose to improve Transformer by incorporating the

[1] https://data.quora.com/First-Quora-Dataset-ReleaseQuestion-Pairs

category information, for capturing both content-based and category-based semantic relations of two questions. To realize this idea, we develop the category-highlighting Transformer network (CHT) with two cascaded units, i.e., category identification unit and category-highlighting Transformer unit. The first unit is to construct a category-based semantic representation for a question, while the second unit is to "deeply" encode both content-based and category-based semantic relations of two questions.

Category identification unit. In the datasets of question retrieval, the questions are not equipped with explicit categories. Inspired by the topic model where both sentences and words are represented as latent topics [33], the category identification unit is developed to automatically identify categories for a question, and then construct category-based semantic representations for the question and its embedded words. Specifically, we first define some question categories for a question dataset, i.e., "time", "location", "people", "solution", "cause", "object" and "other". Note that 1) fine-grained categories can also be defined and applied to our model; 2) we do not label each question with categories as training data. Based on the predefined categories, the category identification unit estimates the relevance between a question and each category, and then uses these relevance scores as weights to sum the embeddings of the predefined categories. The summed embedding is used for representing the category-based semantics of the question. The relevance between a word and the summed embedding as well as the word and predefined categories is estimated and used for constructing the category-based semantic representation of the word.

Category-highlighting Transformer unit. For deeply modelling both category-based and content-based semantic relations of two questions, we develop the cascaded category-highlighting Transformers to model "individual" semantics of a question and the "joint" semantics of two questions. The individual semantics focus on the meaning of one question itself, while the joint semantics focus on the deep semantic relations of two questions. In each Transformer, we leverage the category-based representations of words to improve the self-attention unit, so that both content-based and category-based semantics are captured and modelled. Specifically, based on the category-based representations of words, the category-based attentive similarities among words are estimated. As well, the context-based attentive similarities among words are estimated by using the scaled Dot-Product operation [23]. Secondly, the two types of attentive similarities are summed by a learnable weight to optimize the embeddings of words.

Our contributions are concluded as follows:

- We propose new insights for question retrieval from data and technical views.
- We develop the category-highlighting Transformer network to model both content-based and category-based semantic relations of two questions, without the supervision of question-to-category labelling data.
- We are the first to improve the self-attention unit in Transformer by incorporating the question category information.
- We conduct extensive experiments on three public datasets and validate the effectiveness of the category-highlighting Transformer network.

2 Preliminaries

In this section, we introduce the task of question retrieval, and review the related work about the second-stage, i.e., semantically equivalent question identification.

2.1 Question Retrieval

Question retrieval aims to find semantically equivalent questions from a large question repository D for a user question q. It can be formulated as follows:

$$\mathcal{D} = f(q, D) \tag{1}$$

where $\mathcal{D}$ is a subset of D and may contains one or more questions. The $f()$ is a pipeline of selecting candidate questions $\mathcal{D}$ from D, which often includes multiple stages. A classical pipeline consists of two stages, i.e., retrieval stage and identification stage. The retrieval stage is to find relevant candidates from a repository D, and the identification stage is to select the semantically equivalent questions from the relevant candidates as output. In practice, the retrieval stage needs to be completed in a limited time. So it typically uses some term-matching solutions to find relevant candidate questions, such as the vector space model [24], language model [25], and probabilistic model [26], because of their efficiency and effectiveness. In these methods, both user questions and archived questions are represented as bags-of-words (BOW), and many various functions of similarity estimation can be used for calculating the relevance of two BOW representations. Based on the term-matching solutions, numerous methods have been proposed to bridge the lexical gap, such as embedding-based methods [27] and data augmentation methods [28].

2.2 Related Work of the Identification Stage

The identification stage takes a user question q and the candidate question d as input, and identifies the semantic equivalence of the two questions. Finally, the semantically equivalent candidate questions are outputted. The semantic equivalence identification can be formulated as follows:

$$eq(q, d) = f_1(q, d) \tag{2}$$

where $f_1()$ is a function of determining whether q and d are equivalent. To construct an effective function $f_1()$, many solutions have been proposed [5, 32, 8, 10, 1]. These solutions can be grouped into four classes: term-matching solutions translation-model solutions, topic-model solutions and deep learning solutions.

Term-matching solutions. Many traditional techniques of information retrieval are used for estimating the semantic similarity of two questions, such as BM25 [29], language model [25] and vector space model [24]. These solutions are based on an assumption that two questions with higher text similarity are more likely to be semantically equivalent. Because of this point, the term-matching solutions can not overcome the lexical gap challenge.

Translation-model solutions. These solutions use the translation probability from one question to another question to identify the semantic equivalence

of the two questions [5, 32]. The study [32] proposes a retrieval model consisting of a translation-based language model for the question part and a query likelihood model for the answer part. The proposed models improve the traditional term-matching solutions by incorporating the word-to-word translation probabilities which are learned from different parallel corpora. Instead of the word-to-word translation, the study [6] proposes to learn the phrase-to-phrase translation probabilities, and thus identifies the semantic equivalence of two questions more accurately than using the word-to-word translation probabilities.

Topic-model solutions. These solutions [8, 13, 7] identify the semantic relations of two questions in the latent topic space. The work [8] proposes a topic model that incorporates category information into the process of discovering latent topics. The work [13] proposes a Question-Answer Topic Model (QATM) that learns the latent topics from the question-answer pairs, by assuming that a question and its paired answer share the same topic distribution. In the work [7], a topic-based language model is proposed to match questions not only at a term level but also at a topic level.

Deep learning solutions. The deep learning techniques have been widely applied in question retrieval, and achieve better performance than traditional solutions. The work [15] designs a convolutional neural network that learns an optimal representation of question pairs and a function to relate them in a supervised way from the available training data. The work [10] uses a bidirectional LSTM to generate multiple positional sentence representations for each question. In the work [9], the Siamese CNN is proposed to estimate the semantic similarity of two questions. Recently, the pre-training language models have demonstrated strong performance on text representations in many NLP tasks, such as BERT [1], Sentence-BERT [3] and ERNIE [2]. Meanwhile, many models are proposed based on various pre-trained language models to model the semantic relations of two questions. The work [11] presents a BERT-based FAQ retrieval system which uses BERT to estimate the correlation between a question and an answer. The work [16] develops a fully unsupervised method which uses question answer pairs to train two BERT models and uses the BERT models to match user queries to answers and questions, respectively. The work [4] proposes a topic-informed BERT-based architecture to model the semantic relations of the two texts. In work [17], the medical entity information is learned via ERNIE and incorporated into classification models.

3 Category-Highlighting Transformer Network

In this section, we introduce our solution to accomplish the identification stage of question retrieval. To accurately identify the semantically equivalent questions, we propose new insights: 1) *questions with different categories can not be semantically equivalent*; 2) *Transformer can not well capture the category-based semantic relations of questions* (see details in Section 1). Based on the insights, we propose to model both content-based and category-based semantic relations of two questions. To this end, we study the issue of improving Transformer by incorporating and highlighting the question category information. As the basis

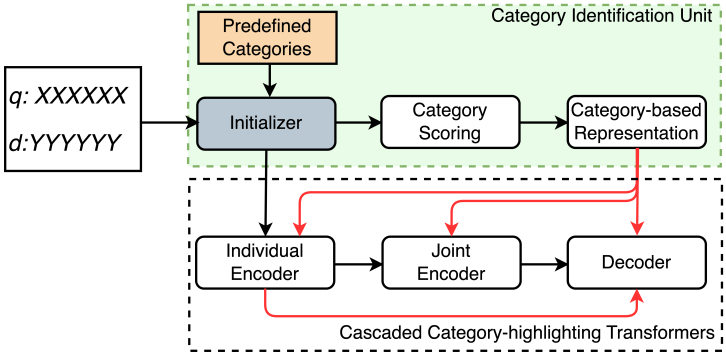


Fig. 1. The architecture of category-highlighting Transformer network.

of this study, we need to answer the following two research questions: 1) since questions in datasets are not equipped with explicit question categories, how to derive the categories or category-based semantics for a question? 2) assuming that questions are equipped with categories (or category-based semantic representations), how to use them to accurately model the category-based semantic relations of two questions?

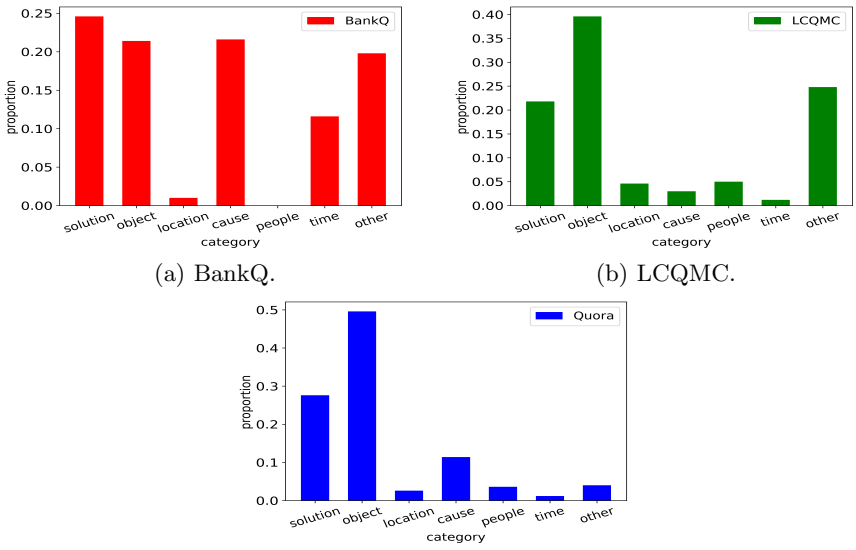
We design the category-highlighting Transformer network (CHT), and show its architecture in Figure 1. For the first research question, CHT exploits a category identification unit to automatically construct the category-based semantic representations for a question and its embedded words. For the second research question, we develop the cascaded category-highlighting Transformers by incorporating the category-based representations of words into the self-attention unit, so that both content-based and category-based relations of questions can be captured and deeply modelled. Moreover, the category-based representations of two questions and the deeply encoding results of the category-highlighting Transformers are combined in a decoder function to estimate the semantic equivalence of two questions. Accordingly, we formulate the category-highlighting Transformer network as follows:

$$eq(q,d) = \sigma_1(w_1[c(q); c(d); che(q,d)], b_1) \tag{3}$$

where $c(q)$ and $c(d)$ are the category-based representations of question q and d. They are constructed by the category identification unit. The $che(q,d)$ is the deeply encoding results over q and d, which contain both content-based semantics and category-based semantics. It is constructed by our cascaded category-highlighting Transformers. The $[;]$ is the concatenation operation. The σ_1 is a $m \times n \times 1$ MLP with activation function ReLU, ReLU and Sigmoid. The w_1 and b_1 are the weight vectors and bias vectors, respectively.

3.1 Category Identification Unit

As motivated in Section 1, in the datasets of question retrieval, both user questions and archived questions are not equipped with explicit question categories. To capture the category-based semantics, we propose to automatically construct



(a) BankQ. (b) LCQMC.

(c) Quora.

Fig. 2. Proportions of questions with different categories on three datasets.

the category-based representations for a question and its embedded words. Initially, we need to define some question categories for a dataset. Inspired by the work [30] that uses the 5W1H ("what", "when", "where", "why", "who", "how") to extract main events from news articles, we define question categories as "time", "location", "people", "solution", "cause", "object" and "other". Note that fine-grained categories can also be defined and applied to our model. To investigate the coverage of each question category, we make a data analysis over three datasets and show the results in Figure 2. Specifically, we randomly sample 500 questions from each dataset, and label each question with these predefined categories. From Figure 2, it can be seen that 1) most of questions belong to the "time", "location", "people", "solution", "cause" and "object" categories, and a few of questions belong to "other" category; 2) the coverage of the "solution" and "object" categories is higher than that of "location", "people" and "time" categories. Besides, we also find that 2.2% questions of BankQ, 1.6% questions of LCQMC and 3.4% questions of Quora belong to multiple categories.

Given a set of predefined question categories C, we construct the category-based semantic representation of a question q as follows:

$$c(q) = \sum_{c \in C} w(c, q) e(c), \quad \sum_{c \in C} w(c, q) = 1. \tag{4}$$

where $e(c)$ is the embedding of the question category c. To initial $e(c)$, we use BERT to encode the category c and take the embedding of $[CLS]$ as $e(c)$. To well construct the semantic representation over a set of categories, $w(c, q)$ is used for adjusting the weight of $e(c)$. An intuition is that $w(c, q)$ is higher, if c is more

relevant to q. So we estimate $w(c, q)$ as follows:

$$w(c, q) = \sigma_2(w_2[e(c); e(q); CoA(e(c), e(q))] + b_2) \tag{5}$$

where σ_2 denotes a MLP network with three layers, w_2 and b_2 are weight vectors. The activation functions of the network are ReLU, ReLU and Sigmoid. The CoA is a function of performing some interactions over $e(c)$ and $e(q)$, such as $e(c)-e(q)$, $|e(c) - e(q)|$ and $e(c) * e(q)$.

Inspired by the topic model where both sentences and words are represented as latent topics [33], we argue that in a question q, a word t_q should have a category-based representation. Since the word is embedded by a question, its category-based representation can be partly derived from the categories of the question, i.e., question-specific category representation. Besides, a word may appear in many questions, and thus it should have a global category representation. Combining the question-specific and global category representations, we construct the category-based representation for a word as follows:

$$c(t_q, q) = w(e(t_q), c(q))c(q) + \sum_{c \in C} w(e(t_q), e(c))e(c), \quad \sum_{c \in C} w(e(t_q), e(c)) = 1. \tag{6}$$

where $c(q)$ is the category-based representation of q. The $w(e(t_q), c(q))$ is the weight of $c(q)$, and $w(e(t_q), e(c))$ is the weight of $e(c)$, which are estimated by Equation 5. The $w(e(t_q), c(q))c(q)$ is to capture the question-specific category semantics, and $\sum_{c \in C} w(e(t_q), e(c))e(c)$ is to model the global category semantics.

3.2 Category-highlighting Transformer

To deeply model the content-based and category-based relations of two questions, we develop the cascaded category-highlighting Transformers as encoders, i.e., individual encoder and joint encoder (see Figure 1). Given a question pair q and d, the individual encoder models q and d individually, and derives the "individual" semantic representation for each question. It is to capture the meanings of a question itself. The joint encoder first concatenates the individual semantic representations of q and d, and then encodes the concatenated results to derive the "joint" semantic representation. It focuses on extracting the deep relations of two questions. So the combination of the individual semantic representation and joint semantic representation can capture the semantic relations of two questions more effectively than any single one. We formulate the cascaded category-highlighting Transformers as follows:

$$che(q, d) = [IT(q); IT(d); JT(IT(q), IT(d))] \tag{7}$$

where IT and JT are two category-highlighting Transformers, denoting the individual encoder and joint encoder respectively. For a question q (or d), IT first encodes the words in q to get their embeddings, and then averages the embeddings as the individual semantics of q. To get the joint semantics of q and d, JT first concatenates $IT(q)$ and $IT(d)$ with a start token [CLS] and separated

token [SEP], and then encodes the concatenated result. The embedding of [CLS] is taken as the joint semantics of q and d.

Original Transformer can effectively encode words by their contexts, and then construct the content-based semantic representations for questions. In the self-attention unit of the original Transformer, the attentive similarity between a word and other words is first calculated by using a scaled Dot-Product [23], and then used for updating the embedding of the word by using a weighted sum method. The scaled Dot-Product operation is performed on the parametrized embeddings of words, i.e., Q and K. Because Q and K are learned by the contexts of a word, the attentive similarity calculated by using a scaled Dot-Product can be considered as a context-based attentive similarity. Since these parametrized embeddings are not specially learning the question category information, Transformer can not well model the category-based semantics of questions.

To deeply model the content-based and category-based semantics relations of two questions, we design the category-highlighting Transformer by improving the self-attention unit with the category-based representations of words. Specifically, we first apply the category-based representations of words to estimate the category-based attentive similarities between a word and other words. Second, the context-based attentive similarities are estimated by the scaled Dot-Product. The category-based attentive similarities and context-based attentive similarities are combined as the final attentive similarities. The category-highlighting attention unit is formulated as follows:

$$Attention(Q, K, V, G, H) = softmax(\frac{QK^T + \lambda GH^T}{\sqrt{d_k}})V, \quad \lambda \in (0, 1) \qquad (8)$$

where $Q = E_q W^Q$, $K = E_q W^K$, $V = E_q W^V$, $G = C_q W^G$, $H = C_q W^H$. The $C_q = \{c(t_q^1, q), \cdots, c(t_q^n, q)\}$ is the category-based representations of words $\{t_q^1, \cdots, t_q^n\}$ in a question q, and $c(t_q^n, q)$ is estimated by Equation 6. The E_q is the context-based representations of words. The W^Q, W^K, W^V, W^G and W^H are projection matrices. The GH^T is to model the category-based attentive similarity, and QK^T is to model the context-based attentive similarity. The parameter λ is the weight of GH^T, to balance the importance of QK^T and GH^T.

4 Experiments

4.1 Experimental Setup

Overview of Experimental Objectives. To accomplish the task of question retrieval, we propose the category-highlighting Transformer network (CHT). In the network, the category identification unit is used for constructing the category-based semantic representations, and the cascaded category-highlighting Transformers are developed to model the deep semantic relations between two questions. Accordingly, our experimental objectives are designed as follows:

•O1: Can CHT better accomplish the question retrieval task than baselines?
•O2: Can CHT effectively model the category-based relations of two questions?
•O3: What are the performances of models on different question categories?

Table 1. Statistical information of datasets.

Dataset	Type	Train	Dev	Test
BankQ	positive	50000	5000	5000
	negative	50000	5000	5000
Quora	positive	139306	5000	5000
	negative	245042	5000	5000
LCQMC	positive	138574	4402	6250
	negative	100192	4400	6250

•O4: Ablation Study. What is the effectiveness of components in CHT?

Datasets Description. We conduct extensive experiments on three public datasets, and show the statistical information of the datasets in Table 1:

• BankQ [19]: It is the largest dataset of semantically equivalent question identification in the financial domain and sampled from the session logs of an online bank custom service system.

• Quora[2]: The dataset is sampled from a Q&A website Quora.com. Each question pair is labeled with a binary value that indicates equivalent or not.

• LCQMC [20]: It is a large-scale Chinese question retrieval dataset and sampled from the largest online Chinese question answering platform, i.e., Baidu Knows.

Comparison Solutions. According to the survey in Section 2.2, we select state-of-the-art solutions as baselines to verify the above experimental objectives:

• BiMPM [18]: BiLSTM model is used for encoding questions, and multiple matching methods are aggregated to model the relations of two questions.

• RE2 [31]: The aligned features, original point-wise features, and contextual features are applied to residual networks.

• BERT [1]: It is a well-known pre-training representation model, widely applied to many NLP tasks and achieves new state-of-the-art results.

• ERNIE [2]: It is a pre-training framework where the lexical, syntactic, and semantic information are learnt by using multi-task learning strategy.

• Sentence-BERT [3]: It is a framework with twin networks, and individually generates an embedding for every sentence.

• tBERT [4]: It first learns topic-based representations for words, and then use them to improve BERT for semantically equivalent question detection.

Performance metrics. Similar to the studies [4, 2, 31], we use accuracy and AUC metrics to measure the effectiveness of all models:

• Accuracy (Acc.): It is a widely used metric for evaluating classification models, and measures the fraction of correct predictions. Based on the positives and negatives, Acc. can be calculated as $Acc. = \frac{(TP+TN)}{TP+FP+TN+FN}$ where TP= true positives, TN = true negatives, FP = false positives, and FN = false negatives.

• Area Under Curve (AUC.)[21]: AUC. measures the two-dimensional area under the ROC curve. It represents the probability that a positive example is positioned in front of a random negative example.

[2] https://data.quora.com/First-Quora-Dataset-ReleaseQuestion-Pairs

Table 2. Performance comparison over BankQ. Since many user questions only have one candidate question, AUC metric can not be calculated on each user question, and then the statistical significance is not tested on the AUC metric.

Model	Acc.(%)	AUC.(%)
BiMPM	79.48†	87.50
RE2	81.07†	88.94
BERT	84.06†	89.34
ERNIE	84.66†	92.31
Sentence-BERT	83.53†	90.95
tBERT	80.65†	88.82
CHT	**85.24**	**92.43**

Table 3. Performance comparison over LCQMC.

Model	Acc.(%)	AUC.(%)
BiMPM	83.59†	93.75
RE2	84.74†	94.78
BERT	86.70†	94.84
ERNIE	87.17†	95.89
Sentence-BERT	84.07†	94.62
tBERT	85.12†	93.13
CHT	**88.98**	**96.23**

Reproducibility. The parameters of all models are assigned the default values for fair comparisons. Specifically, the batch sizes of models on the BankQ and LCQMC datasets are 32 and that of models on Quora datasets is 64. All models are developed in Python 3.8 and Pytorch 1.10 development environment. We set the learning rate to $2e-5$ and use the warm-up learning rate method. We apply dropout before every fully-connected layer of all models and set the dropout probability as 0.3. All models are deployed on the same computing services so that they have the same running conditions. In models, the parameters are optimized by the AdamW optimizer [34]. For fair comparisons, we perform every solution five times, and choose the best model to evaluate on the test set. The statistical significance is tested against CHT by a two-tailed paired t-test, and † denotes the difference at 0.05 level.

4.2 Experimental Results

To verify the experimental objective O1, we perform all models over the three public datasets, and show their performances in Table 2, 3 and 4. According to the metrics achieved by all models, BERT and ERNIE perform better than other baselines. tBERT first learns the topic distribution for each word, and then uses the topic distributions to improve BERT. But it is hard to construct the accurate topic distributions for a word because its context (question) is very short. On the three datasets, the average length of questions is 10.88, 10.93 and 12.68,

Table 4. Performance comparison over Quora.

Model	Acc.(%)	AUC.(%)
BiMPM	85.82†	93.46
RE2	89.20†	95.56
BERT	91.07†	96.96
ERNIE	91.08†	96.84
Sentence-BERT	88.69†	94.84
tBERT	89.31†	95.69
CHT	**91.67**	**97.11**

Table 5. Effectiveness of modelling the category-based relations of two questions. "ND" denotes the negative examples with different categories, "PS" denotes positive examples with the same categories. Since all examples in "ND" are negative and those in "PS" are positive, the AUC metric can not be calculated on "ND" or "PS".

Model	Acc.(%)	Acc.(%)	Acc.(%)
	ND	PS	*other*
BiMPM	84.52	82.16	74.48
RE2	84.59	85.82	75.61
BERT	90.48	83.25	80.52
ERNIE	89.24	86.77	80.31
Sentence-BERT	88.12	83.77	80.45
tBERT	83.47	84.59	76.19
CHT	**90.81**	**86.88**	**80.59**

respectively. Sentence-BERT is a modification of BERT and is to reduce the computation cost of BERT not improve the accuracy of BERT. Compared with BERT and ERNIE, the baseline BIMPM and RE2 are not pre-training models and can not benefit from the pre-trained embeddings. Besides, we can see that our model (CHT) performs better than all baseline models on three datasets. Compared with BERT and ERNIE, CHT not only benefits from the pre-trained embeddings but also benefits from the incorporation of question categories in Transformer. The above comparisons positively verify the experimental objective O1: CHT can better accomplish the question retrieval task than baseline models.

Can CHT effectively model the category-based relations of two questions (experimental objective O2)? To verify this objective, based on the BankQ dataset, we make an analysis. Specifically, we first divide the test examples into three groups by manually labelling, i.e., negative examples with different categories (ND), positive examples with the same categories(PS), and *other*. Secondly, we investigate the performances of models over the three groups and show the results in Table 5. It can be seen that 1) the advantage of CHT on the ND group is larger than that on the PS group. This is because examples with different categories can not be semantically equivalent, i.e., must be negative examples (see the data insight in Section 1), but examples with the same categories may be positive examples or negative examples. CHT can better capture the data insight than baseline models; 2) The advantage of CHT on PS group is larger

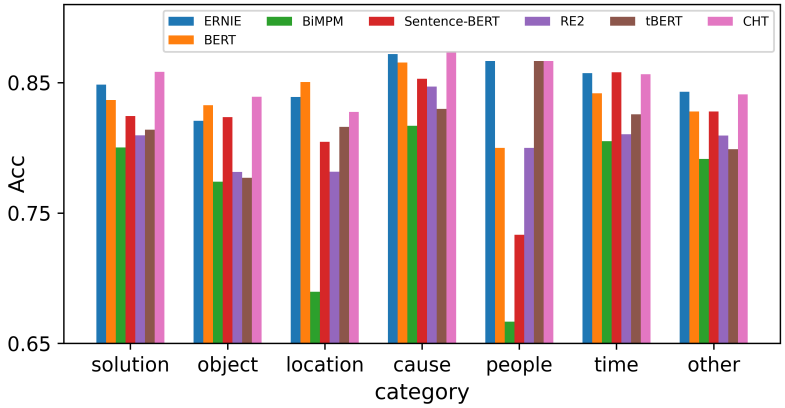


Fig. 3. Acc. comparisons of models on different question categories.

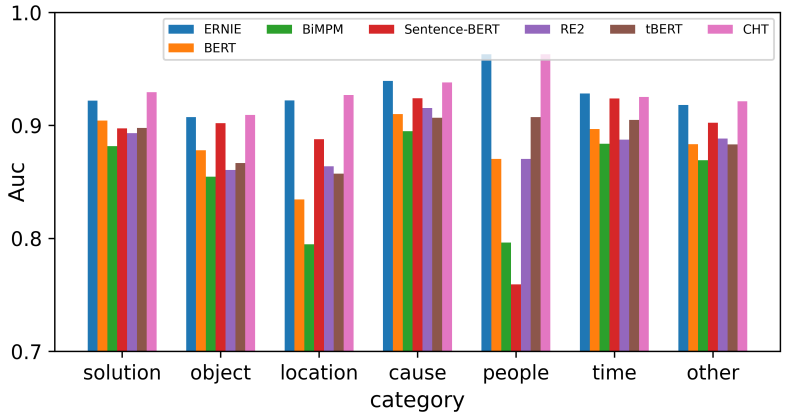


Fig. 4. AUC comparisons of models on different question categories.

than that on the *other* group. The examples with the same categories are more likely to be positive examples. These metric comparisons verify the ability of CHT on modelling category-based semantic relations of two questions.

What are the performances of models on different question categories (the experimental objective O3)? Based on the BankQ dataset, we investigate the performances of models on different question categories, and show the results in Figure 3 and 4. As introduced in Section 3, we define the question categories as "solution", "object", "location", "cause", "people", "time", "other". First, we can see that on the "solution", "object" and "cause" categories, CHT achieves better Acc. metrics than the best baseline ERNIE. On the "location" and "other" categories, the Acc. metrics of ERNIE are better than those of CHT. Second, about AUC metric, on "solution", "object", "location" and "other" categories, CHT performs better than ERNIE, while on the "time" category, ERNIE performs better than CHT. Overall, on the most of question categories, CHT performs better than baseline models.

Table 6. Ablation study: effectiveness of components in CHT.

Model	Acc.(%)	AUC.(%)
CHT-*che*+BERT	84.30	92.16
CHT-*cqd*	84.57	92.31
CHT-*gh*	85.04	92.27
CHT-*ind*	84.72	92.41
CHT-*joint*	84.71	92.31
CHT	**85.24**	**92.43**

To verify the effectiveness of components in CHT, we conduct the ablation study and present the results in Table 6. To capture the category-based semantic relations of two questions, CHT first uses the category identification unit to construct the category-based representations of questions, and then uses the cascaded category-highlighting Transformers to deeply encode the relations of two questions. In Table 6, the notation CHT-*cqd* denotes the category-based representations are not applied, i.e., $c(q)$ and $c(d)$ in Equation 3 are not applied. The CHT-*che*+BERT denotes the cascaded category-highlighting Transformers are not applied, and BERT is used for generating the deep encoding result $che(q, d)$ in Equation 3, i.e., the embedding of [CLS]. It can be seen that the metrics of CHT are higher than those of CHT-*che*+BERT and CHT-*cqd*. This illustrates the combination of the category-based representations and the deeply encoded results is better than any single one. The above comparisons not only verify the rationality of the architecture of CHT, but also verify the effectiveness of both the category identification unit and the cascaded category-highlighting Transformers. Besides, we find that the metrics of CHT-*che*+BERT are lower than those of CHT-*cqd*. The result illustrates that the deeply encoding results by cascaded category-highlighting Transformers are more important to model the relations of two questions than the category-based representations. This is reasonable because the deeply encoding results contain both content-level and category-level semantic relations, while the category-based representations only capture the category-level semantic relations.

In the cascaded category-highlighting Transformers, the first one is to model the individual semantics of a question, and the second one is to model the joint semantics of two questions. We test the two category-highlighting Transformers, and show their results in Table 6 where the CHT-*ind* denotes the individual semantics are not applied, and CHT-*joint* denotes the joint semantics are not applied. Comparing the metrics of CHT, CHT-*ind* and CHT-*joint*, we find that CHT performs better than CHT-*ind* and CHT-*joint*. This illustrates that the combination of the individual semantics and joint semantics can better capture the relations of two questions than any single one. Besides, on AUC metric, CHT-*ind* performs better than CHT-*joint*. This illustrates that the joint semantics are more important for modelling the relations of two questions than the individual semantics.

In the category-highlighting Transformer, we use the category-based representations of words to improve the self-attention unit of the original Trans-

former. Specifically, in Equation 8, both the category-based attentive similarity (GH^T) and context-based attentive similarity (QK^T) are used for updating the embedding of words. Comparing the original Transformer, we incorporate the category-based attentive similarity GH^T into the self-attention unit. To verify the effectiveness of this incorporation, we perform an experiment where GH^T is not applied to CHT and denote the results as CHT-gh in Table 6. Comparing the metrics between CHT and CHT-gh, we find that CHT performs better than CHT-gh. The comparisons illustrate that the incorporation of category-based attentive similarity helps to identify the semantic relations of two questions.

5 Conclusion

In this paper, we propose new insights from data and technical perspectives, to address the lexical gap challenge of question retrieval. To capture these insights, we develop a category-highlighting transformer network (CHT), which models both content-based and category-based semantic relations of two questions. Since questions are not equipped with explicit categories, CHT uses a category identification unit to construct the category-based semantic representations for a question and its embedded words. To deeply model both category-based and content-based relations of two questions, we propose the category-highlighting Transformer by improving the self-attention unit with the category-based representations of words. The cascaded category highlighting Transformers are developed to model "individual" semantics of a question and "joint" semantics of two questions. Extensive experiments demonstrate 1) the category-highlighting transformer network can better accomplish the question retrieval task than baseline models; 2) The incorporation of implicit question categories can effectively identify the semantic distinctions of two questions with high text similarity yet different categories. In the future, we plan to leverage the coarse-to-fine question categories to model the hierarchical semantic relations of questions.

References

1. Jacob Devlin and et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, NAACL-HLT 2019, 4171–4186
2. Yu Sun and et al., ERNIE 2.0: A Continual Pre-Training Framework for Language Understanding, AAAI 2020, 8968–8975
3. Nils Reimers and et al., Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, EMNLP-IJCNLP 2019, 3980–3990
4. Nicole Peinelt and et al., tBERT: Topic Models and BERT Joining Forces for Semantic Similarity Detection, ACL 2020, 7047–7055
5. Jiwoon Jeon and et al., Finding similar questions in large question and answer archives, ACM 2005, 84–90
6. Guangyou Zhou and et al., Phrase-Based Translation Model for Question Retrieval in Community Question Answer Archives, ACL 2011, 653–662
7. Kai Zhang and et al., Question Retrieval with High Quality Answers in Community Question Answering, CIKM 2014, 371–380
8. Li Cai and et al., Learning the Latent Topics for Question Retrieval in Community QA, IJCNLP 2011, 273–281

9. Arpita Das and et al., Together we stand: Siamese Networks for Similar Question Retrieval, ACL 2016
10. Shengxian Wan and et al., A Deep Architecture for Semantic Matching with Multiple Positional Sentence Representations, AAAI 2016, 2835–2841
11. Wataru Sakata and et al., FAQ Retrieval using Query-Question Similarity and BERT-Based Query-Answer Relevance, SIGIR 2019, 1113–1116
12. Preslav Nakov and et al., SemEval-2017 Task 3: Community Question Answering, ACL 2017, 27–48
13. Zongcheng Ji and et al., Question-answer topic model for question retrieval in community question answering, CIKM 2012, 2471–2474
14. Vanessa Murdock and et al., A Translation Model for Sentence Retrieval, EMNLP 2005, 684–691
15. Aliaksei Severyn and et al., Learning to Rank Short Text Pairs with Convolutional Deep Neural Networks, SIGIR 2015, 373–382
16. Yosi Mass and et al., Unsupervised FAQ Retrieval with Question Generation and BERT, ACL 2020, 807–812
17. Bhanu Pratap Singh Rawat and et al., Entity-Enriched Neural Models for Clinical Question Answering, Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing, BioNLP 2020, Online, July 9, 2020, 112–122
18. Zhiguo Wang and et al., Bilateral Multi-Perspective Matching for Natural Language Sentences, IJCAI 2017, 4144–4150
19. Jing Chen and et al., The BQ Corpus: A Large-scale Domain-specific Chinese Corpus For Sentence Semantic Equivalence Identification EMNLP 2018, 4946–4951
20. Xin Liu and et al., LCQMC: A Large-scale Chinese Question Matching Corpus, COLING 2018, 1952–1962
21. Tom Fawcett and et al., An introduction to ROC analysis, Pattern Recognit. Lett. 2006, 861–874
22. Chuanqi Tan and et al., Multiway Attention Networks for Modeling Sentence Pairs, IJCAI 2018, 4411–4417
23. Ashish Vaswani and et al., Attention is All you Need, NIPS 2017, 5998–6008
24. Tsatsaronis, George and et al., A Generalized Vector Space Model for Text Retrieval Based on Semantic Relatedness, EACL 2009, 70–78
25. Song, Fei and et al., A General Language Model for Information Retrieval, CIKM 1999, 316–321
26. Paik, Jiaul H and et al., A Probabilistic Model for Information Retrieval Based on Maximum Value Distribution, SIGIR 2015, 585–594
27. Ganguly, Debasis and et al., Word Embedding Based Generalized Language Model for Information Retrieval, SIGIR 2015, 795–798
28. Kauchak, David and et al., Improving Text Simplification Language Modeling Using Unsimplified Text Data, ACL 2013, 1537–1546
29. Robertson, Stephen and et al., The Probabilistic Relevance Framework: BM25 and Beyond, Found. Trends Inf. Retr., 333–389
30. Felix Hamborg and et al., Giveme5W1H: A Universal System for Extracting Main Events from News Articles, CoRR 2019
31. Runqi Yang and et al., Simple and Effective Text Matching with Richer Alignment Features, ACL 2019
32. Xiaobing Xue and et al., Retrieval models for question and answer archives, SIGIR 2008, 475–482
33. David M. Blei and et al., Latent Dirichlet Allocation, NIPS 2001, 601–608
34. Ilya Loshchilov and et al., Fixing Weight Decay Regularization in Adam, ICLR 2018