Medical Question Retrieval Based on Siamese Neural Network and Transfer Learning Method

Kun Wang[1], Bite Yang[2], Guohai Xu[1], and Xiaofeng He[1(✉)]

[1] School of Computer Science and Software Engineering, East China Normal University, Shanghai, China
51164500119@stu.ecnu.edu.cn, guohai.explorer@gmail.com, xfhe@sei.ecnu.edu.cn

[2] DXY, Hangzhou, Zhejiang, China
yangbt@dxy.cn
http://www.dxy.com

Abstract. The online medical community websites have attracted an increase number of users in China. Patients post their questions on these sites and wait for professional answers from registered doctors. Most of these websites provide medical QA information related to the newly posted question by retrieval system. Previous researches regard such problem as question matching task: given a pair of questions, the supervised models learn question representation and predict it similar or not. In addition, there does not exist a finely annotated question pairs dataset in Chinese medical domain. In this paper, we declare two generation approaches to build large similar question datasets in Chinese health care domain. We propose a novel deep learning based architecture Siamese Text Matching Transformer model (STMT) to predict the similarity of two medical questions. It utilizes modified Transformer as encoder to learn question representation and interaction without extra manual lexical and syntactic resource. We design a data-driven transfer strategy to pre-train encoders and fine-tune models on different datasets. The experimental results show that the proposed model is capable of question matching task on both classification and ranking metrics.

Keywords: Health care · Question matching · Transfer learning

1 Introduction

With the great improvement of public health consciousness, it is hard for traditional offline medical services to meet the rapidly increasing demands. To satisfy public health care demands, some medical communities, such as xywy.com and www.dxy.com, offer abundant medical knowledge, question answering and other online services. As shown that the number of users in online medical communities has increased to 192 million at the end of 2017[1].

[1] https://www.qianzhan.com/analyst/detail/220/181210-db903bba.html.

G. Li et al. (Eds.): DASFAA 2019, LNCS 11448, pp. 49–64, 2019.
https://doi.org/10.1007/978-3-030-18590-9_4

Usually, users can not get a timely answer from online doctors, and instead, they have to wait hours even days for response from online doctors. Actually, the medical communities have a large number of solved medical QA pairs in database. Most of these websites exploit retrieval systems to search related questions for new posted queries, and return topK QA results by computing the word correlation scores. The major challenge of this kind of retrieval system is lexical gap between different expressions of patients. For instance, for the query "Do hypertension patients require long-term medication?", similar question would be "When can hypertension patients stop taking medicine?", however, they have very low similarity score because they have few overlap words. On the other hand, the dissimilar question "Do hyperlipidemia patients require long-term medication?" has a high word overlap with the query. Patients may have different ways to express the same medical contents, which is observed from the collected question corpus. For example, users with medical background may take the professional word "hypoimmunity (免疫力低下)" as a query, while most patients would use "poor immunity (免疫力差)".

In previous works, researchers have proposed many methods to relieve semantic gap of sentence similarity problem, like translation models [9,25], topic model [10,26] and supervised neural networks [21,27]. The above methods have already been applied in general domain, while there are still several challenges when import these approaches into public health domain. Firstly, there is no open source collection of similar question pairs for model training in Chinese health care domain. The sample generation method of question pairs has great influence on the generalization performance of models. Secondly, linguistic resources for Chinese medical NLP tasks are scarce, such as medical encyclopedia and taxonomy, far from complete. Thirdly, patients have different expression ways, which is difficult to be learned only based on a small volume of training data. In addition, typical Siamese neural networks have two encoders with shared weights to process questions separately, without considering the interaction and alignment of information in both questions.

In this paper, we design a Siamese Text Matching Transformer (STMT) neural network to learn question representation and interaction in Chinese health care domain, which fully utilizes the context information and interactive information of question pairs. Specially, the key contributions of this paper are as follow:

- We propose a modified Transformer neural network namely STMT to incorporate context information and interactive information of question pairs to alleviate the problem of ignoring interaction information of typical Siamese neural network. The results on a large Chinese question pairs dataset demonstrate the effectiveness of the proposed model.
- For the lack of linguistic resource in Chinese health care domain, we build a large QA corpus and a huge health care terminology crawled from several well-known Chinese medical websites. We take two different approaches for training dataset annotation. The semantic relatedness between words in medical questions is captured by the pre-trained word embedding.

– A multi-class classification task is applied as a transfer strategy to pre-train encoders of Siamese neural networks, which helps the encoders to learn more general representation of similar questions and distinguish different concepts in medical semantic spaces. The experimental results show the advantages of transfer strategy than random initialization.

The rest of this paper is organized as follows: Sect. 2 discuss related research. Section 3 declares the details of proposed method. Section 4 describes the dataset and experiments. We analyze the results in Sect. 5 and conclude our work in Sect. 6.

2 Related Work

The task of question matching (QM) in community question answering (CQA) systems, has attracted increasing attention in recent years. This task is closely related to paraphrase identification (PI) problem [11]. The goal of PI task is to determine whether or not two sentences have the same meaning, but sentences that are non-paraphrase can still be semantic similar. Nevertheless, the methods share commonness. In addition, methods in QA and natural language inference (NLI) can also be lead into dealing with this problem.

At the early stage, retrieval based methods are widely used to find similar questions [14,16]. However, such approaches only consider lexical keywords, the detailed semantic information in the questions can not be captured. Another unsupervised methods is to train a topic model with unlabeled text corpus, and medical questions can be represented as vectors in latent topic space, and then, the similarity of questions is measured by the similarity of these mapped questions' vectors [10,26]. Some researchers regard this problem as a translation task [5,9,25], which incorporates lexical similarity and semantic similarity into a unified structure. The traditional translation models focus on translation probability of words or phrases among questions. The similarity of questions is the probability of translating one question to another [4]. [3] proposes an end-to-end neural network for question similarity learning, which utilizes RNN as encoder and decoder. The feature engineering based studies utilizes word overlap [8], linguistic features [20] and other hand-crafted statistic information to measure the lexical and phrase level similarity.

More recently, deep learning based methods are widely used to solve this problem [21,27]. Convolutional neural networks (CNN) are applied to extract multiple granularity features for similarity comparison [15,24]. [1,2] use two LSTM to compose two sentences and calculate the similarities between the encoded sentence vectors. These neural networks are called Siamese neural network [7], which process two sentences in parallel. In medical QA domain, [18] proposes an interactive attention based LSTM model to evaluate the similarity of question pairs, which also takes LSTM as question encoder and only focuses on unidirectional attention interactivity. [4] combines translation model and Siamese CNN model to learning question-question similarity and question-answer relations, which is better than traditional retrieval algorithms, like BM25

[16]. BiMPM is design to learning sentence matching from multiple perspectives [23]. [6] propose ESIM model to solve NLI task, which utilize stacked LSTMs and attention mechanism to capture interactive similarity information of sentence pairs. The above two models are the state-of-art method and achieve good performance on different datasets [11,23].

3 Methodology

In this section, we first define the question matching task in Chinese health care domain. Then, we introduce sentence encoders for learning question representations, which are the core component of Siamese neural networks. Next, we describe the similarity measurements and architecture of our proposed model. Further, we explain how we address the problem of dataset generation in Chinese medical domain. Finally, we discuss about a transfer strategy to pre-train encoders to improve the performance of question matching models.

3.1 Task Definition

We first define the task formally and declare some notations that we used in this paper. Given an unsolved medical question $Q = [w_1, w_2, \ldots, w_n]$ containing n words. We embed each word of question as distributed word vector, then $Q = [\boldsymbol{q_1}, \boldsymbol{q_2}, \ldots, \boldsymbol{q_n}]$, where $\boldsymbol{q_i} \in R^{d_e}$ is the word embedding of the i-th word in this question. Define a set of relevant candidate questions $C = \{C_1, C_2, ..., C_m\}$, retrieved from large solved medical question corpus. We need to determine whether or not each candidate C_i is similar to Q and rank them by similarity scores to query Q. We describe the workflow of finding similar medical question in Fig. 1.

3.2 Encoder

In previous research, Siamese RNN structure consists of two share-weights encoders, such as LSTM or GRU. Each encoder process one question in the given question pair [1,2]. We first introduce bidirectional LSTM encoder that we used in baseline and then present the modified Transformer encoder designed in this section.

BiLSTM. We take bidirectional LSTM encoder to learn sentence level representation of medical questions. The unidirectional LSTM model process word sequence from left to right, when LSTM cell goes through the whole question, the information of previous words can be transmitted and accumulated into its memory cell, and the output of last hidden state is used as vector representation of question. The bidirectional LSTM consists of a forward LSTM and a backward LSTM, at each time step t, forward LSTM computes a representation $\overrightarrow{h_t}$ with the left context of word x_t, and the backward LSTM computes a representation

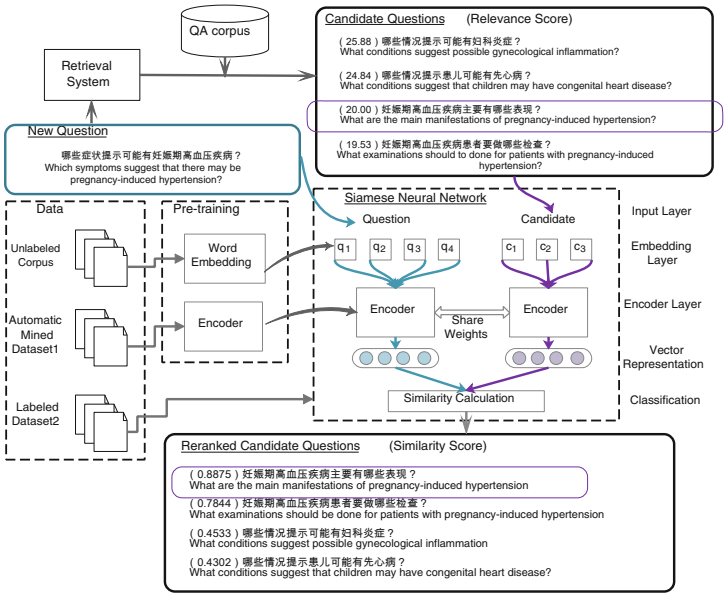


Fig. 1. The workflow of similar medical question retrieval.

$\overleftarrow{h_t}$ of the same sequence in reverse order. Then the representation of word x_t is obtained by concatenating its left and right context representations, namely $h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}]$. Thus, the output of BiLSTM would capture the abundant context information of medical questions at sentence level. The concatenated final hidden state $h_n = [\overrightarrow{h_n}, \overleftarrow{h_n}]$ of BiLSTM is used as the semantic vector representation of questions, thus, we take v_q to represent the question vector, then $v_q = h_n$.

TM-Transformer. Transformer is a neural network architecture proposed by Google for machine translation [19], which adopts multi-head attention to encode sentence instead of RNN models. The basic attention mechanism in Transformer model is scaled dot-product attention, which is described formally as follow,

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

Where $Q \in R^{n \times d_k}$, $K \in R^{m \times d_k}$ and $V \in R^{m \times d_v}$ are sentence embedding matrices. In translation task of [19], the K and V are the same representations of a sentence and Q is the other one of an aligned bilingual sentence pair. The output of such attention is called aligned embeddings.

To enhance the efficiency and effectiveness, instead of taking d_e-dimensional Q, K, V to perform a single attention function, the multi-head attention linearly projects the Q, K and V matrices h times to d_k, d_k and d_v dimensions, respectively. The scaled dot-product attention is performed on each of these projected

queries, keys and values in parallel. The results of h times attention functions are concatenated and linearly projected, resulting in final aligned sentence representations.

$$MultiHead(Q, K, V) = Concat(head_1, \ldots, head_h)W_o \quad (2)$$

$$head_i = Attention(QW_i^q, KW_i^k, VW_i^v) \quad (3)$$

Where the projections of i-th head are parameter matrices $W_q \in R^{d_e \times d_k}$, $W_k \in R^{d_e \times d_k}$, $W_v \in R^{d_e \times d_k}$ and $W_o \in R^{hd_v \times d_e}$. Since that the input is d_e-dimensional question embeddings in this paper, then we set $d_k = d_v = d_e/h$.

Multi-head attention allows the model to jointly process information from different representation subspaces at different time step. Inspire of the encoder and decoder of Transformer in [19], we proposed a multi-head attention based encoder namely TM-Transformer for question matching problem, as show in Fig. 2. In this work, the TM-Transformer model consists of three sub-layers.

(1) The first sub-layer is **Self-Attention encoder**, which means the input of Q, K, V are the same questions, as shown in Fig. 3. This sub-layer is used to reformulate words by capturing context information in sentences.
(2) The second sub-layer is **Inter-Attention encoder**, which has the same structure as the first sub-layer. However, the input of this sub-layer is different that Q is different from K and V. If Q is the embedding of an unsolved medical question, the K and V represent the same candidate questions, or vice versa. In this sub-layer, the inter-attention is used to encode question with the aligned context information from the other question in question pair.
(3) The third sub-layer is **Feed-Forward layer**, which consists of two convolutions with kernel size 1 as linear transformations. The first convolution has dimensionality $d_f = 2 * d_e$ and the dimensionality of the second one is $d_e = 100$.

A global max pooling layer is used to retain important information of the new representation of questions along the words dimension. At last, questions can be represented as final semantic vectors. The number of the TM-Transformer layer is set as $N = 6$ in this paper, which is as same as [19].

Similarity Calculation. For a given question pair $<Q, C_i>$, with the pre-trained encoder model, we map Q and C_i into semantic vectors v_q and v_{c_i}, respectively. In most neural network architectures [6,18,23], a multilayer perceptron (MLP) layer is used for label prediction. We put the concatenated vector $\boldsymbol{v} = [v_q, v_{c_i}]$ into a final MLP classifier in our experiments. The MLP has a hidden layer with *relu* activation and a *softmax* output layer. The final score range in [0, 1], where 1 represents similar and 0 is dissimilar. In this work, we take binary-class entropy loss for training and the entire model is trained end-to-end. Figure 2 shows the proposed neural network architecture.

3.3 Data Generation

In this section, we first introduce the acquisition of the medical QA corpus, and how we build a large medical terminology. Then, we declare how we generate the training and test dataset with two different automated ways.

Medical QA Corpus. In order to satisfy the demands of generating the word embedding and the training datasets, we require a large medical QA corpus that covers most of the medical questions and terms. We have crawled 4 online medical QA websites and acquired privacy-free QA data from cooperative company www.dxy.com. The statistics of obtained QA pairs is shown in Table 1.

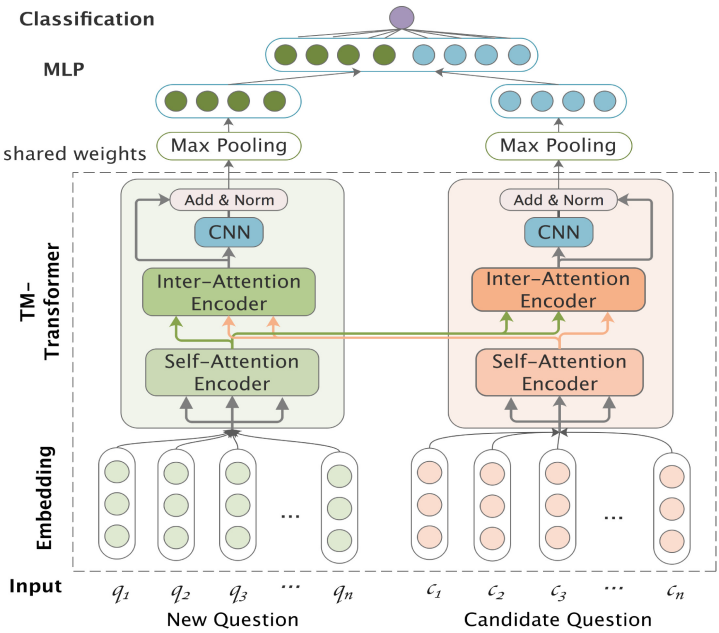


Fig. 2. The framework of Siamese TM-Transformer neural network.

Medical Terminology. According to the analysis of questions in medical QA corpus, we have found that the medical terms in patients queries are usually not exactly equivalent to professional terminology used by medical experts. So in our work, we have collected the medical terms from these sources: (1) the Chinese medical terms of www.dxy.com, which is compiled by professional medical editors, referring to a large number of *medical encyclopedia entries* and *general expressions* of users. (2) the common disease names and drug names crawled from China Public Health Website[2], and the drug names from China Food and

[2] http://www.chealth.org.cn/.

Drug Administration[3]. (3) the disease names, symptom names, and their common aliases and other entities on the medical community websites. Finally, we have built a medical terminology that contains 26,619 diseases, 9,904 symptoms, 1,735 medicines, 8,734 examinations, 7,395 operations, 5,472 body parts and other medical entities.

Training Datasets. In previous research [12], the dataset is generated by sampling approach. The negative samples are generated by randomly sampling medical question pairs that share no common medical entities, which directly neglects dissimilar question pairs with common entities. They randomly replace and drop words in medical questions to generate positive samples. In fact, this sampling approach doesn't consider that the replacing and dropping of words may break the semantic characteristics and syntactic structure of the new generated questions.

Table 1. Statistics of the QA pairs

Websites	# of QA pairs
www.dxy.com	623,395
www.xywy.com	1,357,507
www.39health.com	997,717
www.muzhi.com	6,442
www.120ask.com	68,045
Total	3, 053, 106

Table 2. Brief description of the generated datasets

	Dataset1		Dataset2	
	Train	Test	Train	Test
# of Q	28,045	14,269	165,103	36,094
# of Q&C	-	-	264,685	46,800
# of Group	5,000	3,317	-	-

In our work, we adopt two different generation approaches, which are adapt to different training scenarios. We find that there are some similar questions on each QA web page edited by medical websites, such as xywy.com. In the first generation method, we randomly sample 10,000 different medical questions from the QA corpus as question seeds, and crawl their web pages to acquire their similar questions provided by websites. We remove probably dissimilar questions that have different disease categories with the seed. Finally, about 60,000 questions are automated categorized into 8,317 synonymous sentence groups after filtering out the seeds without similar questions. Several generated question groups are shown in Table 3. In our work, this dataset namly ***Dataset1*** is used for transfer strategy introduced in Sect. 3.4.

Another method generates question pairs with the help of opensource retrieval system. We randomly sample 80 thousand different questions as queries and search their related questions from the medical QA corpus with Solr[4]. The

[3] http://eng.sfda.gov.cn/WS03/CL0755/.
[4] http://lucene.apache.org/solr/.

Table 3. Synonymy question group sample of Dataset1

Group	Medical Questions
0	慢性荨麻疹怎么治疗?(How to treat chronic urticaria?)
0	慢性荨麻疹如何调理?(How to regulate chronic urticaria?)
0	怎样才能根治慢性荨麻疹?(How can I cure chronic urticaria?)
0	得了慢性荨麻疹怎么办?(I have chronic urticaria. What should I do?)
1	高血压伴发的症状.(Symptoms associated with hypertension.)
1	高血压的常见症状有哪些?(What are the common symptoms of hypertension?)
1	高血压的临床表现有什么?(What are the clinical features of hypertension?)

BM25 algorithm built in Solr rank the candidate questions by their *relevance scores* with query. We select the top 20 candidates of each query to form question pairs and represent each question as a mean vector with word embeddings of all words in question. We then calculate *cosine score* between mean vectors of each query and candidate pair. The average of *cosine score* of mean vectors and *relevance score* of BM25 is treated as the final *similarity score* of each question pair, which ranges from 0 to 1. We assume that question pairs with similarity score in a high threshold interval $[0.9, 1)$ are the positive samples and those in a low threshold interval $[0, 0.6]$ as negative samples. Then we randomly sample the positive and negative samples with an approximate ratio of 1:3. Finally, we invite five domain experts of *dxy.com* to validate and remark the auto-tagged labels of samples. We name this dataset as ***Dataset2***.

3.4 Pre-training

Word Embedding. Word Embedding is popular in almost NLP tasks since the Word2vec[5] has been proposed by Google. The distributed representation of words learned by the neural networks would capture the semantic and syntactic information from unlabeled corpus. So that we preprocess the crawled medical QA data as training corpus and take CBOW model in Word2vec to train word embedding, and the dimensionality of word vectors is set to $d_e = 100$.

Transfer Learning Method. The pre-trained word vectors are usually used as the initialized weights of embedding layer in neural networks. It is commonly regarded as a transfer learning approach, which performs better than randomly initialization. Besides to the pre-trained word vectors, we can also transfer weights from pre-trained question encoders of other models to initialize corresponding layers in Siamese neural networks.

Since the process of face feature extraction and comparison of face recognization problem is similar to sentence matching task, in this paper, we adopt a multi-class classification neural network to pre-train encoders mentioned in Sect. 3.2. The encoder of multi-class classification model would help to learn

[5] https://code.google.archive/p/word2vec/.

more generalized representation among different expressions of similar questions. We take ***Dataset1*** that medical questions are categorized into different groups as training corpus for multi-class question classification, which is as same as the process in face recognition models [17]. The extracted features of pre-trained encoder will be used as question representations for similarity measurement.

The classification models would capture reasonable features of most samples, but usually confuse samples near group margins. The reason is that the classification constraint is not able to distinguish the close samples in different categories. The loss function of most multi-class classification tasks is softmax categorical cross entropy. If we set Q as input of encoder, and the cross-entropy loss of general multi-class classification can be described formally as follow,

$$z = Encoder(Q) \tag{4}$$

$$\begin{aligned} f &= softmax(zW) \\ &= softmax((z \cdot w_1), (z \cdot w_2), , (z \cdot w_n)) \end{aligned} \tag{5}$$

$$L_{softmax} = -\sum_{t} log \frac{e^{(z \cdot w_t)}}{\sum_{i=1}^{n} e^{(z \cdot w_i)}} \tag{6}$$

where, W is the weight of final linear layer, $W = (w_1, w_2, ...w_n)$, $(z \cdot w_i)$ is the dot product between z and w_i, f is the target probability distribution of Q and t is the label of Q.

The margin softmax function perform better than softmax function for feature ranking problem. It has beed proved more effective in field of face recognization. There are many angular margin softmax functions, like A-softmax [13] and AM-softmax [22]. We take AM-softmax to improve classification performance in our work. [22] proposes the AM-softmax for learning large-margin features with small intra-class variation and large inter-class difference. In the design of this function, z and w_i are normalized with $l2$ normalization, and the dot product in Eq. 5 is transformed to cosine function. Then a positive number m is used as a margin to tighten the cosine score of f_t and a positive number s is taken as scale rate to the tightened score. Formally, the AM-softmax loss function is shown as follow,

$$L_{AMS} = -\sum_{t} log \frac{e^{s \cdot (cos\theta_t - m)}}{e^{s \cdot (cos\theta_t - m)} + \sum_{i \neq t} e^{s \cdot (cos\theta_i)}} \tag{7}$$

where $cos\theta_i$ is the cosine score of z and w_i. In this paper, we choose the best hyper-parameters $m = 0.35$ and $s = 20$ in experiments. We attempt the Bi-LSTM and the *Self-Attention encoder* of TM-Transformer as the basic encoder in multi-class classification model respectively. Figure 3 is the pre-training model with Self-Attention encoder.

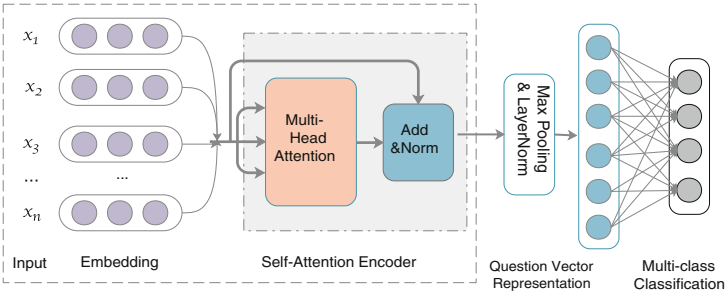


Fig. 3. The architecture of pre-training model with Self-Attention Encoder.

4 Experiments

In this section, we evaluate the proposed model architecture and the transfer learning strategy on large scale QA dataset in the following aspects: (1) We compare the results of Siamese Bi-LSTM and STMT model on the *Dataset2* to validate the effectiveness of our proposed architecture. (2) We pre-train the encoders with the transfer method declared in Sect. 3.4. Then we use pre-trained encoders to initialize corresponding layers of Siamese Bi-LSTM and STMT models and fine-tune model weights on finely labeled dataset.

4.1 Dataset

As mentioned in Sect. 3.3, the *Dataset1* which contains 8,317 similar question groups is used for pre-training encoders. The *Dataset2* that consists of about 300 thousand question pairs is applied to train and fine-tune the Bi-LSTM and TM-Transformer models. Table 2 lists the statistics of the datasets that we used in our experiments.

4.2 Experiment Setup

We randomly choose 5000 question groups from the *Dataset1* as training data for pre-training. In other words, we pre-train a classifier with 5000 target classes. We split the *Dataset2* into two subsets, we randomly select 60% question pairs as train set, 20% as dev set and the rest as test set. For all the experiments, we train the model on train set and tune parameters on dev set and pick the parameters which works the best on dev set. Finally, we train models with all the data in train set and dev set by fixing the best hyper-parameters. We calculate the Precision, Recall and F1 score for evaluating similarity prediction ability and normalized discounted cumulative gain (NDCG@3) as ranking score of these models.

We set the hidden size as 100 in Bi-LSTM architecture. We set head $h = 10$ of the multi-head attention in all the sub-layers of TM-Transformer. In the pre-training stage, we use the same hyper-parameters for above encoders in multi-class classification models. We initialize the embedding layer of all the neural

networks with the pre-trained word embedding, and the other parameters are initialized with Xavier initialization. Once the encoders have been pre-trained, we take weights of encoders to initialize corresponding layers in Siamese neural networks. We take ADAM as optimizer for all neural networks and the initial learning rate is 0.01, which decay along with the training epochs. We train the models for 100 epochs on a batch size of 64.

5 Discussion

5.1 The Effectiveness of TM-Transformer

In this part, we discuss about the effectiveness of our proposed architecture on medical question matching task. For comparison, we use the following four methods as baselines: (1) The logistic regression (LR) model takes the average of all the word vectors of a question as its representation, namely LR-vec. (2) The logistic regression model takes the TFIDF based bag-of-words as question representation, namely LR-tfidf. (3) The Siamese Bi-LSTM model utilize Bi-LSTM as sentence encoder. (4) The traditional Transformer model ($N = 6$ layers) for translation task declared in [19].

The results of these models are shown in Table 4. From the results, we can find that Siamese TM-Transformer model performs better on both classification and ranking metrics than all baseline models. We observe that the F1 score of STMT model exceeds about 2.61% than traditional Transformer, while NDCG score gets 0.89% improvement. It means that the proposed architecture can capture more semantical similarity information between question pairs. Since two attention based sub-layers of TM-Transformer focus on learning question representation and question interaction respectively. The simple sentence representation methods in SVM-vec and SVM-tfidf can not capture question context information and the interactive information between question pairs.

To validate the effect of these sub-layers, we train the TM-Transformer by removing the first sub-layer or the second one and the number of TM-Transformer layers $N = 6$. As shown in Table 4, either of two sub-layers has good performance, and the combination of them achieve a great improvement. In other words, stacked

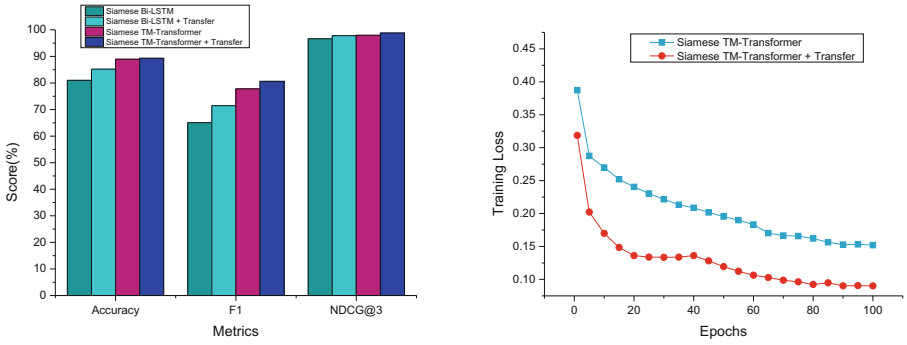
Table 4. Performance of Siamese neural networks on ***Dataset2***

Model	Accuracy	Precision	Recall	F1	NDCG@3
LR-tfidf	78.02	61.68	36.14	45.58	93.13
LR-vec	74.68	62.92	55.62	59.05	93.43
Siamese Bi-LSTM	80.98	60.55	73.57	65.05	96.64
Transformer [19]	85.64	65.92	**90.25**	75.19	97.07
STMT (only Self-Attention Encoder)	82.27	60.91	85.63	70.00	97.33
STMT (only Inter-Attention Encoder)	84.88	70.22	73.46	71.80	97.20
STMT	**88.93**	**76.77**	81.36	**77.80**	**97.96**

multi-head attention layers would help us to capture more semantic information and enhance the fitting and generalization ability of model.

5.2 The Effectiveness of Transfer Learning Method

In this part, we verify the effectiveness of transfer learning method we declared for model pre-training. The baseline are Siamese Bi-LSTM and STMT models, which we initialize model weights with Xavier initialization method in above experiments. We share the same encoder layers between baseline models with the multi-class classification model. And then we take weights of the pre-trained encoders as initialization of baseline models and fine-tune model weights on *Dataset2*. The result is shown in Table 5. There are great improvements of all the basic models with the transfer strategy. We can observe that the F1 score of STMT has a 2.81% increase while Bi-LSTM gets a improvement of 6.41%. For the ranking metrics, the transfer learning approach gives a little bit improvement over baseline, even though the baseline models have achieved good ranking scores already. The NDCG score of Bi-LSTM has a raise of 0.43%, as for TM-Transformer, the improvement is 0.35%. To further verify the effect of transfer strategy, we compare the training loss of TM-Transformer and TM-Transformer with transfer strategy. As shown in Fig. 4(b), the transfer strategy accelerates the model training and achieves better convergence. From the results in Fig. 4(a), it shows that the transfer learning method would increase the performance of different models more or less on question matching task.



(a) Effectiveness (b) Coveragence

Fig. 4. Comparison between plain Siamese neural network and Siamese neural network with transfer strategy.

We also compare the performance of our proposed method with other existing method on our datasets, such as ESIM in [6,11] and BiMPM model in [23]. As shown in Table 5, we find that our STMT model with transfer strategy gets a 6.55% F1 score higher than BiMPM model and 8.85% higher than ESIM

model. We can conclude that our proposed method can relieve question matching problem in medical domain, which perform better than the state-of-art model on other datasets. In the process of generating *Dataset2* based on the crawled medical QA corpus, we can find similar candidates for almost all queries. We also restrict the ratio of randomly selected negative samples. That is why the ranking scores of each model are relatively higher than classification metrics. At the pre-training stage, the feature extraction of multi-class classification is related to the feature ranking of sentence matching but they are not equivalent to each other. In other words, the features extracted by classifier may not work well at similarity ranking stage. So good loss functions, like AM-softmax or even better ones, are really helpful to relieve the gap between feature extraction and feature ranking.

Table 5. Performance of transfer learning method and other existing models

Model	Transfer	Accuracy	Precision	Recall	F1	NDCG@3
Siamese Bi-LSTM	No	80.98	60.55	73.57	65.05	96.64
	Yes	85.20	70.22	72.75	71.46	97.07
STMT	No	88.93	**76.77**	81.36	77.80	97.96
	Yes	**89.30**	74.91	87.25	**80.61**	**98.31**
ESIM [6]	No	85.88	73.13	70.45	71.76	97.18
BiMPM [23]	No	83.70	61.49	**96.33**	74.06	98.09

6 Conclusion

In this study, we investigate a novel Siamese TM-Transformer Neural Network (STMT) for similar health question retrieval in Chinese. Our method improves internal structure of Transformer for question matching task, which overcomes the lack of interactivity of Siamese neural network and the issue of diversity of medical expressions. Besides, we explore to use a transfer strategy to further enhance model performance. For the lack of medical QA datasets in Chinese, we declare two different data generation methods. Experiment results on large scale real world datasets have validates the performance of our method. In general, the proposed model and transfer strategy can also be used to solve text matching tasks in other domain.

Acknowledgment. This work is supported by the National Key Research and Development Program of China under Grant No. 2016YFB1000904.

References

1. Aditya, T.: Siamese recurrent architectures for learning sentence similarity. In: Thirtieth AAAI Conference on Artificial Intelligence, pp. 2786–2792 (2016)
2. Baziotis, C., Pelekis, N., Doulkeridis, C.: Datastories at semeval-2017 task 6: Siamese LSTM with attention for humorous text comparison. In: Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, pp. 390–395 (2017)
3. Borui, Y., Guangyu, F., Anqi, C., Ming, L.: Learning question similarity with recurrent neural networks. In: IEEE International Conference on Big Knowledge, pp. 111–118 (2017)
4. Cai, H., Yan, C., Yin, A., Zhao, X.: Question recommendation in medical community-based question answering. In: Liu, D., Xie, S., Li, Y., Zhao, D., El-Alfy, E.-S.M. (eds.) ICONIP 2017. LNCS, vol. 10638, pp. 228–236. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-70139-4_23
5. Cao, X., Cong, G., Cui, B., Jensen, C.S., Zhang, C.: The use of categorization information in language models for question retrieval. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, pp. 265–274 (2009)
6. Chen, Q., Zhu, X., Ling, Z., Wei, S., Jiang, H., Inkpen, D.: Enhanced LSTM for natural language inference. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL, pp. 1657–1668 (2017)
7. Das, A., Yenala, H., Chinnakotla, M.K., Shrivastava, M.: Together we stand: Siamese networks for similar question retrieval. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL (2016)
8. Eyecioglu, A., Keller, B.: Twitter paraphrase identification with simple overlap features and SVMs. In: Proceedings of the 9th International Workshop on Semantic Evaluation, pp. 64–69 (2015)
9. Jeon, J., Croft, W.B., Lee, J.H.: Finding similar questions in large question and answer archives. In: Proceedings of the 2005 ACM CIKM International Conference on Information and Knowledge Management, pp. 84–90 (2005)
10. Ji, Z., Xu, F., Wang, B., He, B.: Question-answer topic model for question retrieval in community question answering. In: 21st ACM International Conference on Information and Knowledge Management, CIKM 2012, pp. 2471–2474 (2012)
11. Lan, W., Xu, W.: Neural network models for paraphrase identification, semantic textual similarity, natural language inference, and question answering. In: Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, pp. 3890–3902 (2018)
12. Li, Y., et al.: Finding similar medical questions from question answering websites. CoRR abs/1810.05983 (2018)
13. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: deep hypersphere embedding for face recognition. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, pp. 6738–6746 (2017)
14. Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. SIGIR Forum **51**(2), 202–208 (2017)
15. Qiu, X., Huang, X.: Convolutional neural tensor network architecture for community-based question answering. In: Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI, pp. 1305–1311 (2015)
16. Robertson, S.E., Jones, K.S.: Relevance Weighting of Search Terms. Taylor Graham Publishing (1988)

17. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: closing the gap to human-level performance in face verification. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, pp. 1701–1708 (2014)
18. Tang, G., Ni, Y., Xie, G., Fan, X., Shi, Y.: A deep learning-based method for similar patient question retrieval in chinese. In: MEDINFO 2017: Precision Healthcare through Informatics - Proceedings of the 16th World Congress on Medical and Health Informatics, pp. 604–608 (2017)
19. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems, pp. 6000–6010 (2017)
20. Vo, N.P.A., Magnolini, S., Popescu, O.: FBK-HLT: an effective system for paraphrase identification and semantic similarity in Twitter. In: Proceedings of the 9th International Workshop on Semantic Evaluation, pp. 29–33 (2015)
21. Wan, S., Lan, Y., Guo, J., Xu, J., Pang, L., Cheng, X.: A deep architecture for semantic matching with multiple positional sentence representations. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, pp. 2835–2841 (2016)
22. Wang, F., Cheng, J., Liu, W., Liu, H.: Additive margin softmax for face verification. IEEE Signal Process. Lett. **25**(7), 926–930 (2018)
23. Wang, Z., Hamza, W., Florian, R.: Bilateral multi-perspective matching for natural language sentences. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, 19–25 2017, pp. 4144–4150 (2017)
24. Wang, Z., Mi, H., Ittycheriah, A.: Sentence similarity learning by lexical decomposition and composition. arXiv:1602.07019 (2016)
25. Xue, X., Jeon, J., Croft, W.B.: Retrieval models for question and answer archives. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, pp. 475–482 (2008)
26. Zhang, K., Wu, W., Wu, H., Li, Z., Zhou, M.: Question retrieval with high quality answers in community question answering. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, pp. 371–380 (2014)
27. Zhou, G., Zhou, Y., He, T., Wu, W.: Learning semantic representation with neural networks for community question answering retrieval. Knowl.-Based Syst. **93**, 75–83 (2016)