Question Condensing Networks for Answer Selection in Community Question Answering

Wei Wu[1], Xu Sun[1], Houfeng Wang[1,2]

[1]MOE Key Lab of Computational Linguistics, Peking University, Beijing, 100871, China
[2]Collaborative Innovation Center for Language Ability, Xuzhou, Jiangsu, 221009, China
{wu.wei, xusun, wanghf}@pku.edu.cn

Abstract

Answer selection is an important subtask of community question answering (CQA). In a real-world CQA forum, a question is often represented as two parts: a subject that summarizes the main points of the question, and a body that elaborates on the subject in detail. Previous researches on answer selection usually ignored the difference between these two parts and concatenated them as the question representation. In this paper, we propose the Question Condensing Networks (QCN) to make use of the subject-body relationship of community questions. In this model, the question subject is the primary part of the question representation, and the question body information is aggregated based on similarity and disparity with the question subject. Experimental results show that QCN outperforms all existing models on two CQA datasets.

1 Introduction

Community question answering (CQA) has seen a spectacular increase in popularity in recent years. With the advent of sites like Stack Overflow[1] and Quora[2], more and more people can freely ask any question and expect a variety of answers. With the influx of new questions and the varied quality of provided answers, it is very time-consuming for a user to inspect them all. Therefore, developing automated tools to identify good answers for a question is of practical importance.

A typical example for CQA is shown in Table 1. In this example, Answer 1 is a good answer, because it provides helpful information, e.g., "*check it to the traffic dept*". Although Answer 2 is relevant to the question, it does not contain any useful information so that it should be regarded as a bad answer.

From this example, we can observe two characteristics of CQA that ordinary QA does not possess. First, a question includes both a subject that gives a brief summary of the question and a body that describes the question in detail. The questioners usually convey their main concern and key information in the question subject. Then, they provide more extensive details about the subject, seek help, or express gratitude in the question body. Second, the problem of redundancy and noise is prevalent in CQA (Zhang et al., 2017). Both questions and answers contain auxiliary sentences that do not provide meaningful information.

Previous researches (Tran et al., 2015; Joty et al., 2016) usually treat each word equally in the question and answer representation. However, due to the redundancy and noise problem, only part of text from questions and answers is useful to determine the answer quality. To make things worse, they ignored the difference between question subject and body, and simply concatenated them as the question representation. Due to the subject-body relationship described above, this simple concatenation can aggravate the redundancy problem in the question. In this paper, we propose the Question Condensing Networks (QCN) to address these problems.

In order to utilize the subject-body relationship in community questions, we propose to treat the question subject as the primary part of the question, and aggregate the question body information based on similarity and disparity with the question subject. The similarity part corresponds to the information that exists in both question subject and body, and the disparity part corresponds to the additional information provided by the ques-

[1]https://stackoverflow.com/
[2]https://www.quora.com/

Question Subject	Checking the history of the car.
Question body	How can one check the history of the car like maintenance, accident or service history. In every advertisement of the car, people used to write “Accident Free", but in most cases, car have at least one or two accident, which is not easily detectable through Car Inspection Company. Share your opinion in this regard.
Answer1	Depends on the owner of the car.. if she/he reported the accident/s i believe u can check it to the traffic dept.. but some owners are not doing that especially if its only a small accident.. try ur luck and go to the traffic dept..
Answer2	How about those who claim a low mileage by tampering with the car fuse box? In my sense if you’re not able to detect traces of an accident then it is probably not worth mentioning... For best results buy a new car :)

Table 1: An example question and its related answers in CQA. The text is shown in its original form, which may contain errors in typing.

tion body. Both information can be important for question representation. In our model, they are processed separately and the results are combined to form the final question representation.

In order to reduce the impact of redundancy and noise in both questions and answers, we propose to align the question-answer pairs using the multi-dimensional attention mechanism. Different from previous attention mechanisms that compute a scalar score for each token pair, multi-dimensional attention, first proposed in Shen et al. (2018), computes one attention score for each dimension of the token embedding. Therefore, it can select the features that can best describe the word’s specific meaning in the given context. Therefore, we can learn the interaction between questions and answers more accurately.

The main contributions of our work can be summarized as follows:

- We propose to treat the question subject and the question body separately in community question answering. We treat the question subject as the primary part of the question, and aggregate the question body information based on similarity and disparity with the question subject.
- We introduce a new method that uses the multi-dimensional attention mechanism to align question-answer pair. With this attention mechanism, the interaction between questions and answers can be learned more accurately.
- Our proposed Question Condensing Networks (QCN) achieves the state-of-the-art performance on two SemEval CQA datasets, outperforming all exisiting SOTA models by a large margin, which demonstrates the effectiveness of our model.

2 Task Description

A community question answering consists of four parts, which can be formally defined as a tuple of four elements (S, B, C, y). $S = [\boldsymbol{s}^1, \boldsymbol{s}^2, ..., \boldsymbol{s}^l]$ denotes the subject of a question whose length is l, where each $\boldsymbol{s}^i$ is a one-hot vector whose dimension equals the size of the vocabulary. Similarly, $B = [\boldsymbol{b}^1, \boldsymbol{b}^2, ..., \boldsymbol{b}^m]$ denotes the body of a question whose length is m. $C = [\boldsymbol{c}^1, \boldsymbol{c}^2, ..., \boldsymbol{c}^n]$ denotes an answer corresponding to that question whose length is n. $y \in \mathcal{Y}$ is the label representing the degree to which it can answer that question. $\mathcal{Y} = \{\textit{Good}, \textit{PotentiallyUseful}, \textit{Bad}\}$ where *Good* indicates the answer can answer that question well, *PotentiallyUseful* indicates the answer is potentially useful to the user, and *Bad* indicates the answer is just bad or useless. Given $\{S, B, C\}$, the task of CQA is to assign a label to each answer based on the conditional probability $Pr(y|S, B, C)$.

3 Proposed Model

In this paper, we propose Question Condensing Networks (QCN) which is composed of the following modules. The overall architecture of our model is illustrated in Figure 1.

3.1 Word-Level Embedding

Word-level embeddings are composed of two components: GloVe (Pennington et al., 2014)

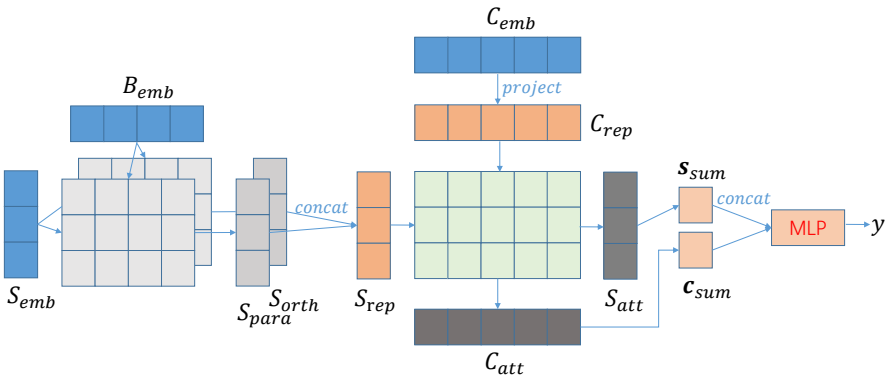


Figure 1: Architecture for Question Condensing Network (QCN). Each block represents a vector.

word vectors trained on the domain-specific unannotated corpus provided by the task [3], and convolutional neural network-based character embeddings which are similar to (Kim et al., 2016). Web text in CQA forums differs largely from normalized text in terms of spelling and grammar, so specifically trained GloVe vectors can model word interactions more precisely. Character embedding has proven to be very useful for out-of-vocabulary (OOV) words, so it is especially suitable for noisy web text in CQA.

We concatenate these two embedding vectors for every word to generate word-level embeddings $\boldsymbol{S}_{emb} \in \mathbb{R}^{d\times l}, \boldsymbol{B}_{emb} \in \mathbb{R}^{d\times m}, \boldsymbol{C}_{emb} \in \mathbb{R}^{d\times n}$, where d is the word-level embedding size.

3.2 Question Condensing

In this section, we condense the question representation using subject-body relationship. In most cases, the question subject can be seen as a summary containing key points of the question, the question body is relatively lengthy in that it needs to explain the key points and add more details about the posted question. We propose to cheat the question subject as the primary part of the question representation, and aggregate question body information from two perspectives: similarity and disparity with the question subject. To achieve this goal, we use an orthogonal decomposition strategy, which is first proposed by Wang et al. (2016), to decompose each question body embedding into a parallel component and an orthogonal component based on every question subject embedding:

$$\boldsymbol{b}^{i,j}_{para} = \frac{\boldsymbol{b}^{j}_{emb} \cdot \boldsymbol{s}^{i}_{emb}}{\boldsymbol{s}^{i}_{emb} \cdot \boldsymbol{s}^{i}_{emb}} \boldsymbol{s}^{i}_{emb} \quad (1)$$

$$\boldsymbol{b}^{i,j}_{orth} = \boldsymbol{b}^{j}_{emb} - \boldsymbol{b}^{i,j}_{para} \quad (2)$$

All vectors in the above equations are of length d. Next we describe the process of aggregating the question body information based on the parallel component in detail. The same process can be applied to the orthogonal component, so at the end of the fusion gate we can obtain S_{orth} and S_{orth} respectively.

The decomposed components are passed through a fully connected layer to compute the multi-dimensional attention weights. Here we use the scaled tanh activation, which is similar to Shen et al. (2018), to prevent large difference among scores while it still has a range large enough for output:

$$\boldsymbol{a}^{i,j}_{para} = c \cdot \tanh\left(\left[W_{p1}\boldsymbol{b}^{i,j}_{para} + \boldsymbol{b}_{p1}\right]/c\right) \quad (3)$$

where $W_{p1} \in \mathbb{R}^{d\times d}$ and $\boldsymbol{b}_{p1} \in \mathbb{R}^{d}$ are parameters to be learned, and c is a hyper-parameter to be tuned.

The obtained word-level alignment tensor $\boldsymbol{A}_{para} \in \mathbb{R}^{d\times l\times m}$ is then normalized along the third dimension to produce the attention weights over the question body for each word in the question subject. The output of this attention mechanism is a weighted sum of the question body embeddings for each word in the question subject:

$$\boldsymbol{w}^{i,j}_{para} = \frac{\exp\left(\boldsymbol{a}^{i,j}_{para}\right)}{\sum_{j=1}^{m}\exp\left(\boldsymbol{a}^{i,j}_{para}\right)} \quad (4)$$

$$\boldsymbol{s}^{i}_{ap} = \sum_{j=1}^{m} \boldsymbol{w}^{i,j}_{para} \odot \boldsymbol{b}^{j}_{emb} \quad (5)$$

[3] http://alt.qcri.org/semeval2015/task3/index.php?id=data-and-tools

where $\odot$ means point-wise product. This multi-dimensional attention mechanism has the advantage of selecting features of a word that can best describe the word's specific meaning in the given context. In order to determine the importance between the original word in the question subject and the aggregated information from the question body with respect to this word, a fusion gate is utilized to combine these two representations:

$$F_{para} = \sigma\left(W_{p2}S_{emb} + W_{p3}S_{ap} + \boldsymbol{b}_{p2}\right) \quad (6)$$

$$S_{para} = F_{para} \odot S_{emb} + (1 - F_{para}) \odot S_{ap} \quad (7)$$

where $W_{p2}, W_{p3} \in \mathbb{R}^{d\times d}$, and $\boldsymbol{b}_{p2} \in \mathbb{R}^{d}$ are learnable parameters of the fusion gate, and $F_{para}, S_{emb}, S_{ap}, S_{para} \in \mathbb{R}^{d\times l}$. The final question representation $S_{rep} \in \mathbb{R}^{2d\times l}$ is obtained by concatenating S_{para} and S_{orth} along the first dimension.

3.3 Answer Preprocessing

This module has two purposes. First, we try to map each answer word from embedding space $\boldsymbol{C}_{emb} \in \mathbb{R}^{d\times n}$ to the same interaction space $\boldsymbol{C}_{rep} \in \mathbb{R}^{2d\times n}$ as the question. Second, similar to Wang and Jiang (2017), a gate is utilized to control the importance of different answer words in determining the question-answer relation:

$$\begin{aligned} C_{rep} =& \sigma\left(W_{c1}C_{emb} + \boldsymbol{b}_{c1}\right) \odot \\ & \tanh\left(W_{c2}C_{emb} + \boldsymbol{b}_{c2}\right) \end{aligned} \quad (8)$$

where $W_{c1}, W_{c2} \in \mathbb{R}^{d\times 2d}$ and $\boldsymbol{b}_{c1}, \boldsymbol{b}_{c2} \in \mathbb{R}^{2d}$ are parameters to be learned.

3.4 Question Answer Alignment

We apply the multi-dimensional attention mechanism to the question and answer representation S_{rep} and C_{rep} to obtain word-level alignment tensor $\boldsymbol{A}_{align} \in \mathbb{R}^{2d\times l\times n}$. Similar to the multi-dimensional attention mechanism described above, we can compute attention weights and weighted sum for both the question representation and the answer representation :

$$\tilde{\boldsymbol{a}}^{i,j}_{align} = W_{a1}\boldsymbol{s}^{i}_{rep} + W_{a2}\boldsymbol{c}^{j}_{rep} + \boldsymbol{b}_{a} \quad (9)$$

$$\boldsymbol{a}^{i,j}_{align} = c \cdot \tanh\left(\tilde{\boldsymbol{a}}^{i,j}_{align}/c\right) \quad (10)$$

$$\boldsymbol{s}^{i}_{ai} = \sum_{j=1}^{n} \frac{\exp\left(\boldsymbol{a}^{i,j}_{align}\right)}{\sum_{j=1}^{n}\exp\left(\boldsymbol{a}^{i,j}_{align}\right)} \odot \boldsymbol{c}^{j}_{rep} \quad (11)$$

$$\boldsymbol{c}^{j}_{ai} = \sum_{i=1}^{l} \frac{\exp\left(\boldsymbol{a}^{i,j}_{align}\right)}{\sum_{i=1}^{l}\exp\left(\boldsymbol{a}^{i,j}_{align}\right)} \odot \boldsymbol{s}^{i}_{rep} \quad (12)$$

where $W_{a1}, W_{a2} \in \mathbb{R}^{2d\times 2d}$ and $\boldsymbol{b}_{a} \in \mathbb{R}^{2d}$ are parameters to be learned. To attenuate the effect of incorrect attendance, input and output of this attention mechanism are concatenated and fed to the subsequent layer. Finally, we obtain the question and answer representation $S_{att} \in \mathbb{R}^{4d\times l} = [S_{rep}; S_{ai}]$, $C_{att} \in \mathbb{R}^{4d\times n} = [C_{rep}; C_{ai}]$.

3.5 Interaction Summarization

In this layer, the multi-dimensional self-attention mechanism is employed to summarize two sequences of vectors (S_{att} and C_{att}) into two fixed-length vectors $\boldsymbol{s}_{sum} \in \mathbb{R}^{4d}$ and $\boldsymbol{c}_{sum} \in \mathbb{R}^{4d}$.

$$A_s = W_{s2}\tanh\left(W_{s1}S_{att} + \boldsymbol{b}_{s1}\right) + \boldsymbol{b}_{s2} \quad (13)$$

$$\boldsymbol{s}_{sum} = \sum_{i=1}^{n} \frac{\exp\left(\boldsymbol{a}^{i}_{s}\right)}{\sum_{i=1}^{n}\exp\left(\boldsymbol{a}^{i}_{s}\right)} \odot \boldsymbol{s}^{i}_{att} \quad (14)$$

where $W_{s1}, W_{s2} \in \mathbb{R}^{4d\times 4d}$ and $\boldsymbol{b}_{s1}, \boldsymbol{b}_{s2} \in \mathbb{R}^{4d}$ are parameters to be learned. The same process can be applied to C_{att} and obtain $\boldsymbol{c}_{sum}$.

3.6 Prediction

In this component, $\boldsymbol{s}_{sum}$ and $\boldsymbol{c}_{sum}$ are concatenated and fed into a two-layer feed-forward neural network. At the end of the last layer, the $softmax$ function is applied to obtain the conditional probability distribution $Pr(y|S, B, C)$.

4 Experimental Setup

4.1 Datasets

We use two community question answering datasets from SemEval (Nakov et al., 2015, 2017) to evaluate our model. The statistics of these datasets are listed in Table 2. The corpora contain data from the QatarLiving forum [4], and are publicly available on the task website. Each dataset

[4] http://www.qatarliving.com/forum

Statistics	SemEval 2015			SemEval 2017		
	Train	Dev	Test	Train	Dev	Test
Number of questions	2376	266	300	5124	327	293
Number of answers	15013	1447	1793	38638	3270	2930
Average length of subject	6.36	6.08	6.24	6.38	6.16	5.76
Average length of body	39.26	39.47	39.53	43.01	47.98	54.06
Average length of answer	35.82	33.90	37.33	37.67	37.30	39.50

Table 2: Statistics of two CQA datasets. We can see from the statistics that the question body is much lengthier than the question subject. Thus, it is necessary to condense the question representation.

consists of questions and a list of answers for each question, and each question consists of a short title and a more detailed description. There are also some metadata associated with them, e.g., user ID, date of posting, and the question category. We do not use the metadata because they failed to boost performance in our model. Since the SemEval 2017 dataset is an updated version of SemEval 2016 [5], and shares the same evaluation metrics with SemEval 2016, we choose to use the SemEval 2017 dataset for evaluation.

4.2 Evaluation Metrics

In order to facilitate comparison, we adopt the evaluation metrics used in the official task or prior work. For the SemEval 2015 dataset, the official scores are macro-averaged F1 and accuracy over three categories. However, many recent researches (Barrón-Cedeño et al., 2015; Joty et al., 2015, 2016) switched to a binary classification setting, i.e., identifying *Good* vs. *Bad* answers. Because binary classification is much closer to a real-world CQA application. Besides, the *PotentiallyUseful* class is both the smallest and the noisiest class, making it the hardest to predict. To make it worse, its impact is magnified by the macro-averaged F1. Therefore, we adopt the F1 score and accuracy on two categories for evaluation.

SemEval 2017 regards answer selection as a ranking task, which is closer to the application scenario. As a result, mean average precision (MAP) is used as an evaluation measure. For a perfect ranking, a system has to place all *Good* answers above the *PotentiallyUseful* and *Bad* answers. The latter two are not actually distinguished and are considered *Bad* in terms of evaluation. Additionally, standard classification measures like accuracy and F1 score are also reported.

[5]The SemEval 2017 dataset provides all the data from 2016 for training , and fresh data for testing, but it does not include a development set. Following previous work (Filice et al., 2017), we use the 2016 official test set as the development set.

4.3 Implementation Details

We use the tokenizer from NLTK (Bird, 2006) to preprocess each sentence. All word embeddings in the sentence encoder layer are initialized with the 300-dimensional GloVe (Pennington et al., 2014) word vectors trained on the domain-specific unannotated corpus, and embeddings for out-of-vocabulary words are set to zero. We use the Adam Optimizer (Kingma and Ba, 2014) for optimization with a first momentum coefficient of 0.9 and a second momentum coefficient of 0.999. We perform a small grid search over combinations of initial learning rate [1×10^{-6}, 3×10^{-6}, 1×10^{-5}], L2 regularization parameter [1×10^{-7}, 3×10^{-7}, 1×10^{-6}], and batch size [8, 16, 32]. We take the best configuration based on performance on the development set, and only evaluate that configuration on the test set. In order to mitigate the class imbalance problem, median frequency balancing Eigen and Fergus (2015) is used to reweight each class in the cross-entropy loss. Therefore, the rarer a class is in the training set, the larger weight it will get in the cross entropy loss. Early stopping is applied to mitigate the problem of overfitting. For the SemEval 2017 dataset, the conditional probability over the *Good* class is used to rank all the candidate answers.

5 Experimental Results

In this section, we evaluate our QCN model on two community question answering datasets from SemEval shared tasks.

5.1 SemEval 2015 Results

Table 3 compares our model with the following baselines:

Methods	F1	Acc
(1) JAIST	78.96	79.10
(2) HITSZ-ICRC	76.52	76.11
(3) Graph-cut	80.55	79.80
(4) FCCRF	81.50	80.50
(5) BGMN	77.23	78.40
(6) CNN-LSTM-CRF	82.22	82.24
(7) QCN	**83.91**	**85.65**

Table 3: Comparisons on the SemEval 2015 dataset.

- **JAIST** (Tran et al., 2015): It used an SVM classifier to incorporate various kinds of features , including topic model based features and word vector representations.
- **HITSZ-ICRC** (Hou et al., 2015): It proposed ensemble learning and hierarchical classification method to classify answers.
- **Graph-cut** (Joty et al., 2015): It modeled the relationship between pairs of answers at any distance in the same question thread, based on the idea that similar answers should have similar labels.
- **FCCRF** (Joty et al., 2016): It used locally learned classifiers to predict the label for each individual node, and applied fully connected CRF to make global inference.
- **CNN-LSTM-CRF** (Xiang et al., 2016): The question and its answers are linearly connected in a sequence and encoded by CNN. An attention-based LSTM with a CRF layer is then applied on the encoded sequence.
- **BGMN** (Wu et al., 2017b): It used the memory mechanism to iteratively aggregate more relevant information which is useful to identify the relationship between questions and answers.

Baselines include top systems from SemEval 2015 (1, 2), systems relying on thread level information to make global inference (3, 4), and neural network based systems (5, 6). We observe that our proposed QCN can achieve the state-of-the-art performance on this dataset, outperforming previous best model (6) by 1.7% in terms of F1 and 3.4% in terms of accuracy.

Methods	MAP	F1	Acc
(1) KeLP	88.43	69.87	73.89
(2) Beihang-MSRA	88.24	68.40	51.98
(3) ECNU	86.72	77.67	78.43
(4) LSTM	86.32	74.41	75.69
(5) LSTM-subject-body	87.11	74.50	77.28
(6) QCN	**88.51**	**78.11**	**80.71**

Table 4: Comparisons on the SemEval 2017 dataset.

Notably, Systems (1, 2, 3, 4) have heavy feature engineering, while QCN only uses automatically-learned feature vectors, demonstrating that our QCN model is concise as well as effective. Furthermore, our model can outperform systems relying on thread level information to make global inference (3, 4), showing that modeling interaction between the question-answer pair is useful enough for answer selection task. Finally, neural network based systems (5, 6) used attention mechanism in sentence representation but ignored the subject-body relationship in community questions. QCN can outperform them by a large margin, showing that condensing question representation helps in the answer selection task.

5.2 SemEval 2017 Results

Table 4 compares our model with the following baselines:

- **KeLP** (Filice et al., 2017): It used syntactic tree kernels with relational links between questions and answers, together with some standard text similarity measures linearly combined with the tree kernel.
- **Beihang-MSRA** (Feng et al., 2017): It used gradient boosted regression trees to combine traditional NLP features and neural network-based matching features.
- **ECNU** (Wu et al., 2017a): It combined a supervised model using traditional features and a convolutional neural network to represent the question-answer pair.
- **LSTM**: It is a simple neural network based baseline that we implemented. In this model, the question subject and the question body are concatenated, and an LSTM is used to obtain the question and answer representation.

- **LSTM-subject-body**: It is another neural network based baseline that we implemented. LSTM is applied on the question subject and body respectively, and the results are concatenated to form question representation.

Baselines include top systems from the SemEval 2017 CQA task (1, 2, 3) and two neural network based baselines (4, 5) that we implemented. (5) can outperform (4), showing that treating question subject and body differently can indeed boot model performance. Comparing (6) with (5), we can draw the conclusion that orthogonal decomposition is more effective than simple concatenation, because it can flexibly aggregate related information from the question body with respect to the main subject. In the example listed in Table 1, attention heatmap of A_{orth} indicates that QCN can effectively find additional information like "*maintenance, accident or service history*", while (5) fails to do so.

QCN has a great advantage in terms of accuracy. We hypothesize that QCN focuses on modeling interaction between questions and answers, i.e., whether an answer can match the corresponding question. Many pieces of previous work focus on modeling relationship between answers in a question thread, i.e., which answer is more suitable in consideration of all other answers. As a consequence, their models have a greater advantage in ranking while QCN has a greater advantage in classification. Despite all this, QCN can still obtain better ranking performance.

5.3 Ablation Study

For thorough comparison, besides the preceding models, we implement nine extra baselines on the SemEval 2017 dataset to analyze the improvements contributed by each part of our QCN model:

- **w/o task-specific word embeddings** where word embeddings are initialized with the 300-dimensional GloVe word vectors trained on Wikipedia 2014 and Gigaword 5.

- **w/o character embeddings** where word-level embeddings are only composed of 600-dimensional GloVe word vectors trained on the domain-specific unannotated corpus.

- **subject-body alignment** where we use the same attention mechanism as Question Answer Alignment to obtain weighted sum of the question body for each question subject word, and then the result is concatenated with S_{emb} to obtain question representation S_{rep}.

- **subject-body concatenation** where we concatenate question subject and body text, and use the preprocessing step described in section 3.3 to obtain S_{rep}.

- **w/o multi-dimensional attention** where the multi-dimensional attention mechanism is replaced by vanilla attention in all modules, i.e., attention score for each token pair is a scalar instead of a vector.

- **subject only** where only question subject is used as question representation.

- **body only** where only question body is used as question representation.

- **similarity only** where the parallel component alone is used in subject-body interaction.

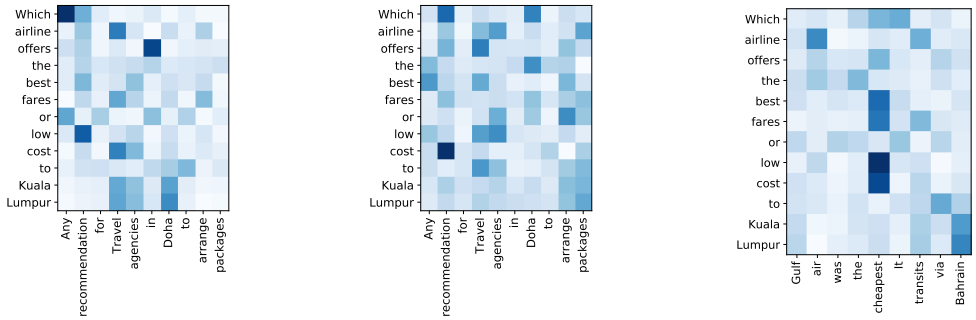
- **disparity only** where the orthogonal component alone is used in subject-body interaction.

Model	Acc
(1) w/o task-specific word embeddings	78.81
(2) w/o character embeddings	78.05
(3) subject-body alignment	77.38
(4) subject-body concatenation	76.06
(5) w/o multi-dimensional attention	78.33
(6) subject only	74.02
(7) body only	75.57
(8) similarity only	79.11
(9) disparity only	78.24
(10) QCN	**80.71**

Table 5: Ablation studies on the SemEval 2017 dataset.

The results are listed in Table 5. We can see that using task-specific embeddings and character embeddings both contribute to model performance. This is because CQA text is non-standard. There are quantities of informal language usage, such as abbreviations, typos, emoticons, and grammatical mistakes. Using task-specific embeddings and character embeddings can help to attenuate the OOV problem.

Using orthogonal decomposition (10) instead of subject-body alignment (3) can bring about significant performance gain. This is because not only



(a) A_{para} (b) A_{orth} (c) A_{align}

Figure 2: Attention probabilities in A_{para}, A_{orth} and A_{align}. In order to visualize the multi-dimensional attention vector, we use the $L2$ norm of the attenion vector for representation.

the similar part of the question body to the question subject is useful for the question representation, the disparity part can also provide additional information. In the example listed in Table 1, additional information like "*maintenance, accident or service history*" is also important to determine answer quality.

QCN outperforms (4) by a great margin, demonstrating that subject-body relationship in community questions helps to condense question representation. Therefore, QCN can identify the meaningful part of the question representation that helps to determine answer quality.

Using the multi-dimensional attention can further boost model performance, showing that the multi-dimensional attention can model the interaction between questions and answers more precisely.

Comparing QCN with (6) and (7), we can conclude that both the subject and the body are indispensable for question representation. (8) outperforms (9), demonstrating the parallel component is more useful in subject-body interaction.

6 Qualitative Study

To gain a closer view of what dependencies are captured in the subject-body pair and the question-answer pair, we visualize the attention probabilities A_{para}, A_{orth} and A_{align} by heatmap. A training example from SemEval 2015 is selected for illustration.

In Figure 2, we can draw the following conclusions. First, orthogonal decomposition helps to divide the labor of identifying similar parts in the parallel component and collecting related information in the question body in the orthogonal component. For instance, for the word "*Kuala*" in the question subject, its parallel alignment score focuses more on "*Doha*" and "*Travel*", while its orthogonal alignment score focuses on "*arrange*" and "*package*", which is the purpose of the travel and therefore is also indispensable for sentence representation. Second, semantically important words such as "*airline*" and "*fares*" dominate the attention weights, showing that our QCN model can effectively select words that are most representative for the meaning of the whole sentence. Lastly, words that are useful to determine answer quality stand out in the question-answer interaction matrix, demonstrating that question-answer relationship can be well modeled. For example, "*best*" and "*low*" are the words that are more important in the question-answer relationship, they are emphasized in the question-answer alignment matrix.

7 Related Work

One main task in community question answering is answer selection, i.e., to rate the answers according to their quality. The SemEval CQA tasks (Nakov et al., 2015, 2016, 2017) provide universal benchmark datasets for evaluating researches on this problem.

Earlier work of answer selection in CQA relied heavily on feature engineering, linguistic tools, and external resource. Nakov et al. (2016) investigated a wide range of feature types including similarity features, content features, thread level/meta features, and automatically generated features for SemEval CQA models. Tran et al. (2015) studied the use of topic model based features and word vector representation based features in the answer re-ranking task. Filice et al. (2016) designed various heuristic features and thread-based features

that can signal a good answer. Although achieving good performance, these methods rely heavily on feature engineering, which requires a large amount of manual work and domain expertise.

Since answer selection is inherently a ranking task, a few recent researches proposed to use local features to make global ranking decision. Barrón-Cedeño et al. (2015) was the first work that applies structured prediction model on CQA answer selection task. Joty et al. (2016) approached the task with a global inference process to exploit the information of all answers in the question-thread in the form of a fully connected graph.

To avoid feature engineering, many deep learning models have been proposed for answer selection. Among them, Zhang et al. (2017) proposed a novel interactive attention mechanism to address the problem of noise and redundancy prevalent in CQA. Tay et al. (2017) introduced temporal gates for sequence pairs so that questions and answers are aware of what each other is remembering or forgetting. Simple as their model are, they did not consider the relationship between question subject and body, which is useful for question condensing.

8 Conclusion and Future Work

We propose Question Condensing Networks (QCN), an attention-based model that can utilize the subject-body relationship in community questions to condense question representation. By orthogonal decomposition, the labor of identifying similar parts and collecting related information in the question body can be well divided in two different alignment matrices. To better capture the interaction between the subject-body pair and the question-answer pair, the multi-dimensional attention mechanism is adopted. Empirical results on two community question answering datasets in SemEval demonstrate the effectiveness of our model. In future work, we will try to incorporate more hand-crafted features in our model. Furthermore, since thread-level features have been explored in previous work (Barrón-Cedeño et al., 2015; Joty et al., 2015, 2016), we will verify their effectiveness in our architecture.

9 Acknowledgments

Our work is supported by National Natural Science Foundation of China under Grant No.61433015 and the National Key Research and Development Program of China under Grant No.2017YFB1002101. The corresponding authors of this paper are Houfeng Wang.

References

Alberto Barrón-Cedeño, Simone Filice, Giovanni Da San Martino, Shafiq R. Joty, Lluís Màrquez, Preslav Nakov, and Alessandro Moschitti. 2015. Thread-level information for comment classification in community question answering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers*. pages 687–693.

Steven Bird. 2006. NLTK: the natural language toolkit. In *ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006*.

David Eigen and Rob Fergus. 2015. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. pages 2650–2658.

Wenzheng Feng, Yu Wu, Wei Wu, Zhoujun Li, and Ming Zhou. 2017. Beihang-msra at semeval-2017 task 3: A ranking system with neural matching features for community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, Vancouver, Canada, August 3-4, 2017*. pages 280–286.

Simone Filice, Danilo Croce, Alessandro Moschitti, and Roberto Basili. 2016. Kelp at semeval-2016 task 3: Learning semantic relations between questions and answers. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*. pages 1116–1123.

Simone Filice, Giovanni Da San Martino, and Alessandro Moschitti. 2017. Kelp at semeval-2017 task 3: Learning pairwise patterns in community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, Vancouver, Canada, August 3-4, 2017*. pages 326–333.

Yongshuai Hou, Cong Tan, Xiaolong Wang, Yaoyun Zhang, Jun Xu, and Qingcai Chen. 2015. HITSZ-ICRC: exploiting classification approach for answer selection in community question answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015, Denver, Colorado, USA, June 4-5, 2015*. pages 196–202.

Shafiq R. Joty, Alberto Barrón-Cedeño, Giovanni Da San Martino, Simone Filice, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2015. Global thread-level inference for comment classification in community question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*. pages 573–578.

Shafiq R. Joty, Lluís Màrquez, and Preslav Nakov. 2016. Joint learning with global inference for comment classification in community question answering. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*. pages 703–713.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M. Rush. 2016. Character-aware neural language models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.*. pages 2741–2749.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR* abs/1412.6980.

Preslav Nakov, Doris Hoogeveen, Lluís Màrquez, Alessandro Moschitti, Hamdy Mubarak, Timothy Baldwin, and Karin Verspoor. 2017. Semeval-2017 task 3: Community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, Vancouver, Canada, August 3-4, 2017*. pages 27–48.

Preslav Nakov, Lluís Màrquez, Walid Magdy, Alessandro Moschitti, Jim Glass, and Bilal Randeree. 2015. Semeval-2015 task 3: Answer selection in community question answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015, Denver, Colorado, USA, June 4-5, 2015*. pages 269–281.

Preslav Nakov, Lluís Màrquez, Alessandro Moschitti, Walid Magdy, Hamdy Mubarak, Abed Alhakim Freihat, Jim Glass, and Bilal Randeree. 2016. Semeval-2016 task 3: Community question answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*. pages 525–545.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*. pages 1532–1543.

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 *AAAI Conference on Artificial Intelligence*.

Yi Tay, Luu Anh Tuan, and Siu Cheung Hui. 2017. Cross temporal recurrent networks for ranking question answer pairs. *CoRR* abs/1711.07656.

Quan Hung Tran, Vu Tran, Tu Vu, Minh Nguyen, and Son Bao Pham. 2015. JAIST: combining multiple features for answer selection in community question answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015, Denver, Colorado, USA, June 4-5, 2015*. pages 215–219.

Shuohang Wang and Jing Jiang. 2017. A compare-aggregate model for matching text sequences. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Zhiguo Wang, Haitao Mi, and Abraham Ittycheriah. 2016. Sentence similarity learning by lexical decomposition and composition. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*.

GuoShun Wu, Yixuan Sheng, Man Lan, and Yuanbin Wu. 2017a. ECNU at semeval-2017 task 3: Using traditional and deep learning methods to address community question answering task. In *Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, Vancouver, Canada, August 3-4, 2017*. pages 365–369.

Wei Wu, Houfeng Wang, and Sujian Li. 2017b. Bi-directional gated memory networks for answer selection. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. LNAI 10565, Springer*. pages 251–262.

Yang Xiang, Xiaoqiang Zhou, Qingcai Chen, Zhihui Zheng, Buzhou Tang, Xiaolong Wang, and Yang Qin. 2016. Incorporating label dependency for answer quality tagging in community question answering via CNN-LSTM-CRF. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*. pages 1231–1241.

Xiaodong Zhang, Sujian Li, Lei Sha, and Houfeng Wang. 2017. Attentive interactive neural networks for answer selection in community question answering. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*. pages 3525–3531.