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, School of EECS, Peking University
[2]Collaborative Innovation Center for Language Ability

{wu.wei, xusun, wanghf}@pku.edu.cn

Abstract

Task:

Answer Selection in Community Question Answering

Problem:

A question includes both a subject that gives a brief summary of the question and a body that describes the question in detail.

The problem of redundancy and noise is prevalent in CQA. Both questions and answers contain auxiliary sentences that do not provide meaningful information.

Proposal:

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.

Example

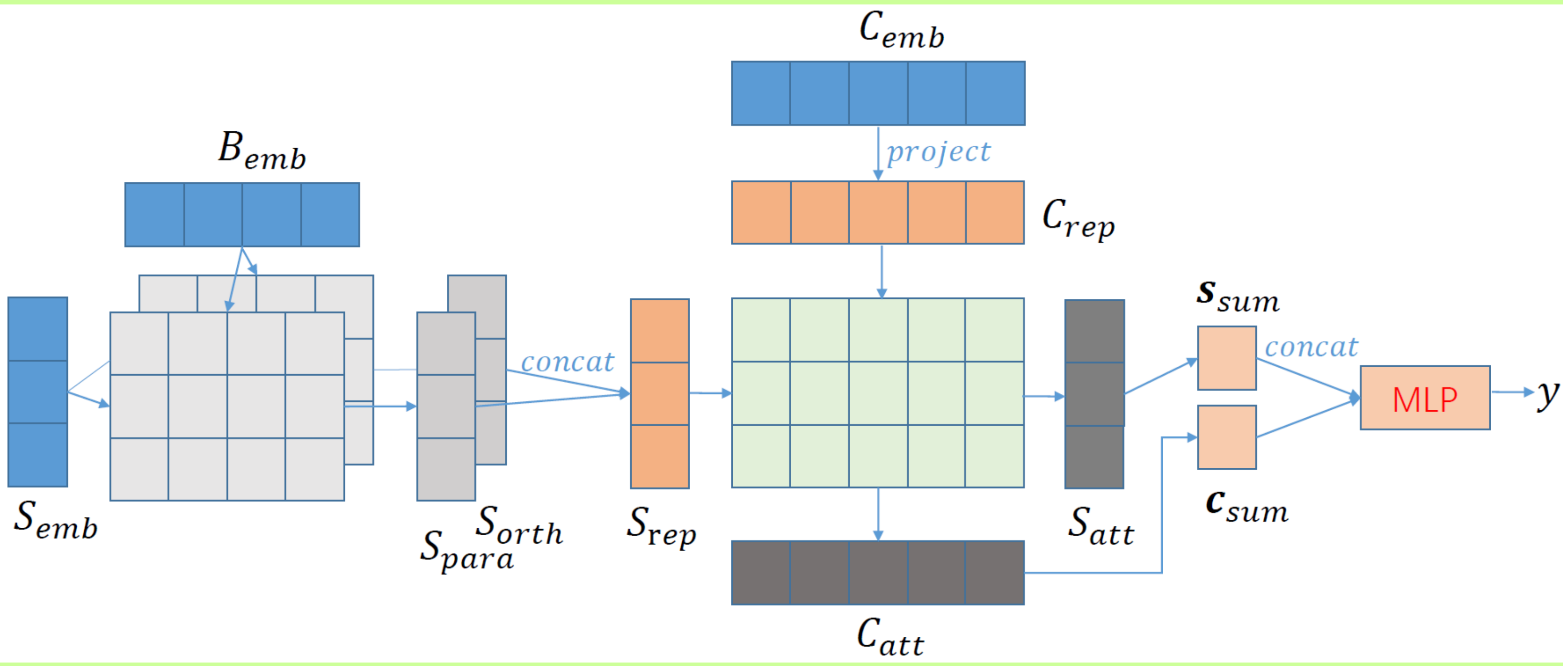
Question Subject: Checking the history of the car.

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.

Answer 1: 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..

Answer 2: 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 :)

Proposed Model



The overall architecture of our model is illustrated above. We propose Question Condensing Networks (QCN) which is composed of the following modules:

- **Question Condensing**

  We propose to cheat question subject as the primary part of the question representation, and aggregate question body information from two perspectives: similarity and disparity with question subject.

$$b_{para}^{i,j} = \frac{b_{emb}^{j} \cdot s_{emb}^{i}}{s_{emb}^{i} \cdot s_{emb}^{i}} s_{emb}^{i} \qquad w_{para}^{i,j} = \frac{\exp(a_{\mathrm{para}}^{i,j})}{\sum_{j=1} \exp(a_{para}^{i,j})}$$

$$b_{orth}^{i,j} = b_{emb}^{j} - b_{para}^{i,j} \qquad s_{ap}^{i} = \sum\nolimits_{j=1} w_{para}^{i,j} \odot b_{emb}^{i,j}$$

- **Question Condensing**

$$s_{ai}^{i} = \sum_{j=1}^{n} \frac{\exp(a_{align}^{i,j})}{\sum_{j=1}^{n} \exp(a_{align}^{i,j})} \otimes c_{rep}^{i} \qquad \mathrm{c}_{ai}^{i} = \sum_{j=1}^{l} \frac{\exp(a_{align}^{i,j})}{\sum_{j=1}^{l} \exp(a_{align}^{i,j})} \otimes s_{rep}^{i}$$

- **Interaction Condensing**

$$A_s = W_{s2} \tanh(W_{s1} S_{att} + b_{s1}) + b_{s2} \qquad s_{sum} = \sum_{i=1}^{n} \frac{\exp(a_s^{i})}{\sum_{i=1}^{n} \exp(a_s^{i})} \otimes s_{att}^{i}$$

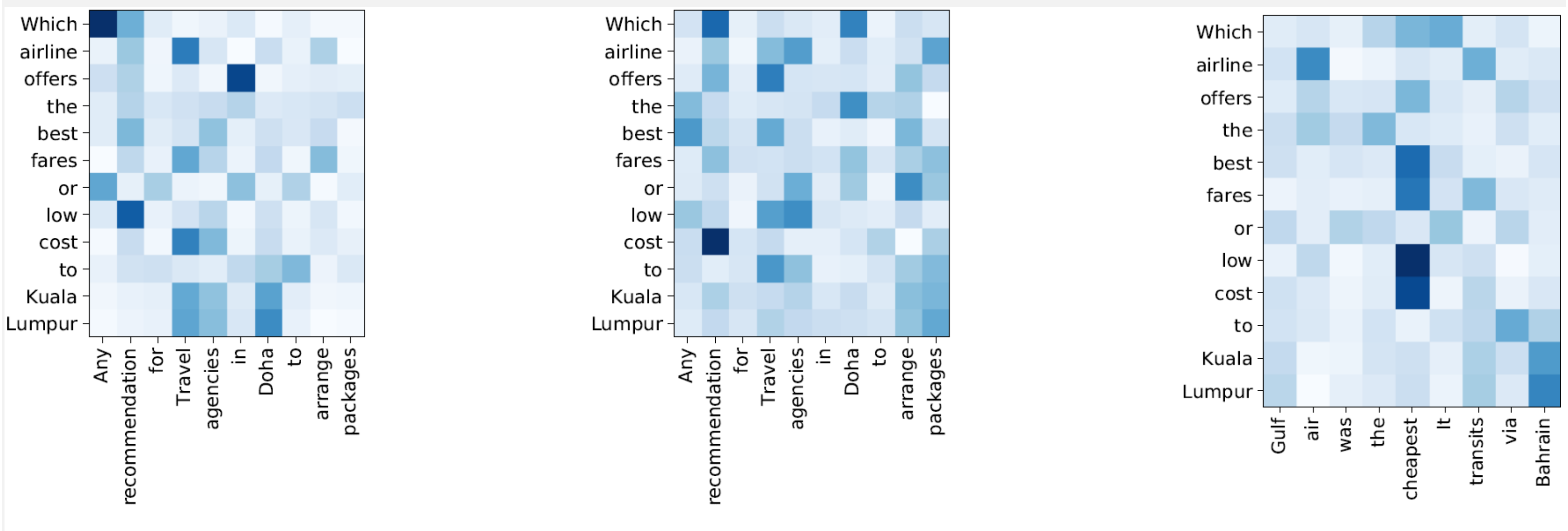
Experimental Results

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**

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**

Results of our model and the baselines on SemEval 2015 and SemEval 2017 datasets.QCN has a great advantage in terms of accuracy. We hypothesize that QCN focuses on modeling interaction between questions and answers, whether an answer can match the corresponding question.

Qualitative Study



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

Attention probabilities in A_{para}, $A_{orth}$and A_{align} is illustrated above.
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. 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.

Conclusion

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.