1 Ablation Test of Multi-Scale Convolution

2 Statistical Significant Experiments

2.1 TrecQA

We conduct statistical significant experiments on TrecQA dataset, and report the p-value in the experiments.

We use bootstrap method in this experiments. We split train dataset using k-fold cross validation method. Specifically, we use the length of each question list as the label, and then apply stratified k-fold method to split the dataset to make sure a question list is not split. Then we train our model using the same hyperparameters on k-fold sub-dataset, and evaluate the model performance. The result is demonstrated in Table 1.

To compare with the PairwiseRank + MP CNN [Rao et al.(2016)Rao, He, and Lin], which achieve 0.780 in terms of MAP and 0.834 in terms of MRR on TrecQA, we also conduct one-sample t-test and report the p-value in Table 1. Under the null hypothesis that MNAC is not significantly better than PairwiseRank + MP CNN, and given the level of significance α=0.05, the p-value is 0.00018 in terms of MAP and 0.0023 in terms of MRR. Therefore, we have strong confidence to reject the null hypothesis and declare that the performance of our model significantly outperforms the baseline model.

	MAP	MRR
kfold-1	0.788	0.842
kfold-2	0.785	0.836
kfold-3	0.783	0.841
kfold-4	0.787	0.845
kfold-5	0.789	0.844
kfold-6	0.793	0.829
kfold-7	0.795	0.84
kfold-8	0.785	0.842
kfold-9	0.783	0.843
kfold-10	0.789	0.848
Mean	0.7877	0.8410
Std	0.00400	0.00527
p-value	**0.00018**	**0.0023**

Table 1: Statistical Test of MNAC on TrecQA(raw)

3 Analysis & Case study

3.1 QuoraQP

For the Paraphrase Identification task, most of the negative samples in QuoraQP have the same intent but differ in qualifiers such as year/place/person. BiMPM is dedicated to capturing the similarity between two sentences in multiple levels of granularity with complicated stacked structure and multiple matching operations (quite slow in the training phase). However, BiMPM cannot outperform our model on wikiQA because semantic relatedness is much more important than lexical similarity in QA task and they could not be formed word-byword usually.. Our model demonstrates superiority in semantic understanding with efficient sentence-level semantic extraction mechanisms. The slight word-level difference may not be well captured in our model, MNAC can still distinguish the paraphrases with a more accurate and simple way.

References

[Rao et al.(2016)Rao, He, and Lin] Jinfeng Rao, Hua He, and Jimmy Lin. 2016. Noise-contrastive estimation for answer selection with deep neural networks. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 1913–1916. ACM.