**Table 1: Results of models and baselines with ablation study. MC and BR denote Mutual Coverage and Bag Representation respectively. "BR w/o Cov" denotes Bag Representation component without coverage module. ‡ and § means the results are significant with p-value $< 0.05$ measured by the Student's paired t-test over the best baseline and the base model respectively.**

| | AliMe | | | | | Quora | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Model | MRR | $R_{10}@1$ | $R_{10}@2$ | $R_{10}@5$ | $R_2@1$ | MRR | $R_{10}@1$ | $R_{10}@2$ | $R_{10}@5$ | $R_2@1$ |
| Q-Q Mean | 0.6122 | 0.5050 | 0.5623 | 0.7287 | 0.8473 | 0.8350 | 0.7847 | 0.8133 | 0.8973 | 0.9480 |
| Q-Q Max | 0.6470 | 0.5477 | 0.6000 | 0.7590 | 0.8523 | 0.8438 | 0.7980 | 0.8227 | 0.8980 | 0.9493 |
| Bag-Con | 0.6552 | 0.5610 | 0.6087 | 0.7607 | 0.8553 | 0.8026 | 0.7420 | 0.7800 | 0.8740 | 0.9287 |
| Base | 0.6845 | 0.6027 | 0.6397 | 0.7700 | 0.8707 | 0.8184 | 0.7643 | 0.7973 | 0.8800 | 0.9337 |
| Base+MC | 0.6936 | 0.6137 | 0.6497 | 0.7807 | **0.8823** | 0.8640 | 0.8247 | 0.8480 | 0.9083 | **0.9587** |
| Base+BR | 0.6913 | 0.6103 | 0.6443 | 0.7833 | 0.8783 | 0.8628 | 0.8213 | 0.8477 | 0.9123 | 0.9497 |
| Base+(BR w/o Cov) | 0.6849 | 0.6013 | 0.6410 | 0.7810 | 0.8727 | 0.8280 | 0.7763 | 0.8093 | 0.8833 | 0.9430 |
| QBM (Base+BR+MC) | **0.7007**‡§ | **0.6197**‡§ | **0.6600**‡§ | **0.7923**‡§ | **0.8823**‡ | **0.8656**‡§ | **0.8253**‡§ | **0.8510**‡§ | **0.9137**§ | 0.9520‡§ |

**Table 2: Some words and their corresponding weights ($e$ in Equation 4) in mutual coverage module. The average weight across the whole vocabulary is also presented here.**

| AliMe | | Quora | |
|---|---|---|---|
| 的 ('s) | 1.180 | The | 0.006 |
| 和 (And) | 1.237 | And | 0.894 |
| 退款 (Refund) | 5.042 | Where | 1.366 |
| 机票 (Air ticket) | 6.484 | America | 2.018 |
| Average | 2.202 | Average | 0.899 |

degree of the original query and the new "question", namely Bag-Con (Bag Concatenation).

### 3.4 Evaluation

Following Qiu et al. [4], we evaluate the model performance on five automatic evaluation metrics: MRR, $R_{10}@1$, $R_{10}@2$, $R_{10}@5$, and $R_2@1$. $R_n@k$ calculates the recall of the true positive pre-defined questions among the $k$ selected candidates from $n$ available candidates. And Mean Reciprocal Rank (MRR) is another popular measurement for ranking problems.

## 4 RESULTS AND ANALYSIS

**Results and Ablation Study** The results are shown in Table 1. Our model (QBM) performs best compared to baselines (Q-Q Mean, Q-Q Max, Bag-con). Comparing Bag-Con and Base model, we find that modelling the query-question relationship following aggregation works better. We assume that the pooling-based aggregation can reduce the redundant information cross sentences in a bag. Considering the Q-Q matching based methods and query-bag based methods. In AliMe dataset, the query-bag matching outperforms the Q-Q matching based methods which shows the necessity to perform query-bag matching. The ablation study shows that the mutual coverage component and bag representation component achieve better performance than the base model, especially in the Quora dataset. The two components work independently and their combination gets the best performance.

**Effectiveness of the Mutual Coverage** To intuitively learn the coverage weight, we sample some words with their weights in Table 2. It shows that the words like "The" have low weight, which confirms that they contribute little to the matching. "Refund" in E-commerce is a very important element in a user query sentence. And "America" is important in Quora, because question like "what is the capital in <location>?" is highly related to location "<location>".

**Analysis of the Bag Representation** Coverage is also applied in the bag representation layer. The results of the bag representation without coverage component (Base+(BR w/o Cov)) is shown in Table 1. Compared with the Base+BR and BR without coverage, it shows that the coverage component contributes a lot on both the two datasets. The bag representation with coverage (Base+BR) gains improvement over Base model, especially in Quora dataset.

## 5 CONCLUSION

In this paper, we propose the QBM model which performs the query-bag matching in information-seeking conversation. Experiments show that the proposed mutual coverage component improves the model performance. And the model can automatically discover important words in the query or bag from both the coverage weighting component and the word-level bag representation. This work also shows that learning the query-bag matching directly in some scenarios may outperform the query-question matching in ranking bags. One advantage of our model is that it is extensible in replacing the query-question matching component.

## REFERENCES


[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *ICLR* (2015).
[2] Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for Natural Language Inference. In *ACL*. 1657–1668.
[3] Hang Li and Jun Xu. 2012. Machine Learning for Query-Document Matching in Web Search. (2012).
[4] Minghui Qiu, Liu Yang, Feng Ji, Wei Zhou, Jun Huang, Haiqing Chen, Bruce Croft, and Wei Lin. 2018. Transfer Learning for Context-Aware Question Matching in Information-seeking Conversations in E-commerce. In *ACL*. 208–213.
[5] Gaurav Singh Tomar, Thyago Duque, Oscar Täckström, Jakob Uszkoreit, and Dipanjan Das. 2017. Neural Paraphrase Identification of Questions with Noisy Pretraining. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*. Association for Computational Linguistics, 142–147.
[6] Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017. Sequential Matching Network: A New Architecture for Multi-turn Response Selection in Retrieval-Based Chatbots. In *ACL*. 496–505.
[7] Runqi Yang, Jianhai Zhang, Xing Gao, Feng Ji, and Haiqing Chen. 2019. Simple and Effective Text Matching with Richer Alignment Features. In *ACL*. 4699–4709.
[8] Jianfei Yu, Minghui Qiu, Jing Jiang, Jun Huang, Shuangyong Song, Wei Chu, and Haiqing Chen. 2018. Modelling domain relationships for transfer learning on retrieval-based question answering systems in e-commerce. In *WSDM*. 682–690.