# Overview of the NLPCC 2017 Shared Task: Open Domain Chinese Question Answering

Nan Duan(✉) and Duyu Tang

Microsoft Research Asia, Beijing, China
{nanduan, dutang}@microsoft.com

**Abstract.** In this paper, we give the overview of the open domain Question Answering (or open domain QA) shared task in the NLPCC 2017. We first review the background of QA, and then describe two open domain Chinese QA tasks in this year's NLPCC, including the construction of the benchmark datasets and the evaluation metrics. The evaluation results of submissions from participating teams are presented in the experimental part.

**Keywords:** Question answering · Knowledge-based QA
Document-based QA · Table-based QA

## 1 Background

Question Answering (or QA) is a fundamental task in Artificial Intelligence, whose goal is to build a system that can automatically answer natural language questions. In the last decade, the development of QA techniques has been greatly promoted by both academic field and industry field.

In the academic field, with the rise of large scale curated knowledge bases, like Yago, Satori, Freebase, etc., more and more researchers pay their attentions to the knowledge-based QA (or KBQA) task, such as semantic parsing-based approaches [1–7] and information retrieval-based approaches [8–16]. Besides KBQA, researchers are interested in document-based QA (or DBQA) as well, whose goal is to select answers from a set of given documents and use them as responses to natural language questions. Usually, information retrieval-based approaches [18–22] are used for the DBQA task.

In the industry field, many influential QA-related products have been built, such as IBM Watson, Apple Siri, Google Now, Facebook Graph Search, Microsoft Cortana and XiaoIce etc. These kinds of systems are immerging into every user's life who is using mobile devices.

Under such circumstance, in this year's NLPCC shared task, we call the open domain QA task that cover both KBQA and DBQA tasks. Our motivations are two-folds:

1. We expect this activity can enhance the progress of QA research, esp. for Chinese;
2. We encourage more QA researchers to share their experiences, techniques, and progress.

X. Huang et al. (Eds.): NLPCC 2017, LNAI 10619, pp. 954–961, 2018.
https://doi.org/10.1007/978-3-319-73618-1_86

Besides these two tasks mentioned above, we also prepared a new task: Table-based QA (TBQA). However, there is no final submission for this task, so we skip the description of this task in this paper.

The remainder of this paper is organized as follows. Section 2 describes two open domain Chinese QA tasks. In Sect. 3, we describe the benchmark datasets constructed. Section 3 describes evaluation metrics, and Sect. 4 presents the evaluation results of different submissions. We conclude the paper in Sect. 5, and point out our plan on future QA evaluation activities.

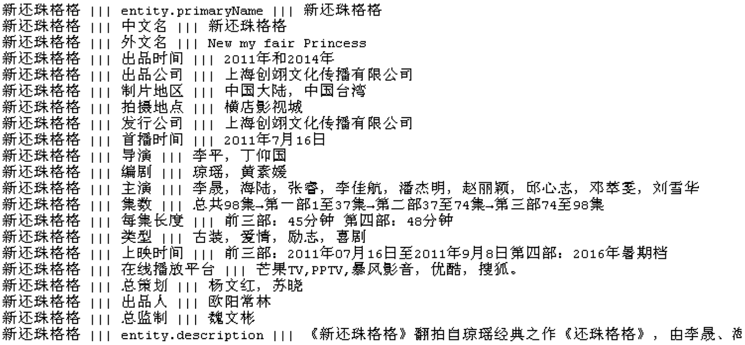
## 2 Task Description

The NLPCC 2017 open domain QA shared task includes two QA tasks for Chinese language: knowledge-based QA (KBQA) task and document-based QA (DBQA) task.

### 2.1 KBQA Task

Given a question, a KBQA system built by each participating team should select one or more entities as answers from a given knowledge base (KB). The datasets for this task include:

- **A Chinese KB.** It includes knowledge triples crawled from the web. Each knowledge triple has the form: <Subject, Predicate, Object>, where 'Subject' denotes a subject entity, 'Predicate' denotes a relation, and 'Object' denotes an object entity. A sample of knowledge triples is given in Fig. 1, and the statistics of the Chinese KB is given in Table 1.



```
新还珠格格 ||| entity.primaryName ||| 新还珠格格
新还珠格格 ||| 中文名 ||| 新还珠格格
新还珠格格 ||| 外文名 ||| New my fair Princess
新还珠格格 ||| 出品时间 ||| 2011年和2014年
新还珠格格 ||| 出品公司 ||| 上海创翊文化传播有限公司
新还珠格格 ||| 制片地区 ||| 中国大陆，中国台湾
新还珠格格 ||| 拍摄地点 ||| 横店影视城
新还珠格格 ||| 发行公司 ||| 上海创翊文化传播有限公司
新还珠格格 ||| 首播时间 ||| 2011年7月16日
新还珠格格 ||| 导演 ||| 李平，丁仰国
新还珠格格 ||| 编剧 ||| 琼瑶，黄素媛
新还珠格格 ||| 主演 ||| 李晟，海陆，张睿，李佳航，潘杰明，赵丽颖，邱心志，邓萃雯，刘雪华
新还珠格格 ||| 集数 ||| 总共98集→第一部1至37集→第二部37至74集→第三部74至98集
新还珠格格 ||| 每集长度 ||| 前三部：45分钟 第四部：48分钟
新还珠格格 ||| 类型 ||| 古装，爱情，励志，喜剧
新还珠格格 ||| 上映时间 ||| 前三部：2011年07月16日至2011年9月8日第四部：2016年暑期档
新还珠格格 ||| 在线播放平台 ||| 芒果TV,PPTV,暴风影音，优酷，搜狐。
新还珠格格 ||| 总策划 ||| 杨文红，苏晓
新还珠格格 ||| 出品人 ||| 欧阳常林
新还珠格格 ||| 总监制 ||| 魏文彬
新还珠格格 ||| entity.description ||| 《新还珠格格》翻拍自琼瑶经典之作《还珠格格》，由李晟、海
```

**Fig. 1.** An example of the Chinese KB.

- **Training set and testing set.** We assign a set of knowledge triples sampled from the Chinese KB to human annotators. For each knowledge triple, a human annotator will write down a natural language question, whose answer should be the object entity of the current knowledge triple. In last year's NLPCC KBQA task, we

**Table 1.** Statistics of the Chinese KB.

| | |
|---|---|
| **# of Subject Entities** | 8,721,640 |
| **# of Triples** | 47,943,429 |
| **# of Averaged Triples per Subject Entity** | 5.5 |

released 14,609 labeled QA pairs as training set, and 9,870 labeled QA pairs as testing set. In this year, we provide a new testing set, which includes 7,631 labeled QA pairs. We follow the same way to annotate this dataset as we did last year. Besides, we also used Automatic Question Generation technique to generate a set faked questions, and mixed them into human labeled questions to form a larger testing set. These generated questions and their corresponding answers will be ignored in the evaluation phase. The statistic of labeled QA pairs and an annotation example are given in Table 2:

**Table 2.** Statistics of the KBQA datasets.

<table>
<tr><td><b># of Labeled Q-A Pairs<br>(training set, 2016)</b></td><td colspan="2">14,609</td></tr>
<tr><td><b># of Labeled Q-A Pairs<br>(testing set, 2016)</b></td><td colspan="2">9,870</td></tr>
<tr><td><b># of Labeled Q-A Pairs<br>(testing set, 2017)</b></td><td colspan="2">7.631</td></tr>
<tr><td rowspan="3"><b>An Example</b></td><td><b>Triple</b></td><td><微软，创始人，比尔盖茨></td></tr>
<tr><td><b>Labeled Question</b></td><td>微软公司的创始人是谁？</td></tr>
<tr><td><b>Golden Answer</b></td><td>比尔盖茨</td></tr>
</table>

In KBQA task, any data resource can be used to train necessary models, such as entity linking, semantic parsing, etc., but answer entities should come from the provided KB only.

## 2.2 DBQA Task

Given a question and its corresponding document, a DBQA system built by each participating team should select one or more sentences as answers from the document. The datasets for this task include:

- **Training set and testing set.** We assign a set of documents to human annotators. For each document, a human annotator will (1) first, select a sentence from the document, and (2) then, write down a natural language question, whose answer should be the selected sentence. In last year's NLPCC DBQA task, we released 8,772 labeled Q-document pairs as training set, and 5,779 labeled Q-document pairs as testing set. In this year, we provide a new testing set as well, which includes 2,500 labeled QA pairs. Like KBQA, we released a larger testing set by adding some automatically generated questions and ignored them during the evaluation phase. The statistic of labeled QD pairs and an annotation example are given in Table 3:

**Table 3.** Statistics of the DBQA datasets.

| | |
|---|---|
| **# of Labeled Q-D Pairs (training set, 2016)** | 8,772 |
| **# of Labeled Q-D Pairs (testing set, 2016)** | 5,779 |
| **# of Labeled Q-D Pairs (testing set, 2017)** | 2,500 |
| **A Q-D Pair Example** | 俄罗斯贝加尔湖的面积有多大? \t 贝加尔湖，中国古代称为北海，位于俄罗斯西伯利亚的南部。 \t **0**<br>俄罗斯贝加尔湖的面积有多大? \t 贝加尔湖是世界上最深，容量最大的淡水湖。 \t **0**<br>俄罗斯贝加尔湖的面积有多大? \t 贝加尔湖贝加尔湖是世界上最深和蓄水量最大的淡水湖。 \t **0**<br>俄罗斯贝加尔湖的面积有多大? \t 它位于布里亚特共和国(Buryatiya) 和伊尔库茨克州(Irkutsk) 境内。 \t **0**<br>俄罗斯贝加尔湖的面积有多大? \t 湖型狭长弯曲，宛如一弯新月，所以又有'月亮湖'之称。 \t **0**<br>俄罗斯贝加尔湖的面积有多大? \t 湖长636公里平均宽48公里，最宽79.4公里，面积3.15万平方公里。 \t **1**<br>俄罗斯贝加尔湖的面积有多大? \t 贝加尔湖湖水澄澈清冽，且稳定透明（透明度达40.8米），为世界第二。 \t **0** |

As shown in the example in Table 3, a question (the 1st column), question's corresponding document sentences (the 2nd column), and their answer annotations (the 3rd column) are provided. If a document sentence is the correct answer of the question, its annotation will be 1, otherwise its annotation will be 0. The three columns will be separated by the symbol '\t'.

In DBQA task, any data resource can be used to train necessary models, such as paraphrasing model, sentence matching model, etc., but answer sentences should come from the provided documents only.

## 3 Evaluation Metrics

The quality of a KBQA system is evaluated by **Averaged F1**, and the quality of a DBQA system is evaluated by **MRR**, **MAP**, and **ACC@1**.

- **Averaged F1**

$$AveragedF1 = \frac{1}{|Q|}\sum_{i=1}^{|Q|} F_i$$

$F_i$ denotes the F1 score for question $Q_i$ computed based on $C_i$ and $A_i$. $F_i$ is set to 0 if $C_i$ is empty or doesn't overlap with $A_i$. Otherwise, $F_i$ is computed as follows:

$$F_i = \frac{2 \cdot \frac{\#(C_i,A_i)}{|C_i|} \cdot \frac{\#(C_i,A_i)}{|A_i|}}{\frac{\#(C_i,A_i)}{|C_i|} + \frac{\#(C_i,A_i)}{|A_i|}}$$

where $\#(C_i, A_i)$ denotes the number of answers occur in both $C_i$ and $A_i$. $|C_i|$ and $|A_i|$ denote the number of answers in $C_i$ and $A_i$ respectively.

- **MRR**

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|} \frac{1}{rank_i}$$

$|Q|$ denotes the total number of questions in the evaluation set, $rank_i$ denotes the position of the first correct answer in the generated answer set $C_i$ for the $i^{th}$ question $Q_i$. If $C_i$ doesn't overlap with the golden answers $A_i$ for $Q_i$, $\frac{1}{rank_i}$ is set to 0.

- **MAP**

$$MAP = \frac{1}{|Q|}\sum_{i=1}^{|Q|} AveP(C_i, A_i)$$

$AveP(C, A) = \frac{\sum_{k=1}^{n}(P(k) \cdot rel(k))}{min(m,n)}$ denotes the average precision. $k$ is the rank in the sequence of retrieved answer sentences. $m$ is the number of correct answer sentences. $n$ is the number of retrieved answer sentences. If $min(m, n)$ is 0, $AveP(C, A)$ is set to 0. $P(k)$ is the precision at cut-off $k$ in the list. $rel(k)$ is an indicator function equaling 1 if the item at rank $k$ is an answer sentence, and 0 otherwise.

- **ACC@N**

$$Accuracy@N = \frac{1}{|Q|}\sum_{i=1}^{|Q|} \delta(C_i, A_i)$$

$\delta(C_i, A_i)$ equals to 1 when there is at least one answer contained by $C_i$ occurs in $A_i$, and 0 otherwise.

## 4 Evaluation Results

There are 35 teams submitted their results. Tables 4 and 5 lists the evaluation results of DBQA and KBQA tasks respectively.

**Table 4.** Evaluation results of the DBQA task.

| | MRR | MAP | ACC@1 |
|---|---|---|---|
| Team 1 | 0.720194 | 0.716594 | 0.592 |
| Team 2 | 0.689619 | 0.68576 | 0.5556 |
| Team 3 | 0.685011 | 0.680963 | 0.5512 |
| Team 4 | 0.683674 | 0.680067 | 0.5492 |
| Team 5 | 0.677203 | 0.673271 | 0.54 |
| Team 6 | 0.675772 | 0.670828 | 0.5356 |
| Team 7 | 0.672872 | 0.668659 | 0.5372 |
| Team 8 | 0.664586 | 0.660256 | 0.5244 |
| Team 9 | 0.660674 | 0.658893 | 0.5144 |
| Team 10 | 0.652062 | 0.649218 | 0.5056 |
| Team 11 | 0.583311 | 0.580741 | 0.4284 |
| Team 12 | 0.557158 | 0.556341 | 0.3996 |
| Team 13 | 0.54846 | 0.545021 | 0.372 |
| Team 14 | 0.533575 | 0.531831 | 0.3692 |
| Team 15 | 0.506718 | 0.503114 | 0.3404 |
| Team 16 | 0.494292 | 0.491736 | 0.3288 |
| Team 17 | 0.436557 | 0.434162 | 0.2696 |
| Team 18 | 0.402115 | 0.40085 | 0.2172 |
| Team 19 | 0.384112 | 0.382343 | 0.2016 |
| Team 20 | 0.384112 | 0.382343 | 0.2016 |
| Team 21 | 0.353259 | 0.352269 | 0.1744 |

**Table 5.** Evaluation results of the KBQA task.

| | Average Precision | Average Recall | Average F1 |
|---|---|---|---|
| **Team 1** | 0.472284104 | 0.472284104 | 0.472284104 |
| **Team 2** | 0.412615314 | 0.435984799 | 0.419647927 |
| **Team 3** | 0.401511483 | 0.418817979 | 0.406784423 |
| **Team 4** | 0.395205646 | 0.41410038 | 0.400818095 |
| **Team 5** | 0.372886909 | 0.413052025 | 0.386275281 |
| **Team 6** | 0.351565082 | 0.478050059 | 0.371838481 |
| **Team 7** | 0.357257566 | 0.381994496 | 0.36381076 |
| **Team 8** | 0.339840781 | 0.364696632 | 0.347019363 |
| **Team 9** | 0.339819304 | 0.36522081 | 0.346864086 |
| **Team 10** | 0.329966705 | 0.360896344 | 0.338715307 |
| **Team 11** | 0.328349034 | 0.359061722 | 0.337029202 |
| **Team 12** | 0.313589307 | 0.313589307 | 0.313589307 |
| **Team 13** | 0.269689425 | 0.269689425 | 0.269689425 |
| **Team 14** | 0.213995544 | 0.213995544 | 0.213995544 |

## 5 Conclusion

This paper briefly introduces the overview of this year's two open domain Chinese QA shared tasks. In the future, we plan to provide more QA datasets for Chinese QA field. Besides, we plan to extend the QA tasks from Chinese to English as well, and promote new QA tasks, such as Table-based QA.

## References

1. Wang, Y., Berant, J., Liang, P.: Building a semantic parser overnight. In: ACL (2015)
2. Pasupat, P., Liang, P.: Compositional semantic parsing on semi-structured tables. In: ACL (2015)
3. Pasupat, P., Liang, P.: Zero-shot entity extraction from web pages. In: ACL (2014)
4. Bao, J., Duan, N., Zhou, M., Zhao, T.: Knowledge-based question answering as machine translation. In: ACL (2014)
5. Yang, M.-C., Duan, N., Zhou, M., Rim, H.-C.: Joint relational embeddings for knowledge-based question answering. In: EMNLP (2014)
6. Berant, J., Chou, A., Frostig, R., Liang, P.: Semantic parsing on freebase from question-answer pairs. In: EMNLP (2013)
7. Kwiatkowski, T., Choi, E., Artzi, Y., Zettlemoyer, L.: Scaling semantic parsers with on-the-fly ontology matching. In: EMNLP (2013)

8. Bordes, A., Usunier, N., Chopra, S., Weston, J.: Large-scale simple question answering with memory network. In: ICLR (2015)
9. Weston, J., Bordes, A., Chopra, S., Mikolov, T.: Towards AI-complete Question Answering: A Set of Prerequisite Toy Tasks. arXiv (2015)
10. Dong, L., Wei, F., Zhou, M., Xu, K.: Question answering over freebase with multi-column convolutional neural networks. In: ACL (2015)
11. Yih, W., Chang, M.-W., He, X., Gao, J.: Semantic parsing via staged query graph generation: question answering with knowledge base. In: ACL (2015)
12. Yao, X.: Lean question answering over freebase from scratch. In: NAACL (2015)
13. Berant, J., Liang, P.: Semantic parsing via paraphrasing. In: ACL (2014)
14. Yao, X., Van Durme, B.: Information extraction over structured data: question answering with freebase. In: ACL (2014)
15. Bordes, A., Weston, J., Chopra, S.: Question answering with subgraph embeddings. In: EMNLP (2014)
16. Bordes, A., Weston, J., Usunier, N.: Open question answering with weakly supervised embedding models. In: Calders, T., Esposito, F., Hüllermeier, E., Meo, R. (eds.) ECML PKDD 2014. LNCS (LNAI), vol. 8724, pp. 165–180. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-662-44848-9_11
17. Yang, Y., Yih, W., Meek, C.: WIKIQA: a challenge dataset for open-domain question answering. In: EMNLP (2015)
18. Miao, Y., Yu, L., Blunsom, P.: Neural variational inference for text processing. arXiv (2015)
19. Wang, D., Nyberg, E.: A long short term memory model for answer sentence selection in question answering. In: ACL (2015)
20. Yin, W., Schütze, H., Xiang, B., Zhou, B.: ABCNN: attention-based convolutional neural network for modeling sentence pairs. In: ACL (2016)
21. Yu, L., Hermann, K.M., Blunsom, P., Pullman, S.: Deep learning for answer sentence selection. In: NIPS Workshop (2014)
22. Yan, Z., Duan, N., Bao, J., Chen, P., Zhou, M., Li, Z., Zhou, J.: DocChat: an information retrieval approach for chatbot engines using unstructured documents. In: ACL (2016)