Overview of the NLPCC 2015 Shared Task: Open Domain QA

Nan Duan(✉)

Microsoft Research, Beijing, China
nanduan@microsoft.com

Abstract. In this paper, we give the overview of the open domain Question Answering (or open domain QA) shared task in NLPCC 2015. We first review the background of QA, and then describe open domain QA shared task in this year's NLPCC, including the construction of the benchmark datasets, the auxiliary dataset, and the evaluation metrics. The evaluation results of submissions from participating teams are presented in the experimental part, together with a brief introduction to the techniques used in each participating team's QA system.

Keywords: Question answering · Knowledge base

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 have 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 and etc., more and more researchers pay their attentions to the open domain QA task. The state-of-the-art methods used in open domain QA can by summarized into two categories: semantic parsing-based approaches and information retrieval-based approaches. Semantic parsing-based approaches, such as [1] [2] [3] [4] [5] [6] [7], first transform a natural language question into its corresponding meaning representation, and then use it as a structured query to lookup answers from an existing KB; information retrieval-based approaches, such as [8] [9] [10] [11] [12] [13] [14] [15] [16], first define and generate the representations of answers stored in KB, and then retrieve the most relevant answers from KB by computing the similarity between input questions and the representations of answers. Recently, with the development of the open IE techniques [20] [21] [22] [23] [24] [25], some approaches, such as [17] [18] [19], build QA systems based on extracted knowledge bases, which consist of assertions extracted from unstructured text by open IE. Comparing to curated KBs, extracted KBs can be extracted from arbitrary corpus, so it is very flexible to be applied to any specific domain. But it also suffers the extraction noise issue, which is brought by open IE. In the industry field, many influential QA-related products have been built, such as

J. Li et al. (Eds.): NLPCC 2015, LNAI 9362, pp. 562–570, 2015.
DOI: 10.1007/978-3-319-25207-0_53

IBM Watson, Apple Siri, Google Now, Facebook Graph Search, Microsoft Cortana/XiaoIce and etc. These kind of systems are immerging into every user's life who is using mobile devices. Under such circumstance, in this year's NLPCC shared tasks, we call the open domain QA task, whose motivations are two-folds:

1. We expect this activity can provide more benchmark data for QA research, especially for Chinese;
2. We encourage more QA researchers to share their experiences, new techniques, and latest progress.

The remainder of this paper is organized as follows: Section 2 simply describe this year's open domain QA shared task; in Section 3, we will describe the benchmark datasets used in this year's QA evaluation, Section 4 describe the auxiliary dataset, which are crawled from semi-structured web pages and can be used as a structured database to build a Chinese QA system; in Section 5, we describe several evaluation metrics that are used to measure the QA quality of submissions generated by participating teams, and present the evaluation results of different submissions in Section 6, with a brief introduction to the techniques used by each team; Finally, we will conclude the paper in Section 7.

2 Task Description

This year's QA shared task provides two benchmark datasets, one for English and one for Chinese. For each question in each dataset, the participating teams should provide a list of answers as the prediction. We don't restrict participating teams to use any specified data for answer generation, so any data resources can be used. We evaluate the quality of the generated answers submitted from each team based on golden answers and several evaluation metrics (described in Section 5). Each team is allowed to provide multiple submissions for each dataset, but should specify one of them as their primary result.

3 Benchmark Data

Recently, two benchmark datasets for English have been released and frequently used by the academic field for the open domain QA task, including:

- *SimpleQuestions* dataset [8] consists of a total of 108,442 question-answer pairs, each of which is labeled by human English-speaking annotators based on a *single* fact from Freebase.
- *WebQuestions* dataset [6] consists of a total of 5,810 question-answer pairs. Different from SimpleQuestions, the questions are selected from the Goggle Suggest API, and then labeled by the Amazon Mechanical Turk (AMT) based on Freebase.

In order to encourage more researchers and institutions in China to devote to the QA research, in this year's NLPCC QA shared task, we specially provide a QA benchmark dataset for Chinese, together with another QA benchmark dataset for English as well. There two datasets are described as follows:

- *NLPCC15QuestionsCH* dataset consists of 1,000 question-answer pairs for Chinese. These questions are randomly sampled from a subset of queries coming from Bing China's query log based on the following rules: (1) each query's character length should be between 5 and 25; (2) each query should contain at least one n-gram contained in an answer type name list, such as '谁', '哪里', '哪一年', etc. These names are heuristically extracted and collected based on a simple statistical analysis on the entire query log; (3) each query should be answered based on an auxiliary data only. We will describe the auxiliary data in the later part. Each question is labeled by a list of answers, which should be agreed by three human annotators.
- *NLPCC15QuestionsEN* dataset consists of 68,481 question-answer pairs for English. These questions are collected in a hybrid way: some of the questions come from online QA sites like WikiAnswers and EVI; while the other questions are randomly sampled from a subset of queries coming from Bing US's query log and labeled by human annotators. The answers of the questions in the former part are crawled from the sites directly; while the answers of the questions in the latter part are labeled by based on Freebase. Unfortunately, no team submitted results for this dataset, and we encourage more institutes can leverage this dataset in the future for the QA research.

The answer annotations of all questions have already been provided to the participating teams, as the evaluation procedure has been finished. Two examples of these two datasets mentioned above are shown in Table 1, and their corresponding statistics are shown in Table 2.

Table 1. Two examples of NLPCC15QuestionsCH and NLPCC15QuestionsEN.

NLPCC15QuestionsCH	
<question id="1">	谁能百里挑一是哪个电视台的节目?
<answer id="1">	东方卫视

NLPCC15QuestionsEN	
<question id="1">	Who founded CBS?
<answer id="1">	William S. Paley

Table 2. Statistics of NLPCC15QuestionsCH and NLPCC15QuestionsEN.

	NLPCC15QuestionsCH	NLPCC15QuestionsEN
# of Questions	1,000	68,481
Averaged Question Length	12.8 (characters)	4.8466 (words)
Averaged Answer Numbers per Question	1.4	1.4

4 Auxiliary Data

Freebase is available for all Web users, so all teams can build QA systems for the English dataset based on APIs provided by Freebase. But such convenient APIs are not available for Chinese. In order to facilitate the system construction procedure for the Chinese QA task, we provide an auxiliary data resource, which plays a role as a Chinese knowledge base.

Formally, each entry stored in the auxiliary has the triple form: <Subject, Predicate, Argument>, where 'Subject' denotes an entity, 'Predicate' denotes a relation, and 'Argument' denotes either an entity or a string that gives a description of the subject entity. This data set is extracted from Baidu Baike pages. An example is given below in Figure 1, and some statistics of this data set are shown in Table 3.

```
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 auxiliary data for NLPCC15QuestionsCH.

Table 3. Statistics of auxiliary data for NLPCC15QuestionsCH.

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

Note, the answers of the questions in NLPCC15QuestionsCH are labeled based on this auxiliary data only, in order to ensure the participating teams can achieve reasonable QA quality by just using the auxiliary data we provided. Of course, other data resources are allowed to be used as well, such as other structured knowledge bases, web pages or offline documents.

5 Evaluation Metric

In this year's QA shared task, the quality of different QA systems are measured by the three evaluation metrics described below:

- Mean Reciprocal Rank (MRR)

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

Where $|Q|$ denotes the total number of questions in the dataset, $rank_i$ denotes the position of the first correct answer in the generated answers 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.

- Accuracy@N

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

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

- Averaged F1

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

Where 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.

6 Evaluation Result

There are totally 12 teams registered for the Chinese QA task, and 7 teams registered for the English QA task. However, only 3 teams submitted final results for the Chinese QA task, and no submission is received for the English QA task. Table 4 lists some statistics of these three submissions:

Table 4. Statistics of submissions.

	# of Questions Answered	Average # of Answers per Question
Team 1	432	1.82
Team 2	1,000	1.06
Team 3	972	1.98

From Table 4 we can see that Team 2's QA system generates answers for all the 1,000 questions, and Team 3's QA system only ignored 28 questions. Team 1's QA system only generate answers for 432 questions, but this angle only cannot tell 'good' or 'bad' of a QA system, as saying no also represents an intelligence for an AI system. We can also see that the average number of answers of Team 2 is nearly one, which means they only provide their top-1 results as the predicted answer for most questions, while this number for Team 1 and Team 3 is nearly two.

We then list the evaluation results in Table 5, 6, and 7, based on MRR, Accuracy@N, and F1 score respectively. From these 3 tables we can see that Team 2 performs better than the other two teams on all three metrics.

Table 5. Evaluation results based on MRR.

	MRR
Team 1	0.1430
Team 2	0.5675
Team 3	0.3360

Table 6. Evaluation results based on Accuracy@N.

	ACC@1	ACC@2	ACC@3	ACC@4	ACC@5
Team 1	0.1270	0.1490	0.1590	0.1650	0.1660
Team 2	0.5650	0.5700	0.5700	0.5700	0.5700
Team 3	0.2640	0.3980	0.4130	0.4130	0.4130

Table 7. Evaluation results on all questions based on F1 Score.

	Precision	Recall	F1 Score
Team 1	0.1169	0.1439	0.1196
Team 2	0.5660	0.5103	0.5240
Team 3	0.2890	0.3547	0.2990

Recall that Team 1 only answer 50% of questions, and in order to compare the quality of different QA systems from more perspectives, we also compare Precision, Recall, and F1 Score on answered questions only. Evaluation results are shown in Table 8, from which we can see that Team 2's results don't change, as they generate answers for all questions; Team 3's results change a little, this is due to the reason that they only ignored 28 questions; Team 1's results become much better, almost comparable to Team 3's numbers. This is to say that Team 1 can further improve their system by enlarging the recall first.

Table 8. Evaluation results on answered questions based on F1 Score.

	Precision	Recall	F1 Score
Team 1	0.2706	0.3331	0.2769
Team 2	0.5660	0.5103	0.5240
Team 3	0.2973	0.3650	0.3076

We also investigate the oracle results of different QA systems, and compare them with the corresponding Accuracy@1 in Table 9. From Table 9 we can see that Team 3 has the largest potential to improve their system; while the potential of Team 2's system is very limited. One key reason of this finding is that, for most of questions, Team 2 just submit a single answer as output.

Table 9. Oracle results.

	ACC@1	Oracle
Team 1	0.1270	0.1660
Team 2	0.5650	0.5700
Team 3	0.2640	0.4130

In order to understand the differences between these three QA systems, we tried to find system description papers from this year's submissions during the paper review procedure, by checking whether there is any evaluation result reported based on NLPCC15QuestionsCH. Below gives a brief introduction to the techniques used in Team 2 and Team 3's QA systems. It is a pity that we failed to find the corresponding system description for Team 1.

- Team 2 leverages both triple knowledge and search engine to answer input questions. For the triple knowledge part, the SPE algorithm is used to transform a natural language question into a triple query; For search engine part, the WKE algorithm is used to extract answer candidates from both Baidu Zhixin and unstructured web texts. This method achieves the best result on the Chinese QA dataset. We also

expect Team 2 can show more detailed evaluation results, to compare the impacts of triple knowledge and search engine in their QA system.
- Team 3 leverages 3 steps to predict answer candidates for a given question, including (1) question analysis, which detects the answer type of a given question based on heuristic rules; (2) multi-source retrieval, which extract answer candidates from two main resources, including knowledge triples and social QA collections (i.e. <question, answer> pairs); and (3) candidate ranking, which ranks different answer candidates based on similarity/redundancy features.

We also did some analysis on the answers predicted by three teams, and found that the evaluation results are better actually. This is because that in some cases, an answer is decided to be wrong just because it cannot match the labeled answer in an exact way. Below is an example:

- Question: 哪些城市有迪士尼乐园？
- Golden Answers: ["洛杉矶","奥兰多","东京","巴黎","香港"]
- Predicted Answers: ["美国加州迪士尼乐园","美国奥兰多迪斯尼世界","日本东京迪斯尼乐园","法国巴黎迪斯尼乐园","中国香港迪斯尼乐园"]

This is due to the fact that currently different resources are allowed to be used for answer generation, so answers extracted from different corpus may have different surface forms but identical semantic meaning. Such issue can be alleviated by using a specified KB only for answer extraction.

7 Conclusion

This paper briefly introduce the overview of this year's Open Domain QA shared task. Although there are only 3 teams that submitted results finally, we still see promising results and different techniques used. We are looking forward more organizations can take part in this yearly activity, and more benchmark data sets and techniques will be delivered to the community.

Reference

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 (2015). arXiv
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. 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., Durme, V.: 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, Part I. LNCS, vol. 8724, pp. 165–180. Springer, Heidelberg (2014)
17. Yin, P., Duan, N., Kao, B., Bao, J., Zhou, M.: Answering questions with complex semantic constraints on open KBs. In: CIKM (2015)
18. Fader, A., Zettlemoyer, L., Etzioni, O.: Open question answering over curated and extracted knowledge bases. In: KDD (2014)
19. Fader, A., Zettlemoyer, L., Etzioni, O.: Paraphrase-driven learning for open question answering. In: ACL (2013)
20. Del Corro, L., Gemulla, R.: ClausIE: Clause-based open information extraction. In: WWW (2013)
21. Mausam, Schmitz, M., Bart, R., Soderland, S., Etzioni, O.: Open language learning for information extraction. In: EMNLP (2012)
22. Yahya, M., Berberich, K., Elbassuoni, S.: Natural language questions for the web of data. In: EMNLP-CoNLL (2012)
23. Fader, A., Soderland, S., Etzioni, O.: Identifying relations for open information extraction. In: EMNLP (2011)
24. Wu, F., Weld, D.S.: Open information extraction using wikipedia. In: ACL (2010)
25. Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M., Etzioni, O.: Open information extraction from the web. In: IJCAI (2007)