NLPCC 2015 Shared Task Guideline:

Open Domain Question Answering

1. Task Description

The task of Question Answering (or QA) is to build systems that can automatically answer natural language questions. Like last year, we call **Open Domain Question Answering task** in this year's NLPCC 2015 shared tasks. We hope this activity can provide more benchmark data for QA research, and encourage more QA researchers to share their experiences, techniques, and progress.

Two testing data sets (one for English and one for Chinese) will be provided to all participating teams, each of which contains a list of questions. For each question, each participating team should submit a list of ranked answers generated by their QA system. We will evaluate the quality of the generated answers submitted from each team based on golden answers and several QA metrics, and then provide a comprehensive report. Each team is allowed to provide multiple submissions for each testing data set, but should specify one of them as their primary result.

2. Data Description

2.1 Testing Data

This year's QA task provides two testing data sets for English and Chinese respectively, and two examples below are given to describe the data format for each of them:

<question id="1">	In what movies has Greg Grunberg starred with Majel Barrett?
<answer id="1">	Star Trek

<question id="1">	微软公司的创始人是谁？
<answer id="1">	比尔盖茨
<answer id="2">	保罗艾伦

Each testing set includes a list of questions. We have labeled golden answers to each question, but this information will NOT be provided to participants, until we finish the entire evaluation procedure and report results to all teams. The format of the submission result should follow the same format described above. If no answer can be generated for a given question, then just set the value of **<answer id="1">** to an empty string.

The questions in the English testing data comes from the following sources:

- Bing's query log
- QA website
- Automatically generated based on knowledge base

The questions in the Chinese testing data comes from the following sources:

- Bing's query log
- Automatically generated based on knowledge base

2.2 Auxiliary Data[1]

Besides two testing data sets described above, we also provide two auxiliary data resources for download. Such data could be used for extracting answer candidates or training QA systems:

- Freebase
  It is a curated knowledge base that is commonly used in the QA field recently.
- WebQAPairs
  It includes Q-A pairs which are either crawled from QA websites or labeled by Microsoft.

Note that, testing questions and their corresponding golden answers have been consciously selected and labeled, 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.

3. Evaluation Metric

The quality of different QA systems are measured by three metrics described below:

- **Mean Reciprocal Rank (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 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**

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

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

[1] Currently, we provide auxiliary data for the English open domain QA task only.

- **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.