NLPCC-ICCPOL 2016 Shared Task Guideline:

Open Domain Question Answering

1. Task Description

The question answering (or QA) task targets on building systems that can answer natural language questions. In NLPCC 2016, we call the *Open Domain Question Answering* shared task, which includes the following two QA sub-tasks for **Chinese** language:

1) **Knowledge-based QA (or KBQA) task**
   When predicting answers to each question, a KBQA system built by each participating team IS LIMITED TO select entities as answers from a given knowledge base (KB). We will provide a Chinese knowledge base, which includes knowledge triples crawled from web. Other resources are allowed to use to train necessary models, such as entity linking, semantic parsing, and etc., but answer entities should come from the provided KB only.
   We will release a runnable KBQA system this year!

2) **Document-based QA (or DBQA) task**
   When predicting answers to each question, a DBQA system built by each participating team IS LIMITED TO select sentences as answers from the question's given document. We will provide a document for each question, which consists of a set of document sentences. Other resources are allowed to use to train necessary models, such as sentence matching model, but answer sentences should come from the provided documents only.
   We will release a runnable DBQA system this year!

We try our best to attract teams to participate in this year's QA task, including labeling new data sets and releasing baseline systems. We hope this can provide more benchmark data for QA research, and encourage more QA researchers to share their experiences, techniques, and progress.

2. Data Description

For KBQA task, we will provide a training set and a testing set. An example in training set is given below, which describes the data format:

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

In training set, both questions and their golden answers will be provided. If multiple answers exist, they will be separated by the symbol '\t'.

In testing set, only questions will be provided. We have labeled golden answers to questions,

but this information will NOT be provided to participants, until the entire evaluation is finished.

The format of the submission result should follow the same format above. If no answer can be generated for a given question, just set the value of **<answer id="1">** to an empty string.

For DBQA task, we will provide a training set and a testing set. An example in training set is given below, which describes the data format:

```
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
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
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
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
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
俄罗斯贝加尔湖的面积有多大？ \t 贝加尔湖长 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
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
```

In training set, questions (in the 1st column), document sentences (in the 2nd column), and their answer annotations (in the 3rd column) will be provided. If a document sentence is the correct answer of the question, its answer annotation will be 1, otherwise its answer annotation will be 0. The three columns will be separated by the symbol '\t'.

In testing set, only questions and their document sentences will be provided. We have labeled golden answer annotations to questions, but this information will NOT be provided to participants, until the entire evaluation is finished.

The submission result file should follow the format below: **each line contains a score only**, which denotes the relevance score between a question and a document sentence at the same line. These scores will be used to rank all answer sentences of a given question by the evaluation toolkit. Please carefully check that, **the number of lines in the submission result file should be the same to the number of lines in the testing set file**.

```
0.2343556
0.3434554
0.5634232
0.2324467
0.1283477
1.2384834
0.4754545
```

3 Evaluation Metric

The quality of a KBQA system will be evaluated by **MRR**, **Accuracy@N**, and **Averaged F1**.
The quality of a DBQA system will be evaluated by **MRR** and **MAP**.

- **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 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.

- **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.

- **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.

- **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.