NLPCC 2014 Shared Tasks Guidelines

Chinese Entity Linking

1. Overview

The main goal for the Entity Linking Evaluation at NLPCC 2014 is to evaluate the current advance of techniques in aligning named entities from various text resources to entities in a reference Chinese knowledge base.

2. The Task

In this Entity Linking task, each query consists of a name string, a background document and a pair of UTF-8 character offsets indicating the start and end positions of the string in the background document. An EL system is expected to output the ID of a knowledge base entry which the query name string should refer to, or a NIL tag if such knowledge base entries do not exist. The reference knowledge base used in this task is built from the InfoBoxes of the Chinese part of Wikipedia dumps in 2013.

3. Data

The reference knowledge base used in this evaluation includes about 400,000 entities based on InfoBoxes from a 2013 dump of Chinese Wikipedia. Each entity in this knowledge base will include a name string, a KB entry ID, and a set of assertions in the form of <subject, predicate, object>. This reference knowledge base will be available along with the training dataset in late April.

The background document collection used in this task will include TWO different types of text resources in Chinese, micro-blog messages (from Sina Weibo) and news articles. Entities will generally occur in multiple queries with different name variants or across different background documents. All datasets will be in XML format with UTF-8 coding. An example query is:

```
<weibo id="1">
        <content >北京时间 3 月 12 日，2013 亚冠联赛小组赛第二轮，广州恒大足球俱乐部客场挑战全北现代，广州恒大首发已经公布。</content>
    <name id="1" >广州恒大足球俱乐部</name>
            <startoffset id="1" >25</startoffset>
            <endoffset id="1" >34</endoffset>
```

```
                    <kb id=”1” >KBxxxx<kb>
          <name id=”2” >全北现代</name>
                    <startoffset id=”2” >38</startoffset>
                    <endoffset id=”2” >42</endoffset>
                    <kb id=”2” >KByyyy<kb>
          <name id=”3” >广州恒大</name>
                    <startoffset id=”3” >43</startoffset>
                    <endoffset id=”3” >47</endoffset>
                    <kb id=”3” >KBxxxx<kb>
</weibo>
```

In this example, the two name strings “广州恒大足球俱乐部” and “广州恒大” should refer to the same entity EN:WIKI:广州恒大足球俱乐部 in the reference knowledge base, while the name string “全北现代”refers to the entity EN:WIKI:全北现代汽车足球俱乐部 in the knowledge base.

4. **Scoring Metric**

For a set of query name strings with background documents, an EL system is required to judge whether each query can be linked to any entry in the reference knowledge base. We will apply the micro-averaged accuracy across name strings to evaluate the performance of a system. For example, there are in total 5 name strings in the test data and one of our teams has submitted the following system output:

Test Instances	System Output	Gold-Standard	Result
wb1_n1	KB1000	KB1000	Correct
wb2_n1	KB1222	KB1222	Correct
wb2_n2	KB2000	KB2000	Correct
wb3_n1	KB3111	KB3111	Correct
wb3_n2	NIL	KB3111	Wrong

Then the averaged accuracy for this submission will be 4/5=0.8

5. Submission

Only one submission file is allowed for each team, which should be formulated as the following

form:

id	system-id	doc-id	name-id	KB-id
1	TeamABC	weibo-1	1	KBWKxxxxx
2	TeamABC	weibo-2	1	KBWKxxxxx
3	TeamABC	weibo-4	2	NIL

…

Each row will correspond to a query name string, with fields corresponding to the result id, system id, background document id, string id and the entity id in the KB, separated by the \tab symbol.