NLPCC 2014 Shared Tasks Guidelines

Cross-Lingual Knowledge Linking

1. Introduction

Wikipedia is a large-scale multilingual knowledge base; articles describing the same subjects in different languages are connected by cross-lingual links in Wikipedia. For example the article *Beijing*[1] in English is linked to its Chinese version *北京*[2]. These cross-lingual links benefit many NLP applications such as cross-lingual information retrieval, cross-lingual semantic relatedness computation and machine translation. Because cross-lingual links are created manually by users of Wikipedia, some newly created articles or articles with less editors may have no cross-lingual links to its equivalent articles in other languages. Therefore cross-lingual knowledge linking can automatically enrich the cross-lingual links in Wikipedia.

2. Task Definition

Given the datasets of English Wikipedia and Chinese Wikipedia, a set of known cross-lingual links, the task is to discover the equivalent English articles for a set of specified Chinese articles.

The datasets of both English Wikipedia and Chinese Wikipedia are in XML format. Each dataset contains a list of articles presented in the following format:

```
<article title="…">
    <abstract>…</abstract>
    <infobox>att1=value1;att2=value2;…</infobox>
    <outlinks>link1;link2;…</outlinks>
    <category>category1;category2;…</category>
    <redirection>…</redirection>
</article>
```

[1] http://en.wikipedia.org/wiki/Beijing
[2] http://zh.wikipedia.org/wiki/北京

- **title**: each article has unique title within a dataset.
- **abstract**: the first paragraph of an article summarizing its most important information.
- **infobox**: a list of attribute-value pairs that record important facts about the subject of the current article.
- **outlinks**: a list of articles' titles that are linked by the current article; the outlinks are listed in the order as they occurred in the article.
- **category:** a list of categories to which the current article belongs to.
- **redirection:** if this article page is a redirect page in Wikipedia, redirection is the article to which this article redirects.

The above fields are not necessarily all contained in each article. For example, if one article has no infobox in its page, there will be no <infobox>…</infobox> xml tags for this article in the xml file.

Cross-lingual links are given in a CSV (Comma Separated Value) file, each line specifies one cross-lingual links by using titles of two articles.

4. Evaluation Metrics

Precision, recall and F1-measure will be used as evaluation metrics in this task.

5. Results Submission

The results of cross-lingual knowledge linking between Chinese Wikipedia and English Wikipedia should be recorded in a single CSV file. Each line in the file corresponds to one cross-lingual link, which should include the system name, the Chinese article title and the predicted English article title.