Crosslink Evaluation @NTCIR 9

1. **Assessment and Evaluation**

## Assessment

There will be two types of assessments: automatic assessment using the Wikipedia ground truth (existing cross-lingual links); and manual assessment done by human assessors. With the results from the above assessments, the performance of cross-lingual link discovery system then can be evaluated using two sets of *qrel*s.

### Automatic Assessment

The set of links used as Wikipedia ground truth is derived from the existing links in the topics and their counterparts in the target corpus. For instance, if the English topic page is “Solar Eclipse” then we define the ground truth set of Chinese links as the set of links out of the Chinese Solar Eclipse page日食to other pages in the Chinese collection. Similarly, if any English Wikipedia page linked by the “Solar Eclipse” English page has a counterpart in the Chinese Wikipedia, such a link also becomes part of the ground truth. For the purpose of evaluation we make the assumption that a good CLLD system will be able to find the same set of Chinese language links starting from the orphaned English text. This may not be very precise—for instance the two pages may not necessarily be exact translations of each other. However, this is likely to be sufficient to provide a good set of useful links.

### Manual Assessment

For the manual assessment, all submissions will be pooled and a GUI tool for efficient assessment will be used. In manual assessment, either the anchor candidate or the target link could be identified relevant (or non-relevant). Once the anchor candidate is assessed as non-relevant, all anchors and associated links inside this anchor will become non-relevant.

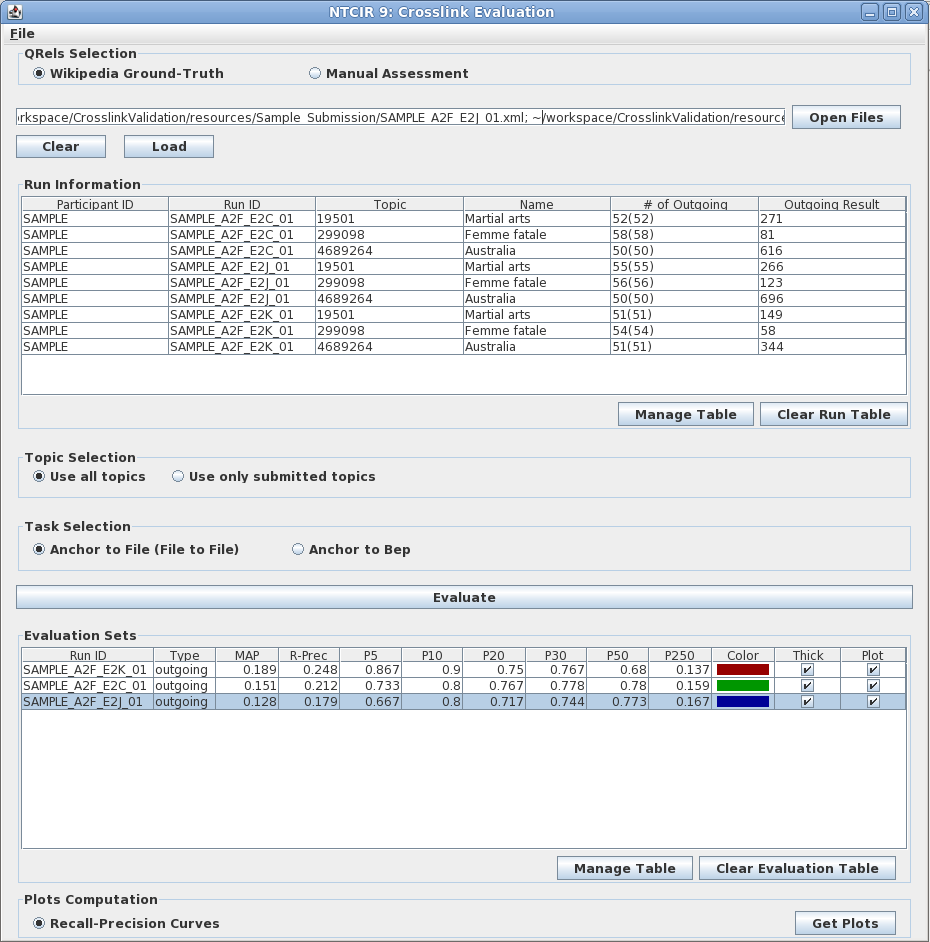
## Evaluation

### Metrics

As with other traditional information retrieval evaluations, *Precision*, *Recall* and *Mean Average Precision* (MAP) metrics are used to quantify the performance of the CLLD system.

1. **Using the Evaluation Software**

## Evaluation Software GUI



The evaluation result for each run including the score calculated using different metrices

The evaluation type, normally the Wikipedia ground-truth only allows file-to-file level evaluation

The loaded run information including # of links, topic id and name for each topic

The assessment type