Crosslink Evaluation @NTCIR 9

Version: 1.0

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 different 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

### Evaluation Type

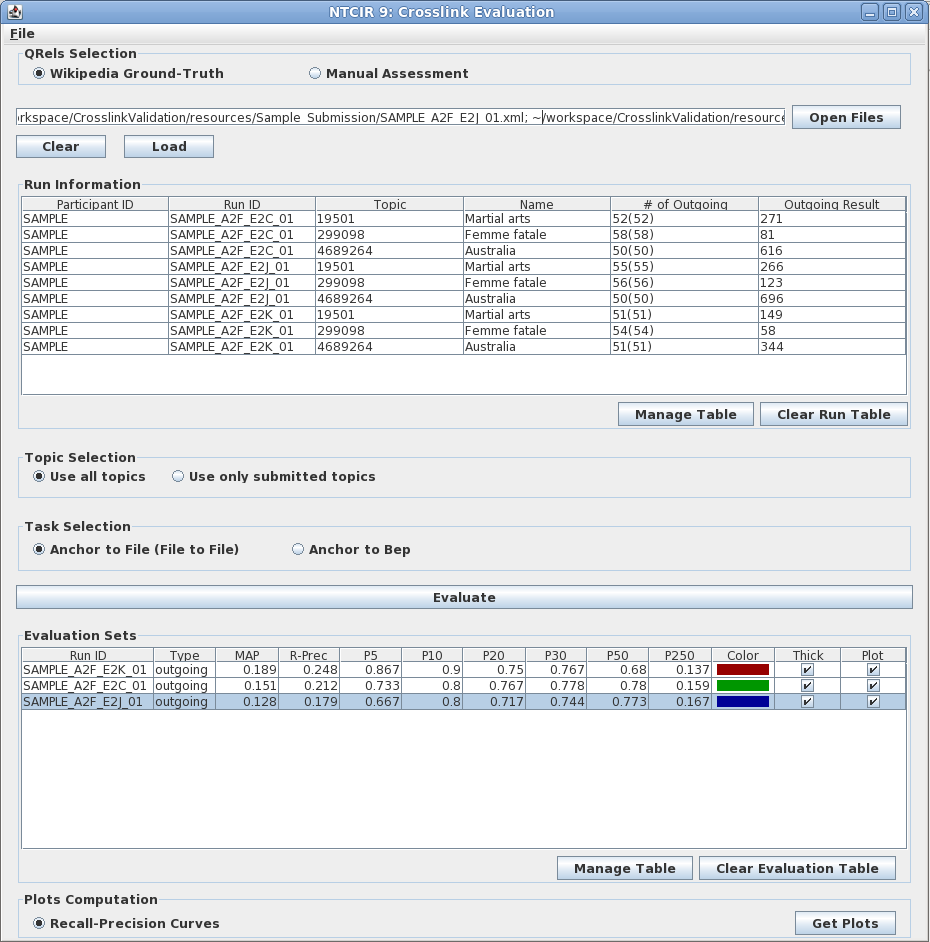
If

### 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 metrics

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

Figure 1 Evaluation Tool GUI

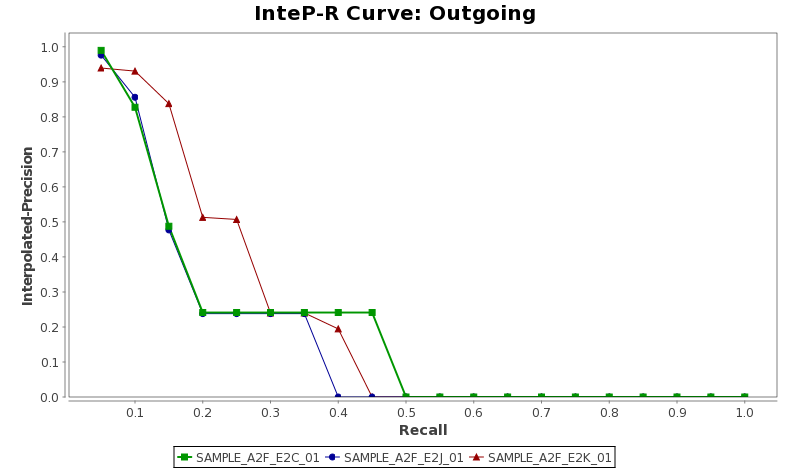


Figure 2 the plot of interpolated Precision - Recall curve

## Lauching the program

* On Unix-like system (e.g. Linux)

$sh run.sh

* On Windows

Double click on the run.bat or execute it in the command line console.

## Get Started

1. Select the *qrel* set (result from the different type of assessment) for evaluation. There are two options: Wikipedia ground-truth and manual assessment result.
2. Select the submission file(s) that you would like to evaluate by clicking on the “Open Files” button.
3. After the submission files are chosen, click on the “Load” button, the run information including number of links, topic id and name for each topic will be displayed in the “Run Information” table.
4. When Wikipedia ground-truth *qrel* is used for evaluation, only one type of evaluations is available. The score for each run will be calculated based on file-to-file level which means that whether anchors are correctly specified is not considered.

If the result of manual assessment is chosen for evaluation, system performance can be either evaluated in file-to-file level or anchor-to-bep level.

1. Then click on the “Evaluate” button, scores of different metrics will be showed in the “Evaluation Sets” table.
2. (Optional) By clicking on the “Get Plots” Button, a graphic of interpolated precision-recall curve for the selected run(s) will be displayed. So with the curve, the quality of different runs or the performance of different systems can be compared easily.

Note:

1. All the tables in the GUI can be exported in CSV format.
2. The popup window for interpolated precision-recall curve can be either saved or printed for your own reference.