Overview of the NTCIR-9 Crosslink Task: Cross-lingual Link Discovery

Ling-Xiang Tang[1], Shlomo Geva[1], Andrew Trotman[2], Yue Xu[1], Kelly Y. Itakura[1]

[1]Faculty of Science and Technology,

Queensland University of Technology,
Brisbane, Australia

{l4.tang, s.geva, yue.xu, kelly.itakura}@qut.edu.au

[2]Department of Computer Science,
University of Otago,
Dunedin, New Zealand

andrew@cs.otago.ac.nz

ABSTRACT

This paper presents an overview of NTCIR-9 Cross-lingual Link Discovery (Crosslink) task. The overview includes: the motivation of cross-lingual link discovery; the Crosslink task definition; the run submission specification; the assessment and evaluation framework; the evaluation metrics; and the evaluation results of submitted runs.

Cross-lingual link discovery (CLLD) is a way of automatically finding potential links between documents in different languages. The goal of this task is to create a reusable resource for evaluating automated CLLD approaches. The results of this research can be used in building and refining systems for automated link discovery. The task is focused on linking between English source documents and Chinese, Korean, and Japanese target documents.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *text analysis*.

I.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *linguistic processing*.

General Terms

Experimentation.

Keywords

Wikipedia, Cross-lingual Link Discovery, Anchor Identification, Link Recommendation, Validation Tool, Assessment Tool, Evaluation Tool, Evaluation Metrics.

1. INTRODUCTION

Cross-lingual link discovery (CLLD) is a way of automatically finding potential links between documents in different languages. It is not directly related to traditional cross-lingual information retrieval (CLIR). While CLIR can be viewed as a process of creating a virtual link between the provided cross-lingual query and the retrieved documents, CLLD actively recommends a set of meaningful anchors in the source document and uses them as queries (with the contextual information) from the article to establish links with documents in other languages.

Wikipedia is an online multilingual encyclopaedia that contains a very large number of articles covering most written languages and so it includes extensive hypertext links between documents of same language for easy reading and referencing. However, the pages in different languages are rarely linked except for the cross-lingual link between pages about the same subject. This could pose serious difficulties to users who try to seek information or knowledge from different lingual sources. Figure 1 shows a snippet of *Martial Art* Wikipedia article in which anchors are only linked to related English articles about different types of martial arts; direct links to other related Chinese/Japanese/Korean articles do not exist in Wikipedia.

Information flow could be easily blocked between articles of different languages in knowledge sharing due to the language barrier. To tackle this problem, cross-lingual link discovery aims to break this language barrier. With CLLD users are able to discover documents in languages which they either are familiar with, or which have a richer set of documents than in their language of choice.

For English there are several link discovery tools, which assist topic curators in discovering prospective anchors and targets for a given document. No such tools yet exist, that support the cross linking of documents from multiple languages. This task aims to incubate the technologies assisting CLLD and enhance the user experience in viewing or editing documents in cross-lingual manner.

The remainder of this paper is organized as follows: First, we define cross-lingual link discovery task in Section 2. The submission speciation is outlined in Section 3. The assessment and evaluation frame work is discussed in Section 4. Evaluation results are given in Section 5. We then conclude in Section 6.

2. CROSSLINK TASK

2.1 Task Definition

Generally, the link between documents can be classified as either outgoing or incoming, but this task is mainly focused on

NTCIR-9 Workshop Meeting, 2012, Tokyo, Japan

the outgoing link starting from English source documents and pointing to Chinese, Korean, or Japanese target documents. The CLLD task is comprised of following three subtasks:

- English to Chinese CLLD (E2C)
- English to Japanese CLLD (E2J)
- English to Korean CLLD (E2K)

For each subtask, English documents are provided as topics; and for each topic it is required to identify prospective anchors and recommend links for them in the CJK document collections.

2.1.1 Topic

The English topics are actual Wikipedia articles in xml format consisting of rich structured information. All the existing links are removed.

2.1.2 Anchor

An anchor is a piece of text that is relevant to the topic and worthy of being linked to other documents for further reading. To submit a run for a given task, participants are required to choose the most suitable anchors in English topic documents, and for each anchor identify the most relevant documents in the target language corpus. Up to 250 anchors are allowed for each topic.

2.1.3 Cross-lingual Link

The target links of recommended anchors have to be chosen from the provided standard CJK document collections. Up to 5 targets are allowed for each anchor. So there is a total of up to 1250 outgoing links per topic.

A cross-lingual link can be symbolized as: $a \rightarrow d$, where a is the anchor, d is the cross-lingual CJK target document.

2.2 Document Collections

The training and test collections for the three subtasks are exactly the same. The collections are formed of search engine friendly xml files created from Wikipedia mysql database dumps taken on June 2010. The original article text containing unique Wikipedia mark-ups is converted into XML using the YAWN system [1]. The details of the collections are given in Table 1.

Table 1. CJK Wikipedia document collections

Language	# doc	Size	Dump Date
Chinese	318736	2.7G	27/06/2010
Japanese	716,088	6.1G	24/06/2010
Korean	201596	1.2G	28/06/2010

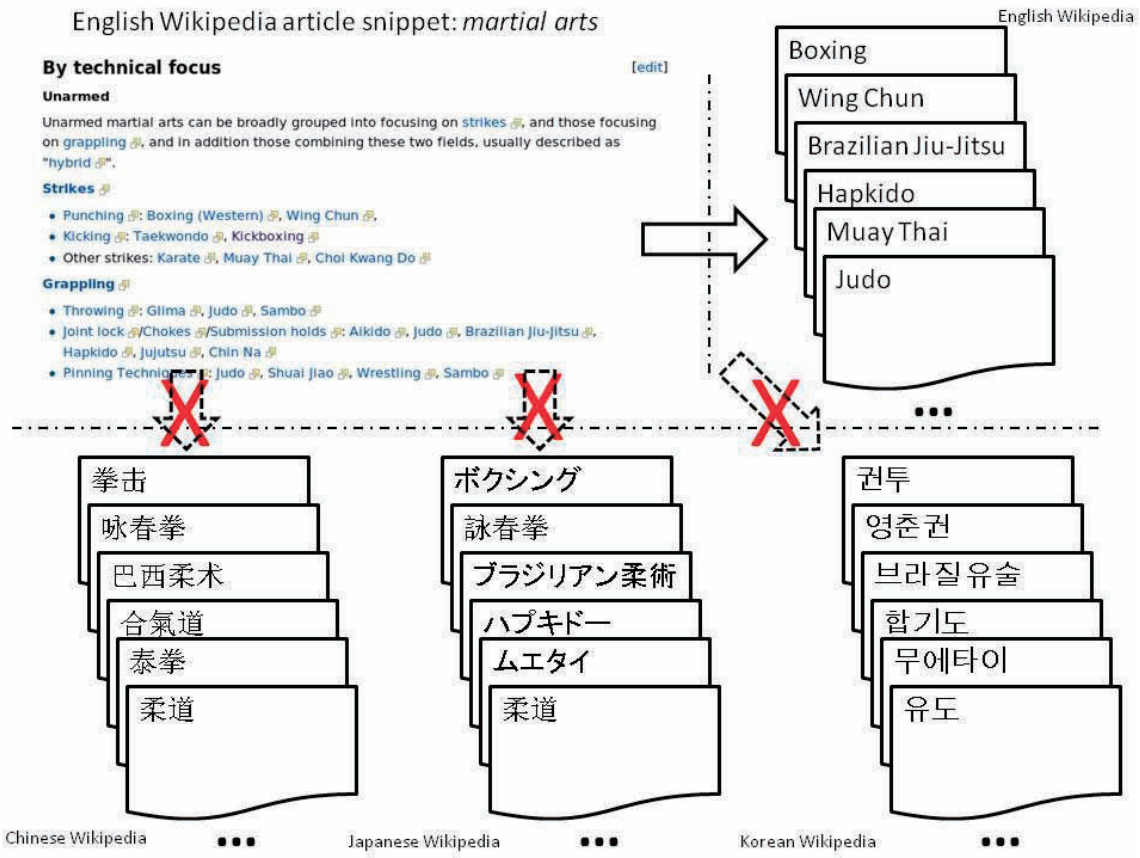


Figure 1. Cross-lingual Linking in Wikipedia

2.3 Topics

Only three topics are used for system training. The training topic details are given in Table 2.

Table 2. Training topics

#	Title	ID
1	Australia	4689264
2	Femme fatale	299098
3	Martial arts	19501

A set of 25 articles are randomly chosen from the English Wikipedia and used as test topics. All test topics are prepared in a form of XML file without *link* tags, which means the previously existing links in topics are removed.

3. SUBMISSION

3.1 Rules

Topic files including their CJK counterparts must be removed from document collections either physically or virtually.

Special case links (numbers, years, dates, and century links) are not recommended for being included in the runs. There were several reasons for this decision. First, they are special cases that can be handled uniformly in addition to the algorithms we are comparing. Second, Wikipedia has special rules on their use; chronological items should not be linked "unless their content is germane" [1]. If those links are included in the submissions, they will be rejected and not considered in either assessment or evaluation.

3.2 Run Specification

XML is used for formatting run results. The specification of submission is similar with that of the INEX Link-the-Wiki task[2].

3.2.1 Submission XML File DTD

The document type declaration (DTD) for experimental run XML file is given in Table 3. Only submissions complying with this DTD can be recognised by the assessment and evaluation tools.

The root element *crosslink-submission* should contain information about participant's ID, run ID (which should include university affiliation), the task which should be either **A2F** or **A2B** and the default target language should be given in language abbreviation which could be either **zh**, **ja**, or **ko**. The linking algorithm should be described in *description* node. The *collections* element contains a list of document collections used in the run. Generally, the collection element should contain text from one of the following: ***Chinese Wikipedia***, ***Japanese Wikipedia*** or ***Korean Wikipedia***. Each topic should be contained in a topic element which should contain an anchor element for each anchor-text that should be linked. Each *anchor* element should include *offset*, *length* and *name* attributes for detailed information of the recommended anchor, and should also have one or more *tofile* sub-elements with the target document ID contained within them. The *tofile* element should contain following information: language id, title and bep (specified in *lang*, *title* and *bep_offset* attributes separately) of the linked document.

Table 3. Submission XML File DTD

```
<!ELEMENT crosslink-submission (details, description, collections,
topic+)>
<!ATTLIST crosslink-submission
  participant-id CDATA #REQUIRED
  run-id CDATA #REQUIRED
  task (A2F| A2B) #REQUIRED
  default_lang (zh|ja|ko) ) #REQUIRED
>
<!ELEMENT details (machine, time)>
<!ELEMENT machine (cpu, speed, cores, hyperthreads, memory)>
<!ELEMENT cpu (#PCDATA)>
<!ELEMENT speed (#PCDATA)>
<!ELEMENT cores (#PCDATA)>
<!ELEMENT hyperthreads (#PCDATA)>
<!ELEMENT memory (#PCDATA)>
<!ELEMENT time (#PCDATA)>
<!ELEMENT description (#PCDATA)>
<!ELEMENT collections (collection+)>
<!ELEMENT collection (#PCDATA)>
<!ELEMENT topic (outgoing)>
<!ATTLIST topic
  file CDATA #REQUIRED
  name CDATA #REQUIRED
>

<!ELEMENT outgoing (anchor+)>

<!ELEMENT anchor (tofile+)>
<!ATTLIST anchor
  name CDATA #REQUIRED
  offset CDATA #REQUIRED
  length CDATA #REQUIRED
>
<!ELEMENT tofile (#PCDATA)>
<!ATTLIST tofile
  bep_offset CDATA #REQUIRED
  lang (zh|ja|ko)#REQUIRED
  title CDATA #REQUIRED
>
```

3.2.2 Anchor

The position (zero-based offset) of an anchor is calculated by counting the number of bytes from the beginning of the provided topic file in binary form. Offset of an anchor could be different and will not be properly recognized by the assessment and evaluation framework if it is calculated by counting the number of characters because a character can use more than one byte of storage size due to various encoding schemes. Similarly, the length of an anchor is the number of bytes occupied by that anchor text.

If an anchor has a name that doesn't match the text in the topic with the given offset and length, it will be discarded in the pooling,

As topics are presented in XML format, recommend anchors can be overlapped with XML tags. However, the "text" of anchors

[1] http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style_(linking)#Specific_cases

included in submissions must not contain those XML tags or special characters due to the XML well-formedness issue.

If anchor text and title names contain special characters such as "&", "<", """, they should be replaced with the correspondent predefined XML entities.

For anchors containing XML tags or entities, their anchor length must also include the length of the tags or entities.

For example, if an anchor "*A Sample <it>Anchor*" is found, anchor should be present as following:

```
<anchor offset="768" length="19" name="A Sample Anchor">
```

After removing XML tags, the anchor name in the above example should be "A Sample Anchor" with length of 19.

If anchors contain incomplete XML tags, they will be discarded in the pooling. For example, "*A Sample <i*" is an incorrect anchor.

3.3 Submission Validation Tool

Unlike other information retrieval tasks, in cross-lingual link discovery recommended anchors have to have correct offsets, and then the anchors can be displayed properly in the manual assessment tool for assessment. So it is important all anchor offsets are correctly specified. A submission validation tool is provided for this purpose. The tool GUI snapshot is given in Figure 2. With the validation tool, the recommended anchors can be self verified by the participants.

3.4 Participant Submissions

At the end of experimentation season, in total 57 runs from 11 teams were received. The groups and their affiliation names are given in Table 4.

Table 4. Crosslink Participants

GROUP	ORGANISATION
DUIIS	Daegu University
HITS	Heidelberg Institute for Theoretical Studies
IISR	Yuan Ze University
ISTIC	Institute of Scientific and Technical Information of China
KMI	The Open University
kslab_nut	Nagaoka University of Technology
KSLP	Kyungsung University
nthuisa	Academia Sinica
QUT	Queensland University of Technology
UKP	TU Darmstadt
WUST	Wuhan University of Science and Technology

The run statistics of each sub-task is given in Table 5.

Table 5. Submission statistics of participants

Group	En-2-Zh	En-2-Ja	En-2-Ko
DUIIS	0	0	2
HITS	3	3	3
IISR	0	0	5
ISTIC	1	0	0
KMI	4	0	0
kslab_nut	0	1	0
KSLP	0	0	5
nthuisa	3	0	0
QUT	5	2	1
UKP	5	5	5
WUST	4	0	0
Sub-total	25	11	21
Total	57		

The system descriptions of each run are outlined in Table 25, 26, 27. These texts, which are enclosed in "description" tag, are extracted from the submitted run files.

4. 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. For the latter, 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 irrelevant. If an anchor candidate is assessed as irrelevant, all anchors and their associated links will become irrelevant. With the *qrel* generated from the assessment results, the performance of cross-lingual link discovery system then can be evaluated using *Precision*, *Recall* and *Mean Average Precision* (MAP) metrics.

4.1 Wikipedia Ground-Truth

The set of links used as the ground truth (GT) is derived from the existing links in the topics, and their counterparts in the target corpus.

The ground truth is the combination of links from both corpora. For instance, if an English topic page is "Martial Art" then we define the ground truth set of Chinese links as the set of links out of the Chinese / Japanese / Korean "Martial Art" article (Chinese: 武術; Japanese: 格闘技; Korean: 무예). Similarly, for all links out of the English "Martial Art" article, if the target page has a counterpart in CJK Wikipedia then the link also becomes part of the ground truth.

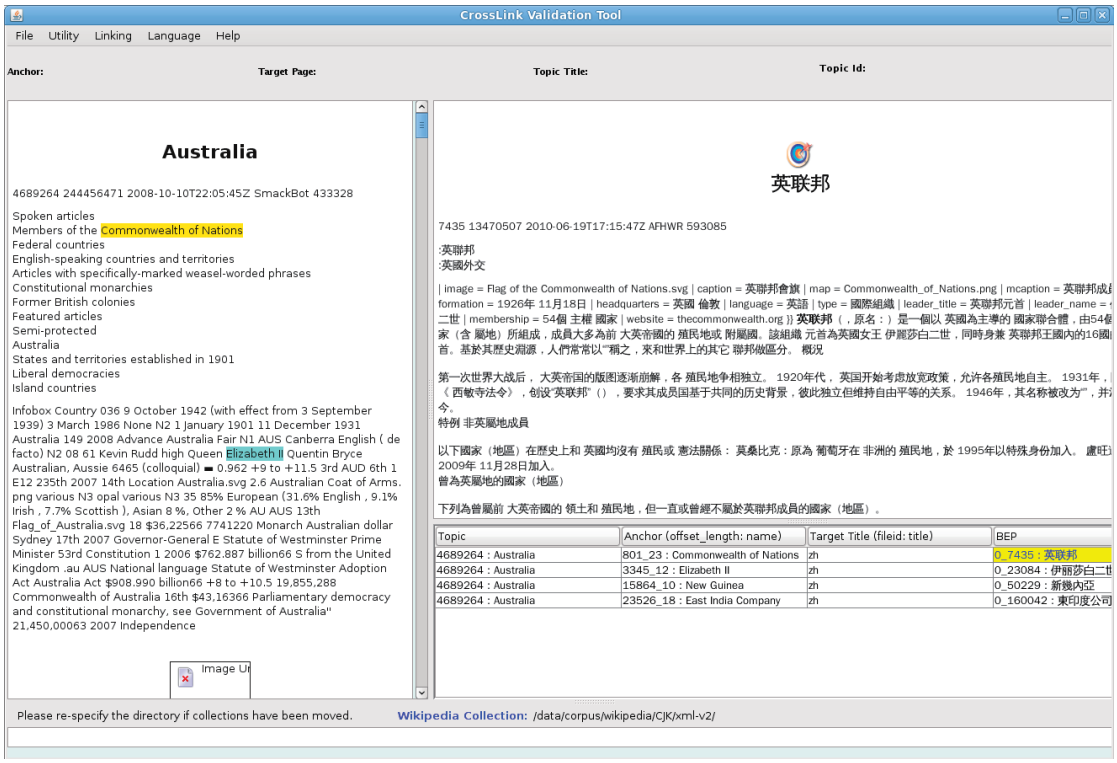


Figure 2. Crosslink Run Self Validation Tool

We accept that the ground truth may not be complete. It may also contain links from the English version to which there is no equivalent text to link from in the CJK versions. It may also not contain the kinds of links that users click on. We do not believe that this will adversely affect the relative rank order of different CLLD systems.

Evaluation using ground-truth could be biased towards links already in Wikipedia. Huang et al[3] suggested that manual topical assessment of machine generated links could result in substantially different relevant link set.

4.2 Manual Assessment

4.2.1 Human-in-the-loop Assessment

The Wikipedia ground truth is easy to get, but not necessarily reflecting user preferences optimally. Much of it is automatically generated. Even manually generated Wikipedia links, by the content contributors, may be disliked by most assessors. The NTCIR assessors will make the decision about the quality of both anchors and targets.

For manual assessment we will involve human assessors. First, all the prospective anchors and the corresponding links will be pooled. Then an assessor will inspect each anchor and its corresponding prospective links, accepting or rejecting them one by one. This is not dissimilar to the assessment approaches used in CLIR evaluations. Given an anchor and its context (our "query"), the assessor judges the relevance of the target document. The assessment can be done by designated independent assessors.

4.2.2 Assessment Tool

The design of the assessment tool is similar with the submission validation tool. The GUI snapshots of English-2-Chinese, English-2-Japanese and English-2-Korean assessment tool are given in Figure 3, 4, and 5 respectively. There isn't any functionality difference in assessment of each language subtask.

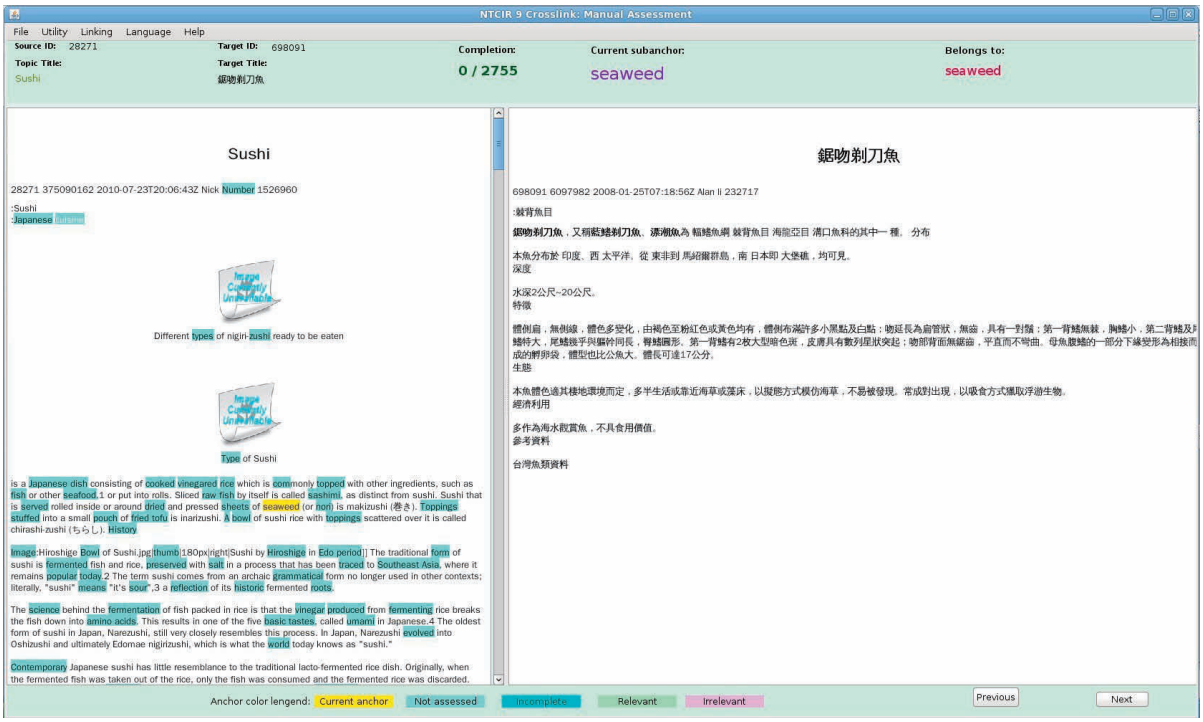


Figure 3. Crosslink Manual Assessment Tool (English-to-Chinese)

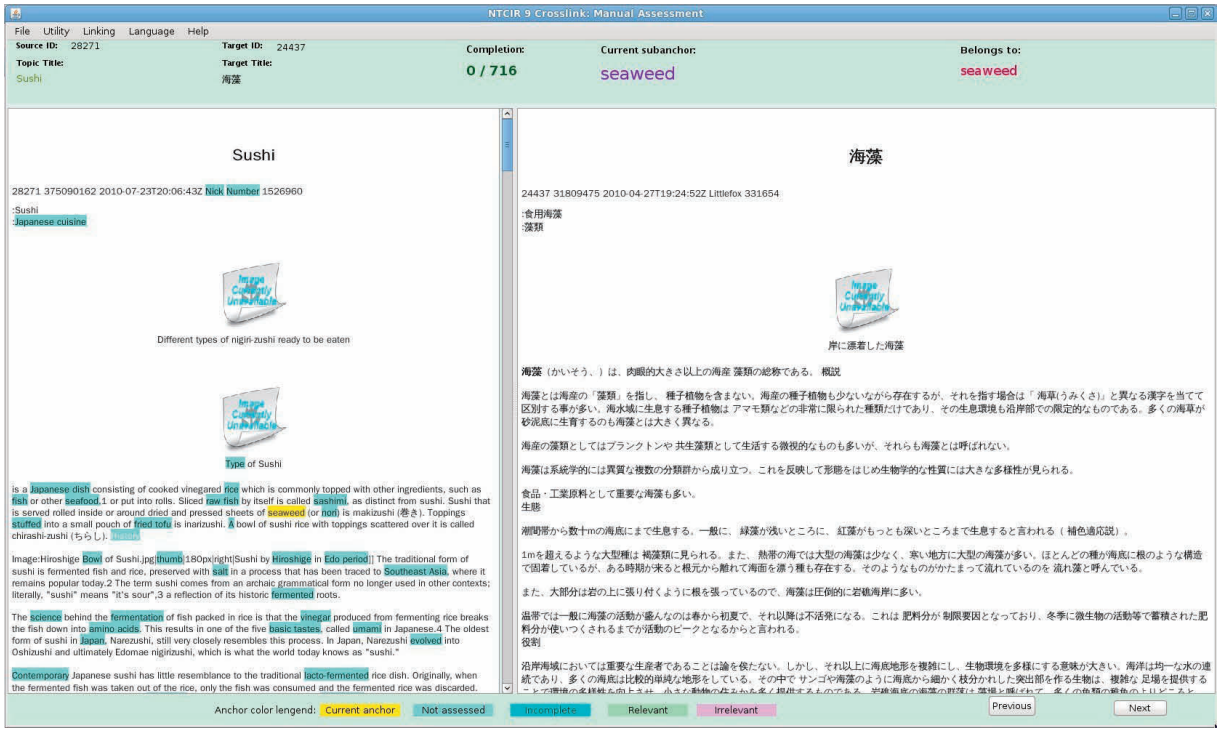


Figure 4. Crosslink Manual Assessment Tool (English-to-Japanese)

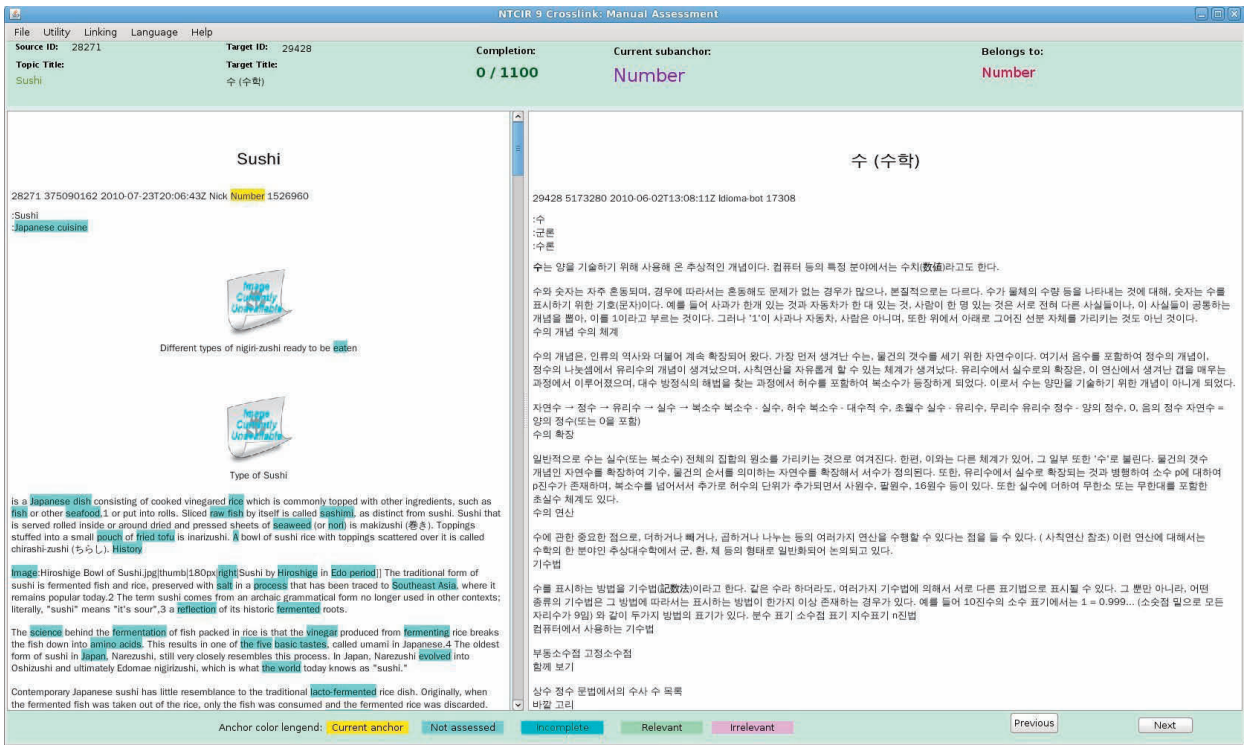


Figure 5. Crosslink Manual Assessment Tool (English-to-Korean)

5. EVALUATION

5.1 Evaluation Methods

The performance of the different systems is evaluated in both file-to-file (F2F) and anchor-to-file (A2F) levels. In F2F evaluation, performance is measured simply on the basis of "this file should link to that file" regardless of the anchors. In A2F evaluation, the correctness of anchors must however be considered.

With automatic assessment results we can have only F2F evaluation, because evaluating a run using Wikipedia ground-truth is difficult for several reasons. An anchor can occur multiple times in a document in subtly different linguistic forms. It is unreasonable to score multiple identical links and also unreasonable not to score different linguistic variants. The best approach to measuring this imprecision is unclear and has been studied at the INEX Link Discovery Track [4-6] where it changed from year to year.

In A2F evaluation, an anchor can be either relevant or irrelevant to the topic; if an anchor is considered irrelevant, then its entire associated links are irrelevant. With the manual assessment results, correctness of all pooled anchors and relevancy of the associated links can be judged. So there could be two sets of evaluation: one in file-2-file level; the other in anchor-to-file level.

5.2 Metrics

In this task, *Precision-at-N, R-Prec,* and *Mean Average Precision* (MAP) are the main metrics used to quantify the performance of the different CLLD systems. As other traditional information retrieval evaluation tasks, *precision* and *recall* are the two underneath key metrics to measure run performance on each topic. But it needs to be noted that precision and recall are computed differently in each evaluation (file-to-file or anchor-to-file).

5.2.1 Precision and Recall

File-to-File Evaluation

$$Precision_{f2f} = \frac{number\ \ of\ correct\ \ links}{number\ \ of\ identified\ \ links} \quad (4)$$

and,

$$Recall = \frac{number\ \ of\ correct\ \ links}{number\ \ of\ links\ in\ qrel} \quad (5)$$

The precision and recall are computed in ***link*** level for each topic.

Anchor-to-File Evaluation

For the anchor-to-file evaluation, we adopted a similar precision calculation formula as used in INEX 2009[7]. Both relevancies of anchor and link need to be considered. The score of anchor is defined as:

$$f_{anchor}(i) = \begin{cases} 1, & \text{if relevant with} \geq 1 \text{ relevant link(s)} \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

For example, if an anchor *i* itself in a topic is relevant, and it has at least one relevant link, then $f_{anchor}(i) = 1$.

Since we allow multiple targets (links) for each anchor, for each recommended link *j* if the link *j* is relevant to the anchor, then $f_{link}(j) = 1$.

$$f_{link}(j) = \begin{cases} 1, & \text{if relevant} \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

So the precision in anchor-to-file evaluation for a topic is then:

$$Precision_{a2f} = \frac{\sum_{i=1}^{n} (f_{anchor}(i)) \times \frac{\sum_{j=1}^{k} f_{link}(j)}{k_i}}{n} \quad (8)$$

Where *n* is the number of identified anchors; *k* is the number of returned links for the *anchor(i)* and k_i is the number of links recommended for this anchor.

Overall, the precision and recall are computed in ***anchor*** level for each topic.

5.2.2 System Evaluation Metrics

For both types of evaluation, *R-Prec,* and *MAP* are defined as:

$$MAP = \frac{\sum_{t=1}^{n} \frac{\sum_{k=1}^{m} P_{kt}}{m}}{n} \quad (9)$$

where *n* is the number of topics; *m* is the number of identified items (links or anchors); P_{kt} is the precision at top *K* items (links or anchors) for topic *t*.

$$R - Prec = \frac{\sum_{t=1}^{n} P_t @ R}{n} \quad (10)$$

where *n* is the number of topics; *m* is the number of identified items (links or anchors); P_{kt} is the precision at top *k* items (links or anchors) for topic *t*; $P_t @ R$ (= *number of correct items (links or anchors) / number of items (links or anchors) in qrel*) is the precision calculated using number of links / anchors in *qrel* as denominator for topic ***t.***

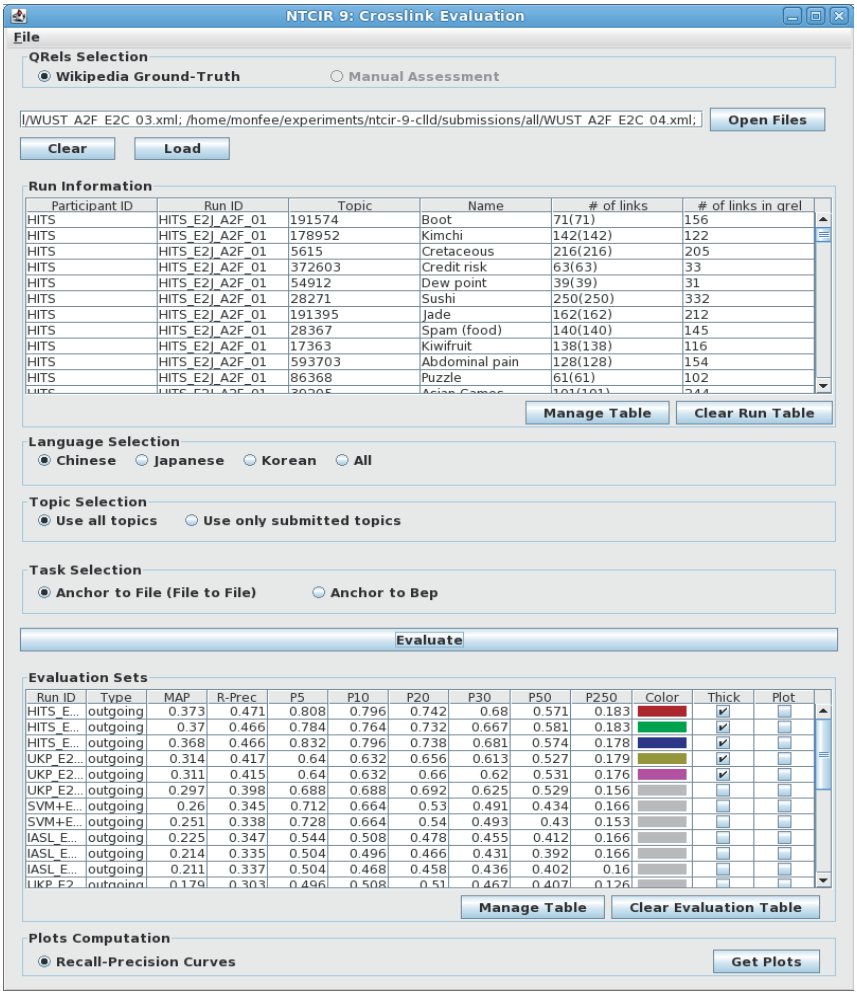


Figure 6. Crosslink Evaluation Tool

Similarly, *Precision-at-N* is computed using the average precision for all topics at the pre-defined top *N* items (links or anchors). In both types of evaluations (F2F and A2F), *N* is defined as 5, 10, 20, 30, 50, and 250 separately.

5.3 Evaluation Tool

A GUI program is employed for evaluation. It can be easily used to compute performance scores (*Precision-at-N, R-Prec,* and *MAP*) of different CLLD systems using *qrels* either from Wikipedia ground-truth or manual assessment. With this tool, a comparative plot of precision-recall curves for different systems can also be generated. The GUI snapshot of this tool is given in Figure 6.

6. EVALUATION RESULTS

Run performance is measured primarily using *MAP* metric. *Precision-at-5* and *R-Prec* are secondly evaluation metrics.

6.1 Evaluation with Wikipedia Ground-Truth

This section presents the evaluation results calculated against the *qrel* created from Wikipedia Ground-truth. The evaluation includes all 25 test topics.

Table 6 shows the scores of *MAP* and *R-Prec* of all runs in three subtasks: English-2-Chinese, English-2-Japanese, and English-2-Korean respectively. The table is sorted on MAP score. The ranking could be different but not much if R-Prec scores are used for sorting, particularly the ranking between different teams.

Table 7, 8, 9 show the evaluation results with Wikipedia ground truth using *Precision-at-N* metric for English-2-Chinese, English-2-Japanese, and English-2-Korean runs respectively.

Table 6. F2F evaluation results with Wikipedia ground truth: MAP, R-PREC

English-2-Chinese			English-2-Japanese			English-2-Korean		
Run-ID	**MAP**	**r-prec**	**Run-ID**	**MAP**	**r-prec**	**Run-ID**	MAP	r-prec
HITS_E2C_A2F_02	0.373	0.471	HITS_E2J_A2F_02	0.316	0.409	HITS_E2K_A2F_02	0.447	0.506
HITS_E2C_A2F_03	0.370	0.466	HITS_E2J_A2F_03	0.313	0.413	HITS_E2K_A2F_01	0.447	0.509
HITS_E2C_A2F_01	0.368	0.466	HITS_E2J_A2F_01	0.310	0.403	HITS_E2K_A2F_03	0.439	0.513
UKP_E2C_A2F_02	0.314	0.417	UKP_E2J_A2F_01	0.246	0.363	DUIIS_A2F_E2K_4Pre	0.370	0.442
UKP_E2C_A2F_01	0.311	0.415	UKP_E2J_A2F_02	0.242	0.361	DUIIS_A2F_E2K_4Rec	0.365	0.438
UKP_E2C_A2F_03	0.297	0.398	UKP_E2J_A2F_03	0.235	0.338	UKP_E2K_A2F_02	0.337	0.440
KMI_SVM_ESA_TERMDB	0.260	0.345	UKP_E2J_A2F_04	0.205	0.318	UKP_E2K_A2F_03	0.335	0.439
KMI_SVM_ESA	0.251	0.338	UKP_E2J_A2F_05	0.189	0.338	UKP_E2K_A2F_01	0.333	0.433
IASL_E2C_01	0.225	0.347	QUT_PNM_JA	0.076	0.143	KSLP_E2K_05	0.328	0.437
IASL_E2C_02	0.214	0.335	TRANSLITERATION_JA	0.047	0.145	KSLP_E2K_04	0.326	0.439
IASL_E2C_03	0.211	0.337	kslab_nut_A2F_E2J_01	0.041	0.084	KSLP_E2K_03	0.318	0.431
UKP_E2C_A2F_04	0.179	0.303				KSLP_E2K_02	0.316	0.427
QUT_LinkProb_ZH	0.179	0.244				UKP_E2K_A2F_04	0.287	0.405
UKP_E2C_A2F_05	0.178	0.321				KSLP_A2F_E2K_01	0.260	0.346
KMI_SVM_TERMDB	0.127	0.211				IISR_ko_title_to_key_name	0.173	0.260
WUST_A2F_E2C_03	0.108	0.207				IISR_kttkn_to_ztttzt	0.172	0.265
WUST_A2F_E2C_01	0.093	0.165				IISR_sttent_to_ztttzt_to_kttkn	0.164	0.263
WUST_A2F_E2C_02	0.089	0.163				UKP_E2K_A2F_05	0.158	0.296
QUT_PNM_ZH	0.088	0.166				IISR_sttent_to_kttkn	0.137	0.261
WUST_A2F_E2C_04	0.076	0.123				QUT_PNM_KO	0.122	0.208
QUT_LinkProbZh2_ZH	0.069	0.154				IISR_singular_term_to_en_title	0.118	0.159
QUT_LinkProbZh_ZH	0.059	0.148						
KMI_ESA_SVM_ESA discovery	0.041	0.120						
ISTIC_A2F_E2C_01	0.032	0.101						
QUT_LinkProbIR_ZH	0.023	0.067						

Table 7. F2F evaluation results with Wikipedia ground truth: *Precision-at-N* (English-2-Chinese)

Run-ID	P5	P10	P20	P30	P50	P250
HITS_E2C_A2F_01	0.832	0.796	0.738	0.681	0.574	0.178
HITS_E2C_A2F_02	0.808	0.796	0.742	0.680	0.571	0.183
HITS_E2C_A2F_03	0.784	0.764	0.732	0.667	0.581	0.183
QUT_LinkProb_ZH	0.776	0.588	0.480	0.404	0.319	0.132
KMI_SVM_ESA	0.728	0.664	0.540	0.493	0.430	0.153
KMI_SVM_ESA_TERMDB	0.712	0.664	0.530	0.491	0.434	0.166
UKP_E2C_A2F_03	0.688	0.688	0.692	0.625	0.529	0.156
UKP_E2C_A2F_02	0.640	0.632	0.656	0.613	0.527	0.179
UKP_E2C_A2F_01	0.640	0.632	0.660	0.620	0.531	0.176
KMI_SVM_TERMDB	0.624	0.552	0.454	0.383	0.302	0.078
QUT_PNM_ZH	0.592	0.472	0.362	0.307	0.242	0.064
WUST_A2F_E2C_03	0.576	0.492	0.406	0.360	0.285	0.077
WUST_A2F_E2C_01	0.576	0.496	0.406	0.353	0.264	0.060
WUST_A2F_E2C_02	0.552	0.480	0.394	0.327	0.247	0.061
IASL_E2C_01	0.544	0.508	0.478	0.455	0.412	0.166
IASL_E2C_03	0.504	0.468	0.458	0.436	0.402	0.160
IASL_E2C_02	0.504	0.496	0.466	0.431	0.392	0.166
UKP_E2C_A2F_04	0.496	0.508	0.510	0.467	0.407	0.126
WUST_A2F_E2C_04	0.496	0.424	0.344	0.304	0.231	0.048
QUT_LinkProbZh2_ZH	0.360	0.284	0.248	0.221	0.187	0.082
UKP_E2C_A2F_05	0.344	0.376	0.394	0.419	0.381	0.156
QUT_LinkProbZh_ZH	0.304	0.208	0.168	0.161	0.156	0.082
KMI_ESA_SVM_ESAdiscovery	0.264	0.240	0.186	0.165	0.138	0.044
QUT_LinkProbIR_ZH	0.184	0.160	0.118	0.109	0.084	0.044
ISTIC_A2F_E2C_01	0.168	0.140	0.138	0.132	0.113	0.046

Table 8. F2F evaluation results with Wikipedia ground truth: *Precision-at-N* (English-2-Japanese)

Run-ID	P5	P10	P20	P30	P50	P250
HITS_E2J_A2F_02	0.840	0.832	0.782	0.724	0.618	0.209
HITS_E2J_A2F_01	0.816	0.824	0.778	0.725	0.626	0.206
HITS_E2J_A2F_03	0.768	0.768	0.756	0.707	0.629	0.210
UKP_E2J_A2F_03	0.624	0.652	0.674	0.636	0.554	0.177
QUT_PNM_JA	0.624	0.504	0.394	0.333	0.262	0.079
UKP_E2J_A2F_04	0.568	0.584	0.622	0.587	0.514	0.168
UKP_E2J_A2F_01	0.544	0.604	0.646	0.639	0.546	0.202
UKP_E2J_A2F_02	0.520	0.584	0.620	0.619	0.534	0.205
kslab_nut_A2F_E2J_01	0.440	0.364	0.256	0.209	0.151	0.053
UKP_E2J_A2F_05	0.344	0.428	0.454	0.483	0.463	0.196
TRANSLITERATION_JA	0.160	0.136	0.126	0.139	0.152	0.099

Table 9. F2F evaluation results with Wikipedia ground truth: *Precision-at-N* (English-2-Korean)

Run-ID	P5	P10	P20	P30	P50	P250
HITS_E2K_A2F_01	0.848	0.764	0.720	0.625	0.520	0.148
HITS_E2K_A2F_02	0.840	0.764	0.712	0.616	0.518	0.151
DUIIS_A2F_E2K_4Pre	0.792	0.768	0.674	0.596	0.479	0.124
DUIIS_A2F_E2K_4Rec	0.784	0.760	0.664	0.583	0.474	0.126
HITS_E2K_A2F_03	0.744	0.764	0.678	0.612	0.521	0.151
KSLP_E2K_02	0.696	0.660	0.616	0.547	0.448	0.116
KSLP_E2K_05	0.680	0.684	0.610	0.544	0.452	0.123
KSLP_E2K_04	0.680	0.688	0.612	0.548	0.454	0.122
KSLP_E2K_03	0.680	0.652	0.622	0.552	0.450	0.117
IISR_singular_term_to_en_title	0.680	0.532	0.372	0.329	0.204	0.041
IISR_ko_title_to_key_name	0.656	0.584	0.498	0.424	0.318	0.066
UKP_E2K_A2F_03	0.648	0.648	0.608	0.552	0.466	0.127
KSLP_A2F_E2K_01	0.632	0.564	0.506	0.444	0.362	0.122
IISR_kttkn_to_ztttzt	0.600	0.588	0.502	0.424	0.322	0.067
UKP_E2K_A2F_04	0.568	0.564	0.558	0.507	0.431	0.120
UKP_E2K_A2F_02	0.560	0.580	0.588	0.531	0.463	0.146
UKP_E2K_A2F_01	0.560	0.600	0.590	0.533	0.461	0.143
QUT_PNM_KO	0.552	0.460	0.384	0.321	0.244	0.062
IISR_sttent_to_ztttzt_to_kttkn	0.544	0.528	0.484	0.413	0.316	0.066
IISR_sttent_to_kttkn	0.400	0.420	0.426	0.359	0.294	0.066
UKP_E2K_A2F_05	0.208	0.252	0.286	0.324	0.303	0.119

The runs with highest performance scores of all measures (*MAP, R-Prec, Precision-at-N*) in the file-to-file evaluation with Wikipedia ground-truth in all three subtasks were submitted by team HITS.

The top three teams of each measure are listed below:

- *English-2-Chinese*

MAP: HITS, UKP, KMI

R-Prec: HITS, UKP, KMI

Precision-at-5: HITS, QUT, KMI

- *English-2-Japanse*

MAP: HITS, UKP, QUT

R-Prec: HITS, UKP, QUT

Precision-at-5: HITS, UKP, QUT

- *English-2-Korean*

MAP: HITS, DUIIS, UKP

R-Prec: HITS, DUIIS, UKP

Precision-at-5: HITS, DUIIS, KSLP

6.2 Evaluation with Manual Assessment Result

This section presents the evaluation results using the *qrel* created from manual assessment results. The evaluation is given in both file-2-file and anchor-to-file level on all 25 test topics.

6.2.1 File-2-File Evaluation

Table 10 shows the scores of *MAP* and *R-Prec* of all runs in three subtasks. The table is sorted on MAP score. Table 11, 12, 13 show the evaluation results using *Precision-at-N* metric for English-2-Chinese, English-2-Japanese, and English-2-Korean runs respectively.

The top three teams of each measure are listed below:

- *English-2-Chinese*

MAP: UKP, KMI, HITS

R-Prec: UKP, KMI, HITS

Precision-at-5: QUT, HITS, KMI

- *English-2-Japanse*

MAP: HITS, UKP, QUT

R-Prec: HITS, UKP, QUT

Precision-at-5: HITS, UKP, QUT

Table 10. F2F evaluation with manual assessment results: MAP, R-PREC

English-2-Chinese			English-2-Japanese			English-2-Korean		
Run-ID	**MAP**	**r-prec**	**Run-ID**	**MAP**	**r-prec**	**Run-ID**	MAP	r-prec
UKP_E2C_A2F_02	0.308	0.429	HITS_E2J_A2F_03	0.451	0.513	UKP_E2K_A2F_02	0.376	0.522
UKP_E2C_A2F_01	0.306	0.424	HITS_E2J_A2F_02	0.435	0.499	UKP_E2K_A2F_01	0.369	0.518
KMI_SVM_ESA_TERMDB	0.258	0.393	HITS_E2J_A2F_01	0.434	0.501	UKP_E2K_A2F_03	0.285	0.425
UKP_E2C_A2F_03	0.250	0.337	UKP_E2J_A2F_01	0.332	0.426	UKP_E2K_A2F_04	0.260	0.404
HITS_E2C_A2F_03	0.245	0.319	UKP_E2J_A2F_02	0.329	0.425	DUIIS_A2F_E2K_4Rec	0.258	0.379
HITS_E2C_A2F_02	0.241	0.315	UKP_E2J_A2F_03	0.297	0.408	DUIIS_A2F_E2K_4Pre	0.252	0.357
KMI_SVM_ESA	0.231	0.344	UKP_E2J_A2F_05	0.265	0.377	UKP_E2K_A2F_05	0.244	0.430
HITS_E2C_A2F_01	0.229	0.296	UKP_E2J_A2F_04	0.265	0.387	HITS_E2K_A2F_03	0.235	0.341
IASL_E2C_01	0.205	0.308	QUT_PNM_JA	0.122	0.187	HITS_E2K_A2F_02	0.234	0.342
QUT_LinkProb_ZH	0.202	0.309	kslab_nut_A2F_E2J_01	0.028	0.075	KSLP_A2F_E2K_01	0.233	0.341
IASL_E2C_02	0.200	0.312	TRANSLITERATION_JA	0.026	0.051	HITS_E2K_A2F_01	0.214	0.307
IASL_E2C_03	0.194	0.301				KSLP_E2K_05	0.184	0.264
UKP_E2C_A2F_05	0.166	0.319				KSLP_E2K_04	0.177	0.252
UKP_E2C_A2F_04	0.145	0.257				KSLP_E2K_02	0.170	0.245
KMI_SVM_TERMDB	0.133	0.192				KSLP_E2K_03	0.169	0.244
QUT_LinkProbZh2_ZH	0.132	0.267				QUT_PNM_KO	0.114	0.182
QUT_LinkProbZh_ZH	0.123	0.274				IISR_ko_title_to_key_name	0.094	0.129
WUST_A2F_E2C_03	0.082	0.124				IISR_kttkn_to_ztttzt	0.092	0.130
QUT_PNM_ZH	0.074	0.123				IISR_sttent_to_ztttzt_to_kttkn	0.092	0.131
WUST_A2F_E2C_01	0.070	0.102				IISR_sttent_to_kttkn	0.090	0.131
WUST_A2F_E2C_02	0.065	0.103				IISR_singular_term_to_en_title	0.050	0.067
KMI_ESA_SVM_ESA discovery	0.054	0.132						
QUT_LinkProbIR_ZH	0.049	0.134						
WUST_A2F_E2C_04	0.038	0.056						
ISTIC_A2F_E2C_01	0.029	0.100						

Table 11. F2F evaluation with manual assessment results: *Precision-at-N* (English-2-Chinese)

Run-ID	P5	P10	P20	P30	P50	P250
QUT_LinkProb_ZH	0.808	0.652	0.570	0.511	0.446	0.286
HITS_E2C_A2F_03	0.752	0.736	0.770	0.753	0.704	0.287
HITS_E2C_A2F_02	0.752	0.772	0.748	0.735	0.701	0.288
HITS_E2C_A2F_01	0.752	0.776	0.764	0.739	0.702	0.268
KMI_SVM_TERMDB	0.752	0.692	0.636	0.613	0.561	0.178
WUST_A2F_E2C_03	0.744	0.684	0.576	0.528	0.440	0.120
WUST_A2F_E2C_01	0.744	0.692	0.572	0.516	0.407	0.096
UKP_E2C_A2F_03	0.736	0.712	0.742	0.715	0.687	0.322
KMI_SVM_ESA	0.728	0.720	0.678	0.668	0.615	0.306
KMI_SVM_ESA_TERMDB	0.720	0.728	0.684	0.648	0.604	0.358
WUST_A2F_E2C_02	0.712	0.668	0.552	0.476	0.381	0.098
UKP_E2C_A2F_01	0.696	0.696	0.728	0.719	0.682	0.409
UKP_E2C_A2F_02	0.688	0.684	0.716	0.712	0.678	0.417
IASL_E2C_01	0.680	0.660	0.610	0.613	0.588	0.297
IASL_E2C_02	0.680	0.640	0.596	0.583	0.566	0.302
QUT_PNM_ZH	0.664	0.608	0.514	0.479	0.420	0.119
IASL_E2C_03	0.648	0.604	0.574	0.588	0.578	0.287
UKP_E2C_A2F_04	0.568	0.536	0.536	0.521	0.519	0.244
QUT_LinkProbZh2_ZH	0.528	0.476	0.416	0.396	0.367	0.248
WUST_A2F_E2C_04	0.520	0.512	0.428	0.364	0.270	0.057
KMI_ESA_SVM_ESAdiscovery	0.464	0.388	0.348	0.321	0.283	0.119
QUT_LinkProbZh_ZH	0.424	0.348	0.346	0.327	0.316	0.252
UKP_E2C_A2F_05	0.408	0.468	0.480	0.499	0.482	0.305
ISTIC_A2F_E2C_01	0.256	0.228	0.238	0.217	0.202	0.091
QUT_LinkProbIR_ZH	0.248	0.232	0.198	0.184	0.172	0.123

Table 12. F2F evaluation with manual assessment results: *Precision-at-N* (English-2-Japanese)

Run-ID	P5	P10	P20	P30	P50	P250
HITS_E2J_A2F_03	0.656	0.652	0.626	0.552	0.472	0.144
HITS_E2J_A2F_01	0.624	0.640	0.580	0.543	0.468	0.142
HITS_E2J_A2F_02	0.608	0.644	0.582	0.536	0.466	0.144
UKP_E2J_A2F_03	0.544	0.564	0.536	0.473	0.378	0.107
UKP_E2J_A2F_04	0.496	0.508	0.498	0.440	0.358	0.103
QUT_PNM_JA	0.496	0.420	0.312	0.245	0.186	0.049
UKP_E2J_A2F_01	0.480	0.564	0.540	0.492	0.391	0.129
UKP_E2J_A2F_02	0.456	0.552	0.524	0.477	0.386	0.131
UKP_E2J_A2F_05	0.264	0.388	0.398	0.400	0.345	0.130
kslab_nut_A2F_E2J_01	0.160	0.144	0.094	0.073	0.054	0.019
TRANSLITERATION_JA	0.072	0.060	0.052	0.053	0.050	0.040

Table 13. F2F evaluation with manual assessment results: *Precision-at-N* (English-2-Korean)

Run-ID	P5	P10	P20	P30	P50	P250
QUT_PNM_KO	0.720	0.672	0.612	0.537	0.463	0.124
IISR_singular_term_to_en_title	0.720	0.604	0.416	0.373	0.232	0.046
IISR_ko_title_to_key_name	0.712	0.692	0.606	0.532	0.402	0.086
HITS_E2K_A2F_03	0.696	0.688	0.680	0.671	0.643	0.239
KSLP_E2K_02	0.680	0.664	0.638	0.621	0.575	0.169
HITS_E2K_A2F_02	0.672	0.672	0.682	0.668	0.635	0.239
KSLP_E2K_03	0.672	0.656	0.640	0.620	0.578	0.168
HITS_E2K_A2F_01	0.656	0.684	0.688	0.671	0.625	0.216
IISR_kttkn_to_ztttzt	0.656	0.672	0.618	0.533	0.405	0.088
IISR_sttent_to_ztttzt_to_kttkn	0.648	0.660	0.610	0.527	0.401	0.088
IISR_sttent_to_kttkn	0.648	0.620	0.606	0.529	0.406	0.088
KSLP_E2K_05	0.640	0.656	0.634	0.631	0.581	0.182
KSLP_E2K_04	0.640	0.656	0.638	0.625	0.577	0.175
DUIIS_A2F_E2K_4Rec	0.632	0.692	0.700	0.687	0.658	0.263
DUIIS_A2F_E2K_4Pre	0.632	0.692	0.716	0.705	0.679	0.247
UKP_E2K_A2F_03	0.600	0.636	0.656	0.675	0.662	0.304
UKP_E2K_A2F_01	0.552	0.584	0.636	0.663	0.655	0.402
KSLP_A2F_E2K_01	0.552	0.576	0.566	0.539	0.483	0.286
UKP_E2K_A2F_02	0.544	0.568	0.632	0.656	0.656	0.413
UKP_E2K_A2F_04	0.528	0.600	0.632	0.651	0.643	0.285
UKP_E2K_A2F_05	0.376	0.392	0.434	0.477	0.501	0.327

- *English-2-Korean*

MAP: UKP, DUIIS, HITS

R-Prec: UKP, DUIIS, HITS

Precision-at-5: QUT, IISR, HITS

In file-to-file evaluation using *qrel* from manual assessment rather than Wikipedia ground-truth, team HITS remains the top performer in English-2-Japanese task; runs from team UKP outperform others and make team UKP the top performer in English-2-Chinese and English-2-Korean tasks. Performance of runs from teams KMI, QUT, and DUIIS shows their methods can also contribute good links.

6.2.2 Anchor-to-File Evaluation

Scores computed with different metrics in anchor-to-file level could more accurately reflect the actual performance of different CLLD systems, as relevancy of both anchors and their associated links is considered.

Table 14 shows the scores of *MAP* and *R-Prec* of all runs in three subtasks. The table is sorted on MAP score. Table 15, 16, 17 show the evaluation results using *Precision-at-N* metric for English-2-Chinese, English-2-Japanese, and English-2-Korean runs respectively.

From these tables, it can be seen that the rankings of runs are different with those in file-to-file evaluation as shown in previous section. It could suggest that some identified anchors in the high ranking runs of F2F evaluation may be considered inappropriate in manual assessment, even though the links associated with them are irrelevant.

The top three teams of each measure are listed below:

- *English-2-Chinese*

MAP: UKP, QUT, HITS

R-Prec: UKP, QUT, KMI

Precision-at-5: KMI, QUT, UKP

- *English-2-Japanse*

MAP: HITS, UKP, QUT

R-Prec: HITS, UKP, QUT

Precision-at-5: HITS, UKP, QUT

- *English-2-Korean*

MAP: UKP, HITS, KSLP

R-Prec: UKP, HITS, KSLP

Precision-at-5: HITS, KSLP, UKP

In English-2-Chinese and English-2-Korean tasks, team UKP is the top performer; in English-2-Japanese task HITS is the top performer. Teams QUT, KMI and KSLP are the middle performers.

Table 14. A2F evaluation with manual assessment results: MAP, R-PREC

English-2-Chinese			English-2-Japanese			English-2-Korean		
Run-ID	**MAP**	**r-prec**	**Run-ID**	**MAP**	**r-prec**	**Run-ID**	MAP	r-prec
UKP_E2C_A2F_02	0.157	0.171	HITS_E2J_A2F_02	0.425	0.059	UKP_E2K_A2F_02	0.232	0.207
UKP_E2C_A2F_01	0.157	0.171	HITS_E2J_A2F_03	0.419	0.062	UKP_E2K_A2F_01	0.226	0.204
UKP_E2C_A2F_03	0.116	0.121	HITS_E2J_A2F_01	0.418	0.060	UKP_E2K_A2F_05	0.150	0.149
QUT_LinkProb_ZH	0.115	0.133	UKP_E2J_A2F_05	0.333	0.040	UKP_E2K_A2F_03	0.145	0.145
HITS_E2C_A2F_03	0.102	0.105	UKP_E2J_A2F_01	0.330	0.047	UKP_E2K_A2F_04	0.127	0.136
HITS_E2C_A2F_02	0.102	0.105	UKP_E2J_A2F_02	0.326	0.046	HITS_E2K_A2F_03	0.124	0.117
KMI_SVM_ESA_TERMDB	0.097	0.114	UKP_E2J_A2F_03	0.276	0.042	HITS_E2K_A2F_02	0.122	0.117
HITS_E2C_A2F_01	0.096	0.098	UKP_E2J_A2F_04	0.254	0.040	HITS_E2K_A2F_01	0.113	0.105
UKP_E2C_A2F_05	0.094	0.122	QUT_PNM_JA	0.087	0.016	KSLP_E2K_05	0.081	0.086
QUT_LinkProbZh_ZH	0.094	0.119	TRANSLITERATION_JA	0.000	0.000	KSLP_A2F_E2K_01	0.079	0.097
QUT_LinkProbZh2_ZH	0.090	0.117				KSLP_E2K_04	0.076	0.080
KMI_SVM_ESA	0.080	0.092				KSLP_E2K_03	0.073	0.076
UKP_E2C_A2F_04	0.073	0.095				KSLP_E2K_02	0.072	0.075
KMI_SVM_TERMDB	0.070	0.075				DUIIS_A2F_E2K_4Rec	0.043	0.036
IASL_E2C_01	0.037	0.036				QUT_PNM_KO	0.043	0.043
IASL_E2C_02	0.034	0.037				DUIIS_A2F_E2K_4Pre	0.041	0.034
IASL_E2C_03	0.033	0.036				IISR_sttent_to_ztttzt_to_kttkn	0.030	0.029
QUT_PNM_ZH	0.030	0.033				IISR_kttkn_to_ztttzt	0.029	0.029
KMI_ESA_SVM_ESA discovery	0.014	0.035				IISR_sttent_to_kttkn	0.029	0.029
WUST_A2F_E2C_03	0.012	0.010				IISR_ko_title_to_key_name	0.029	0.029
WUST_A2F_E2C_02	0.009	0.008				IISR_singular_term_to_en_title	0.014	0.013
WUST_A2F_E2C_01	0.009	0.008						
QUT_LinkProbIR_ZH	0.008	0.026						
WUST_A2F_E2C_04	0.008	0.006						
ISTIC_A2F_E2C_01	0.000	0.000						

Table 15. A2F evaluation with manual assessment results: *Precision-at-N* (English-2-Chinese)

Run-ID	P5	P10	P20	P30	P50	P250
KMI_SVM_TERMDB	0.376	0.368	0.324	0.316	0.297	0.096
KMI_SVM_ESA_TERMDB	0.368	0.368	0.330	0.303	0.269	0.142
KMI_SVM_ESA	0.360	0.364	0.330	0.299	0.260	0.113
QUT_LinkProb_ZH	0.336	0.308	0.294	0.288	0.277	0.172
QUT_LinkProbZh_ZH	0.320	0.244	0.260	0.273	0.269	0.158
QUT_LinkProbZh2_ZH	0.312	0.312	0.304	0.299	0.271	0.155
UKP_E2C_A2F_03	0.240	0.264	0.308	0.303	0.287	0.154
HITS_E2C_A2F_03	0.240	0.236	0.324	0.319	0.294	0.131
UKP_E2C_A2F_02	0.216	0.256	0.298	0.299	0.289	0.221
UKP_E2C_A2F_01	0.216	0.256	0.298	0.299	0.286	0.221
QUT_PNM_ZH	0.208	0.204	0.214	0.220	0.187	0.045
HITS_E2C_A2F_01	0.176	0.252	0.282	0.281	0.290	0.121
UKP_E2C_A2F_04	0.176	0.200	0.234	0.228	0.221	0.121
HITS_E2C_A2F_02	0.160	0.248	0.282	0.281	0.281	0.131
UKP_E2C_A2F_05	0.136	0.144	0.172	0.197	0.185	0.161
QUT_LinkProbIR_ZH	0.104	0.104	0.072	0.073	0.070	0.033
IASL_E2C_01	0.096	0.112	0.090	0.104	0.094	0.044
IASL_E2C_02	0.096	0.096	0.094	0.099	0.094	0.044
IASL_E2C_03	0.088	0.068	0.080	0.083	0.074	0.044
KMI_ESA_SVM_ESAdiscovery	0.088	0.108	0.110	0.108	0.090	0.045
WUST_A2F_E2C_03	0.056	0.068	0.058	0.061	0.055	0.014
WUST_A2F_E2C_02	0.056	0.068	0.056	0.059	0.047	0.011
WUST_A2F_E2C_01	0.056	0.068	0.056	0.059	0.047	0.011
WUST_A2F_E2C_04	0.056	0.076	0.058	0.048	0.038	0.008
ISTIC_A2F_E2C_01	0.000	0.000	0.000	0.000	0.000	0.000

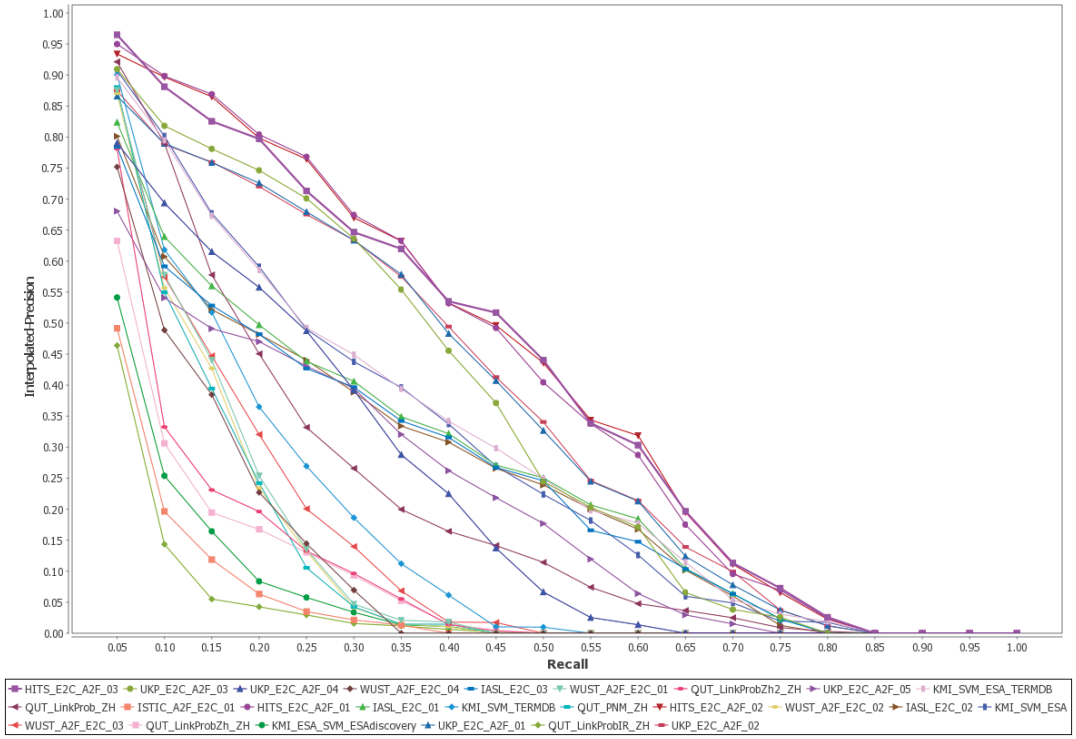
Table 16. A2F evaluation with manual assessment results: *Precision-at-N* (English-2-Japanese)

Run-ID	P5	P10	P20	P30	P50	P250
HITS_E2J_A2F_03	0.344	0.360	0.350	0.316	0.266	0.088
HITS_E2J_A2F_01	0.288	0.296	0.284	0.276	0.260	0.087
HITS_E2J_A2F_02	0.256	0.296	0.292	0.277	0.260	0.088
UKP_E2J_A2F_03	0.200	0.252	0.266	0.244	0.194	0.060
UKP_E2J_A2F_02	0.192	0.248	0.266	0.248	0.201	0.077
UKP_E2J_A2F_01	0.184	0.244	0.266	0.251	0.203	0.077
UKP_E2J_A2F_04	0.168	0.220	0.242	0.225	0.182	0.059
QUT_PNM_JA	0.128	0.124	0.108	0.096	0.077	0.020
UKP_E2J_A2F_05	0.104	0.164	0.168	0.188	0.171	0.082
TRANSLITERATION_JA	0.000	0.000	0.000	0.000	0.000	0.000
HITS_E2J_A2F_03	0.344	0.360	0.350	0.316	0.266	0.088

Table 17. A2F evaluation with manual assessment results: *Precision-at-N* (English-2-Korean)

Run-ID	P5	P10	P20	P30	P50	P250
HITS_E2K_A2F_03	0.368	0.364	0.340	0.332	0.320	0.129
HITS_E2K_A2F_02	0.312	0.268	0.328	0.333	0.320	0.129
HITS_E2K_A2F_01	0.272	0.260	0.332	0.332	0.318	0.116
KSLP_E2K_03	0.240	0.264	0.276	0.265	0.254	0.084
KSLP_E2K_02	0.232	0.252	0.270	0.263	0.250	0.083
UKP_E2K_A2F_03	0.224	0.276	0.316	0.331	0.327	0.160
KSLP_E2K_05	0.200	0.256	0.264	0.268	0.254	0.095
KSLP_E2K_04	0.200	0.264	0.272	0.265	0.248	0.088
IISR_sttent_to_kttkn	0.200	0.168	0.172	0.161	0.134	0.032
IISR_singular_term_to_en_title	0.200	0.192	0.146	0.116	0.071	0.014
UKP_E2K_A2F_02	0.192	0.248	0.324	0.323	0.326	0.252
UKP_E2K_A2F_01	0.192	0.252	0.316	0.316	0.321	0.248
KSLP_A2F_E2K_01	0.184	0.196	0.210	0.219	0.231	0.108
IISR_ko_title_to_key_name	0.184	0.204	0.176	0.163	0.135	0.031
IISR_sttent_to_ztttzt_to_kttkn	0.168	0.188	0.174	0.157	0.136	0.032
UKP_E2K_A2F_04	0.160	0.228	0.290	0.307	0.309	0.150
IISR_kttkn_to_ztttzt	0.152	0.192	0.178	0.161	0.133	0.032
QUT_PNM_KO	0.136	0.200	0.220	0.217	0.193	0.047
UKP_E2K_A2F_05	0.120	0.124	0.140	0.184	0.198	0.189
DUIIS_A2F_E2K_4Rec	0.064	0.076	0.090	0.116	0.115	0.036
DUIIS_A2F_E2K_4Pre	0.056	0.076	0.090	0.117	0.118	0.034

a.



b.

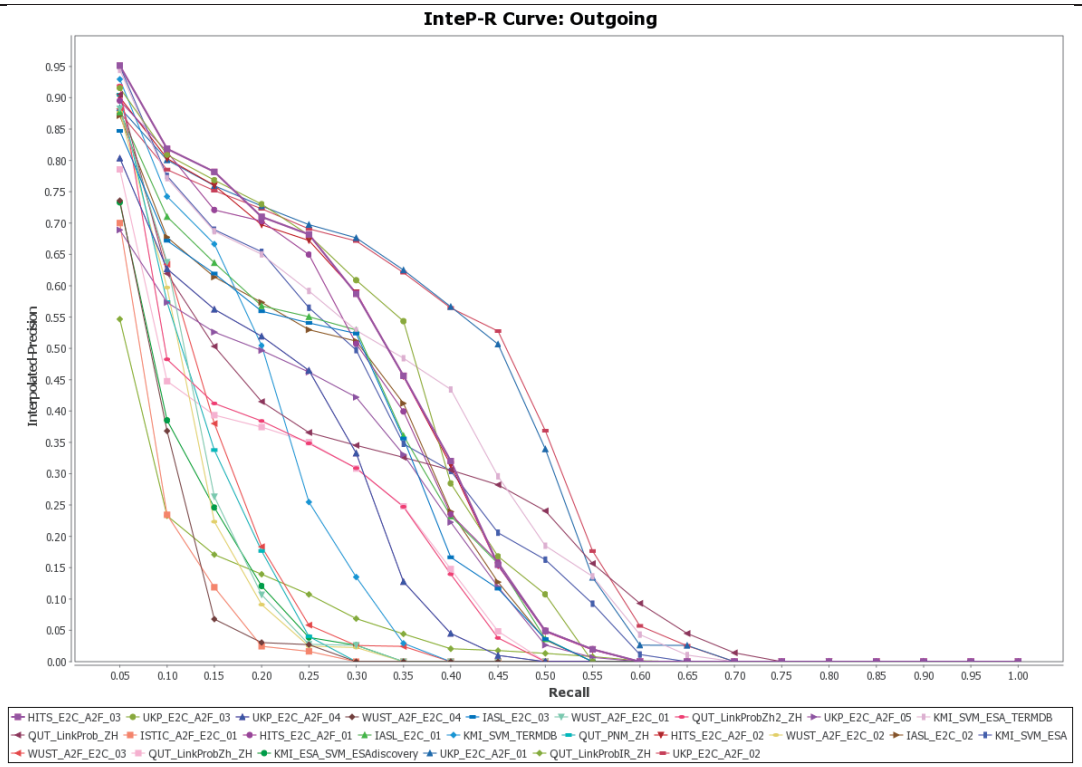
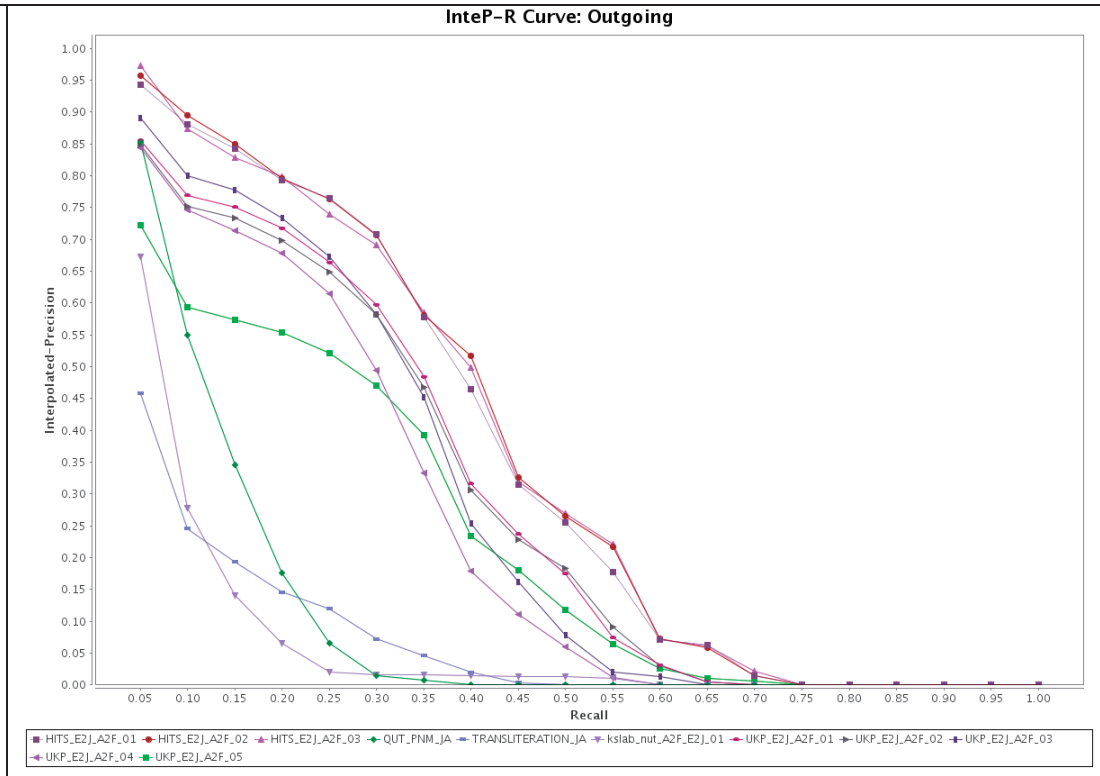


Figure 7. Interpolated Precision-Recall of English-2-Chinese links (plot *a* is P-R curves of runs in f2f evaluation with Wikipedia ground-truth; plot *b* is P-R curves of runs in a2f evaluation with manual assessment results)

a.



b.

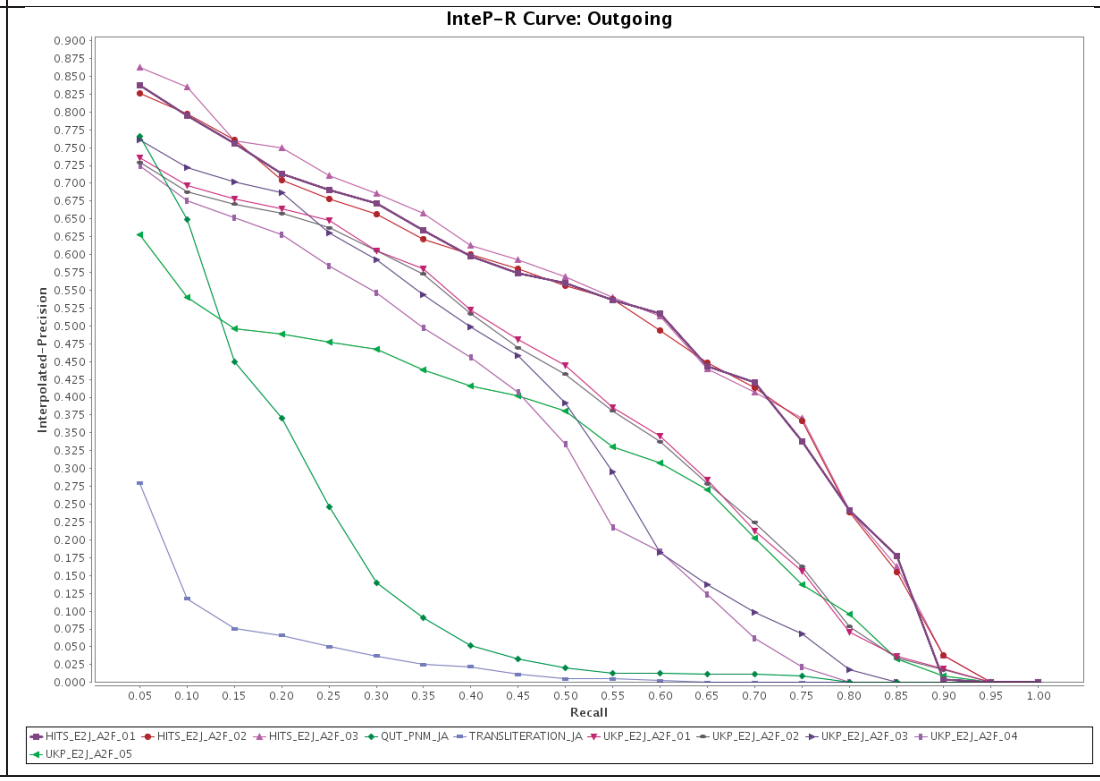
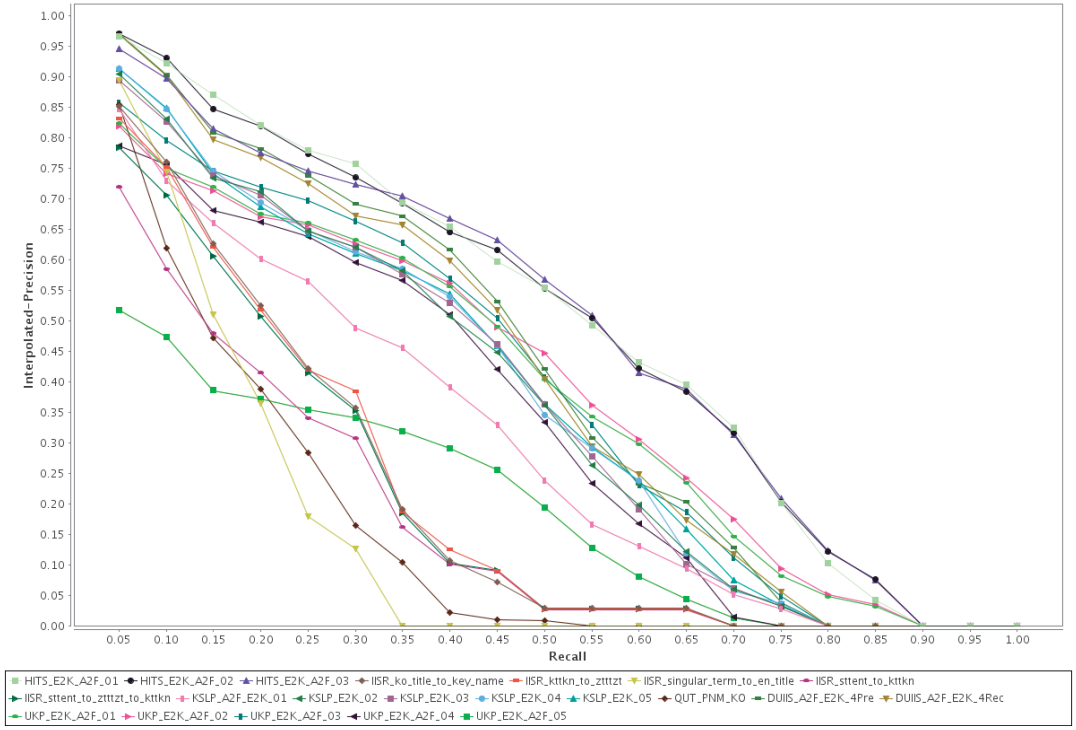


Figure 8. Interpolated Precision-Recall of English-2-Japanese links (plot *a* is P-R curves of runs in f2f evaluation with Wikipedia ground-truth; plot *b* is P-R curves of runs in a2f evaluation with manual assessment results)

a.



b.

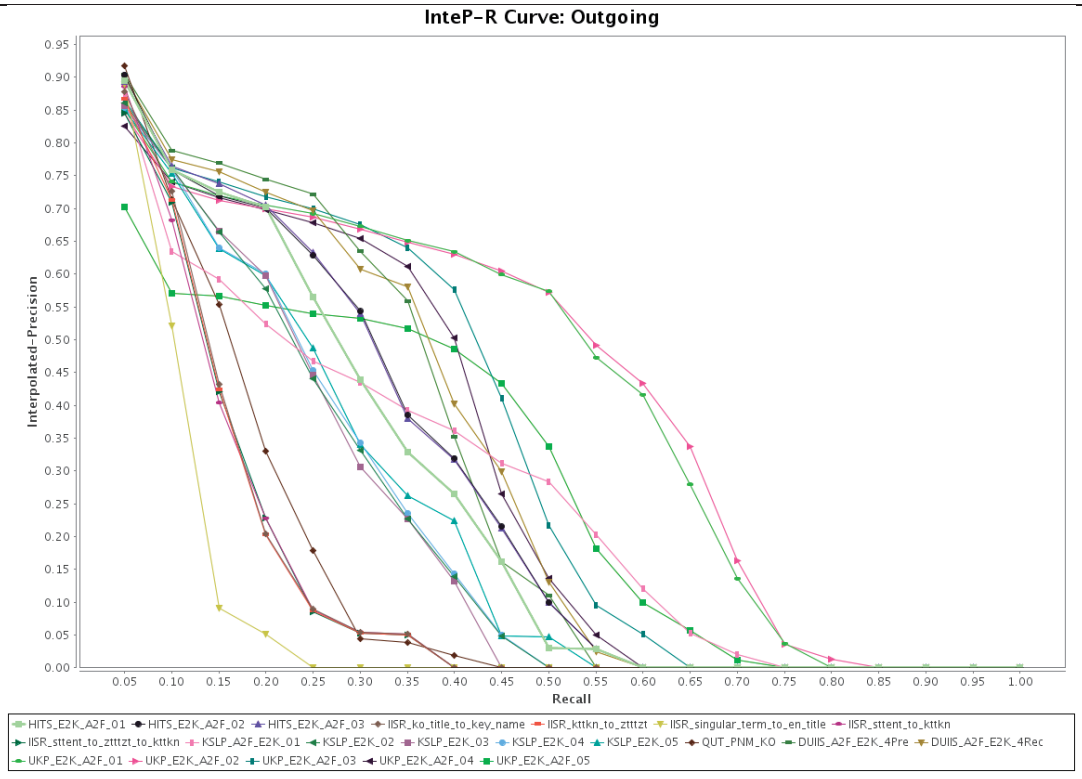


Figure 9. Interpolated Precision-Recall of English-2-Korean links (plot *a* is P-R curves of runs in f2f evaluation with Wikipedia ground-truth; plot *b* is P-R curves of runs in a2f evaluation with manual assessment results)

6.3 Unique Relevant Links

Submitted runs have different rankings in different evaluations with different metrics. To have more straightforward comparisons of performance of all runs, precision-recall curves of runs of both file-to-file and anchor-to-file evaluations are given in figure 7, 8, 9 for English-2-Chinese, English-2-Japanese, and English-2-Korean tasks respectively. It can be easily seen that teams HITS and UKP are the top performers of three language subtasks in both F2F and A2F evaluations.

The scores of different evaluation metrics and precision-recall curves of submitted runs do reflect the performance difference of various CLLD systems in one aspect. Inspired by the NTCIR-8 ACLIA IR4QA task[8], it will be also interesting to see the number of unique relevant links contributed by the runs or teams. The number of unique relevant links that a CLLD system can contribute indicates a scale of how broad a system can expose the knowledge in Wikipedia to users. The larger unique relevant links can be found, the more chances it can have to increase the knowledge exposure in Wikipedia.

The link numbers in *qrels* of Wikipedia ground-truth and manual assessment results for the 25 test topics are given in table 18. From the table, it can be seen that there are much more discovered relevant links in *qrel* of E2C and E2K manual assessment results than that in Wikipedia ground-truth; the number of relevant links in E2J manual assessment results is reasonably less than that in the Wikipedia ground-truth due to a limited number of submissions for English-to-Japanese task.

Table 18. Statistics of *qrels* of Wikipedia ground-truth and manual assessment results

Task	# of Links	Overlapping
E2C ground-truth	2116	1134
E2C manual assessment	4309	
E2J ground-truth	2939	781
E2J manual assessment	1118	
E2K ground-truth	1681	821
E2K manual assessment	2786	

Table 19, 20 and 21 are the statistics of unique relevant links found by each team with the Wikipedia ground-truth in English-2-Chinese, English-2-Japanese, and English-2-Korean tasks respectively. Table 22, 23 and 24 list the number of unique relevant links found by each team with manual assessment results in the three different language tasks.

From the above unique link number tables, it can be seen that most runs are able to contribute unique relevant links, but the number of unique relevant links discovered by each run varies.

The teams with the highest number of unique relevant links found with the Wikipedia ground-truth are:

- UKP for E2C
- QUT for E2J
- HITS for E2K

The teams with the highest number of unique relevant links found with the manual assessment results are:

- QUT for E2C
- HITS for E2J
- KSLP for E2K

Of particular note is that team QUT found 1103 unique relevant links that are more than seven times of links (152) found by team KMI which is ranked second with manual assessment results in English-2-Chinese task.

Table 19. Unique relevant English-2-Chinese links with Wikipedia ground-truth

Team	Run-ID	#	total
UKP	UKP_E2C_A2F_05	91	97
	UKP_E2C_A2F_04	69	
	UKP_E2C_A2F_03	6	
	UKP_E2C_A2F_02	5	
	UKP_E2C_A2F_01	4	
QUT	QUT_LinkProb_ZH	48	95
	QUT_LinkProbIR_ZH	35	
	QUT_LinkProbZh2_ZH	32	
	QUT_LinkProbZh_ZH	29	
	QUT_PNM_ZH	2	
KMI	KMI_SVM_ESA_TERMDB	16	27
	KMI_SVM_ESA	16	
	KMI_ESA_SVM_ESAdiscovery	15	
	KMI_SVM_TERMDB	5	
HITS	HITS_E2C_A2F_03	14	14
	HITS_E2C_A2F_02	14	
	HITS_E2C_A2F_01	11	
NTHUISA	IASL_E2C_01	8	8
	IASL_E2C_02	7	
	IASL_E2C_03	7	
WUST	WUST_A2F_E2C_02	3	4
	WUST_A2F_E2C_04	1	
	WUST_A2F_E2C_03	1	

Table 20. Unique relevant English-2-Japanese links with Wikipedia ground-truth

Team	Run-ID	#	total
QUT	TRANSLITERATION_JA	167	172
	QUT_PNM_JA	9	
kslab_nut	kslab_nut_A2F_E2J_01	159	159
HITS	HITS_E2J_A2F_02	140	141
	HITS_E2J_A2F_03	139	
	HITS_E2J_A2F_01	130	
UKP	UKP_E2J_A2F_05	106	131
	UKP_E2J_A2F_02	101	
	UKP_E2J_A2F_01	94	
	UKP_E2J_A2F_04	63	
	UKP_E2J_A2F_03	61	

Table 21. Unique relevant English-2-Korean links with Wikipedia ground-truth

Team	Run-ID	#	total
HITS	HITS_E2K_A2F_03	62	62
	HITS_E2K_A2F_02	62	
	HITS_E2K_A2F_01	56	
KSLP	KSLP_A2F_E2K_01	25	31
	KSLP_E2K_05	7	
	KSLP_E2K_04	7	
	KSLP_E2K_03	6	
	KSLP_E2K_02	6	
UKP	UKP_E2K_A2F_05	18	24
	UKP_E2K_A2F_02	9	
	UKP_E2K_A2F_01	9	
	UKP_E2K_A2F_04	4	
	UKP_E2K_A2F_03	1	
DUIIS	DUIIS_A2F_E2K_4Rec	6	6
	DUIIS_A2F_E2K_4Pre	5	
IISR	IISR_kttkn_to_ztttzt	5	5
	IISR_ko_title_to_key_name	3	
	IISR_sttent_to_kttkn	3	
	IISR_sttent_to_ztttzt_to_kttkn	3	
QUT	QUT_PNM_KO	1	1

Table 22. Unique relevant English-2-Chinese links with manual assessment results

Team	Run-ID	#	total
QUT	QUT_LinkProbIR_ZH	562	1103
	QUT_LinkProb_ZH	526	
	QUT_LinkProbZh_ZH	261	
	QUT_LinkProbZh2_ZH	256	
	QUT_PNM_ZH	48	
KMI	KMI_ESA_SVM_ESAdiscovery	96	152
	KMI_SVM_ESA_TERMDB	65	
	KMI_SVM_ESA	59	
	KMI_SVM_TERMDB	15	
UKP	UKP_E2C_A2F_05	48	88
	UKP_E2C_A2F_02	47	
	UKP_E2C_A2F_01	43	
	UKP_E2C_A2F_04	20	
	UKP_E2C_A2F_03	19	
HITS	HITS_E2C_A2F_03	40	43
	HITS_E2C_A2F_02	40	
	HITS_E2C_A2F_01	38	
NTHUISA	IASL_E2C_01	10	10
	IASL_E2C_02	10	
	IASL_E2C_03	10	
WUST	WUST_A2F_E2C_02	1	1
	WUST_A2F_E2C_01	1	
	WUST_A2F_E2C_04	1	
	WUST_A2F_E2C_03	1	

Table 23. Unique relevant English-2-Japanese links with manual assessment results

Team	Run-ID	#	total
HITS	HITS_E2J_A2F_02	136	138
	HITS_E2J_A2F_03	134	
	HITS_E2J_A2F_01	130	
UKP	UKP_E2J_A2F_05	109	122
	UKP_E2J_A2F_02	88	
	UKP_E2J_A2F_01	85	
	UKP_E2J_A2F_04	55	
QUT	QUT_PNM_JA	6	6
	TRANSLITERATION_JA	1	

Table 24. Unique relevant English-2-Korean links with manual assessment results

Team	Run-ID	#	total
KSLP	KSLP_A2F_E2K_01	296	362
	KSLP_E2K_05	72	
	KSLP_E2K_04	65	
	KSLP_E2K_03	63	
	KSLP_E2K_02	62	
UKP	UKP_E2K_A2F_05	211	305
	UKP_E2K_A2F_02	156	
	UKP_E2K_A2F_01	148	
	UKP_E2K_A2F_04	60	
	UKP_E2K_A2F_03	40	
HITS	HITS_E2K_A2F_03	105	105
	HITS_E2K_A2F_02	105	
	HITS_E2K_A2F_01	91	
QUT	QUT	37	37
	QUT_PNM_KO	37	
IISR	IISR_ko_title_to_key_name	5	5
	IISR_kttkn_to_ztttzt	5	
	IISR_sttent_to_kttkn	5	
	IISR_sttent_to_ztttzt_to_kttkn	5	
DUIIS	DUIIS_A2F_E2K_4Rec	3	3
	DUIIS_A2F_E2K_4Pre	1	

7. CONCLUSION AND FUTURE WORK

In the environment of globalization nowadays, it is inevitable for knowledge management and discovery to go multilingual, cross-lingual. Link targets in the knowledge base such as Wikipedia should not be restricted to the language of the anchors. Several such links already exist in Wikipedia, but we believe they are underutilised.

In this paper, we provide details of the cross-lingual link discovery task definition, the run submission specification, the assessment and evaluation framework, the evaluation metrics and the evaluation results of the submitted runs. This year's task focuses on cross-lingual linking between English source documents and Chinese, Korean, and Japanese target documents.

The evaluations in both file-to-file and anchor-to-file levels show promising results of participating teams even this is the first year of this task. Particularly, team HITS and UKP achieved very high scores of different evaluation measures (*MAP, R-Prec, Precision-at-N*) in three language subtasks.

With the standard test collections and evaluation data set developed for this task, systems or applications for realising CLLD can be built or further refined to provide a better automated cross-lingual linking for knowledge management and sharing. With the possible future deployment of techniques and approaches used by the participants, we hope user experience in viewing or editing document in different languages could be enhanced; language is then no longer a barrier for knowledge discovery.

Crosslink tasks that link CJK documents to English ones are planned for evaluation round of next year at NTCIR-10.

Table 25. CLLD system descriptions of submitted runs in English-to-Chinese subtask

HITS_E2C_A2F_01	The results are produced using the following approach: -Preprocessing: Parsing of the topic xml files, tokenization (opennlp tools), tagging (tree-tagger)-Term recognition: -Disambiguation relative to enhanced English Wikipedia version: Global graph-based approach using a combined score of different relatedness measures as edge weights and local features as vertex weight. To select concepts a weighted clique algorithm is used.-Mapping of ids (from disambiguation) to the respective language (i.e. zh) using crosslanguage links and other mapping techniques
HITS_E2C_A2F_02	The results are produced using the following approach: -Preprocessing: Parsing of the topic xml files, tokenization (opennlp tools), tagging (tree-tagger)-Term recognition: -Disambiguation relative to enhanced English Wikipedia version: Global graph-based approach using a combined score of different relatedness measures as edge weights and local features as vertex weight. To select concepts a weighted clique algorithm is used.-Mapping of ids (from disambiguation) to the respective language (i.e. zh) using crosslanguage links and other mapping techniques
HITS_E2C_A2F_03	The results are produced using the following approach: -Preprocessing: Parsing of the topic xml files, tokenization (opennlp tools), tagging (tree-tagger)-Term recognition: -Disambiguation relative to enhanced English Wikipedia version: Global graph-based approach using a combined score of different relatedness measures as edge weights and local features as vertex weight. To select concepts a weighted clique algorithm is used.-Mapping of ids (from disambiguation) to the respective language (i.e. zh) using crosslanguage links and other mapping techniques
KMI_SVM_ESA_TERMDB	SVM+ESA+TERMDB
KMI_SVM_ESA	SVM+ESA
KMI_SVM_TERMDB	SVM+ESA+TERMDB
KMI_ESA_SVM_ESAdiscovery	ESA+SVM+ESAdiscovery
QUT_LinkProbIR_ZH	Use the anchors recommended by link probability, and retrieve relevant links using a search engine with anchors as query terms
QUT_LinkProbZh2_ZH	Same as QUT_LinkProbZh_ZH , except for that anchors are sorted based on Chinese link probability table.
QUT_LinkProbZh_ZH	Use two set of link probability tables (one Chinese; one English mining from English Wikipedia corpus from INEX), and tables are connected by translation. Anchors are sorted based on English link probability table.
QUT_LinkProb_ZH	Use link probability for anchor sorting and link recommendation
IASL_E2C_01	Find Missing Link By Similar Pages
IASL_E2C_02	Find Missing Link By Similar Pages 2
IASL_E2C_03	Missing Link By Google Translated and Relevant Ranking
QUT_PNM_ZH	This is a run using the PNM algorithm, and the cross-lingual title-to-target table is generated from the NTCIR 9- Crosslink: Chinese Wikipedia Corpus
ISTIC_A2F_E2C_01	We use the multi-filtering method to solve the CLLD problems. First, the punctuation and the common stopwords could be the separating characters to separate the long paragraph into short text segments. Then, the word frequency statistics could be done to get the low frequency words (some threshold should be set here). These selected low frequency words would be the new separating characters to separate the short text segments above, and we could get some new words and multi-word phrases. In the third stage, according to the words which are above the thresholds, computing the cooccurrence times among these words in each sentence of the documents and acquiring the new multi-words by the sequence of words above in sentence. In the fourth stage, pos tagging would be done first. Then, some kinds of words (such as adjective, adverb and verb, et.al) and some POS collocation modes (such as noun+verb, adj.+adj., et.al) would be used to filter the words and phrased getting from the steps above. In the end, some weighting rules would be used to select the final words and phrases for translation. The translation tool we used is Google translator and the searching platform is established by lucene software package.
UKP_E2C_A2F_01	ngram.anchorprobability.cascade_3.titlematch_1_1
UKP_E2C_A2F_02	ngram.anchorprobability.cascade_3.titlematch_3_3
UKP_E2C_A2F_03	ngram.anchorprobability.interlingualindex.titlematch_1_1
UKP_E2C_A2F_04	ngram.anchorprobability.interlingualindex.cascade_3_1
UKP_E2C_A2F_05	ngram.anchorprobability.cascade_3.incominglinksearch_1_1

WUST_A2F_E2C_01	Using online data collecting and processing algorithm
WUST_A2F_E2C_02	Using online data collecting and processing algorithm
WUST_A2F_E2C_03	Using online data collecting and processing algorithm
WUST_A2F_E2C_04	Using online data collecting and processing algorithm

Table 26. CLLD system descriptions of submitted runs in English-to-Japanese subtask

HITS_E2J_A2F_01	The results are produced using the following approach: -Preprocessing: Parsing of the topic xml files, tokenization (opennlp tools), tagging (tree-tagger)-Term recognition: -Disambiguation relative to enhanced English Wikipedia version: Global graph-based approach using a combined score of different relatedness measures as edge weights and local features as vertex weight. To select concepts a weighted clique algorithm is used.-Mapping of ids (from disambiguation) to the respective language (i.e. ja) using crosslanguage links and other mapping techniques
HITS_E2J_A2F_02	The results are produced using the following approach: -Preprocessing: Parsing of the topic xml files, tokenization (opennlp tools), tagging (tree-tagger)-Term recognition: -Disambiguation relative to enhanced English Wikipedia version: Global graph-based approach using a combined score of different relatedness measures as edge weights and local features as vertex weight. To select concepts a weighted clique algorithm is used.-Mapping of ids (from disambiguation) to the respective language (i.e. ja) using crosslanguage links and other mapping techniques
HITS_E2J_A2F_03	The results are produced using the following approach: -Preprocessing: Parsing of the topic xml files, tokenization (opennlp tools), tagging (tree-tagger)-Term recognition: -Disambiguation relative to enhanced English Wikipedia version: Global graph-based approach using a combined score of different relatedness measures as edge weights and local features as vertex weight. To select concepts a weighted clique algorithm is used.-Mapping of ids (from disambiguation) to the respective language (i.e. ja) using crosslanguage links and other mapping techniques
QUT_PNM_JA	This is a run using the PNM algorithm, and the cross-lingual title-to-target table is genereated from the NTCIR 9- Crosslink: Japanese Wikipedia Corpus
QUT_TRANSLITERATION_JA	The Stanford Named Entity Recogniser is used with the included 4-class CoNLL 2003 Shared Task model to identify named entities. After extracting these from the text, each named entity is translated using Google Translate, and if this fails then a potential list of transliterations is calculated (using a custom-written transliteration module). All terms are then passed to Wikipedia's Japanese search engine to identify suitable pages to link to. The top result for each is considered the best link, and this link is followed and a further 4 links gathered in ascending order. If for some reason 5 links can't been gathered then the remaining search results are considered until all options are expired. XXX At present, the code produces more than 250 links per page. This needs to be fixed
kslab_nut_A2F_E2J_01	This is a sample run using a simple link discovery algorithm
UKP_E2J_A2F_01	ngram.anchorprobability.cascade_3.titlematch_1_1
UKP_E2J_A2F_02	ngram.anchorprobability.cascade_3.titlematch_3_3
UKP_E2J_A2F_03	ngram.anchorprobability.interlingualindex.titlematch_1_1
UKP_E2J_A2F_04	ngram.anchorprobability.interlingualindex.cascade_3_1
UKP_E2J_A2F_05	ngram.anchorprobability.cascade_3.incominglinksearch_1_1

Table 27. CLLD system descriptions of submitted runs in English-to-Korean subtask

HITS_E2K_A2F_01	The results are produced using the following approach: -Preprocessing: Parsing of the topic xml files, tokenization (opennlp tools), tagging (tree-tagger)-Term recognition: -Disambiguation relative to enhanced English Wikipedia version: Global graph-based approach using a combined score of different relatedness measures as edge weights and local features as vertex weight. To select concepts a weighted clique algorithm is used.-Mapping of ids (from disambiguation) to the respective language (i.e. ko) using crosslanguage links and other mapping techniques
HITS_E2K_A2F_02	The results are produced using the following approach: -Preprocessing: Parsing of the topic xml files, tokenization (opennlp tools), tagging (tree-tagger)-Term recognition: -Disambiguation relative to enhanced English Wikipedia version: Global graph-based approach using a combined score of different relatedness measures as edge weights and local features as vertex weight. To select concepts a weighted clique algorithm is used.-Mapping of ids (from disambiguation) to the respective language (i.e. ko) using crosslanguage links and other mapping techniques
HITS_E2K_A2F_03	The results are produced using the following approach: -Preprocessing: Parsing of the topic xml files, tokenization (opennlp tools), tagging (tree-tagger)-Term recognition: -Disambiguation relative to enhanced English Wikipedia version: Global graph-based approach using a combined score of different relatedness measures as edge weights and local features as vertex weight. To select concepts a weighted clique algorithm is used.-Mapping of ids (from disambiguation) to the respective language (i.e. ko) using crosslanguage links and other mapping techniques
IISR_ko_title_to_key_name	Combine ko translated term with Wikipedia en title discovery algorithm
IISR_kttkn_to_ztttzt	Combine singular anchor with multi-languages title discovery algorithm (method 3)
IISR_singular_term_to_en_title	Mapping singular anchor and en translated term algorithm
IISR_sttent_to_kttkn	Combine singular anchor with multi-languages title discovery algorithm (method 2)
IISR_sttent_to_ztttzt_to_kttkn	Combine singular anchor with multi-languages title discovery algorithm (method 4)
KSLP_A2F_E2K_01	This is a run using keyphraseness, lesk-like crosslingual anchoring algorithm
KSLP_E2K_02	This is a run using stepwise translation dictionaries, and contextual WSD with 2nd-order cooccurrences
KSLP_E2K_03	This is a run using stepwise translation dictionaries, and a combination of definitional and contextual WSD with 2nd-order cooccurrences
KSLP_E2K_04	This is a run using stepwise translation dictionaries, and IR-based unified crosslingual anchoring with 2nd-order cooccurrence
KSLP_E2K_05	This is a run using stepwise translation dictionaries, and IR-based unified crosslingual anchoring with 3rd-order cooccurrence
QUT_PNM_KO	This is a run using the PNM algorithm, and the cross-lingual title-to-target table is generated from the NTCIR 9- Crosslink: Korean Wikipedia Corpus
DUIIS_A2F_E2K_4Pre	Using English keyword link probability, Korean keyword link probability, and English-Korean dictionary
DUIIS_A2F_E2K_4Rec	Using English keyword link probability, Korean keyword link probability, and English-Korean dictionary
UKP_E2K_A2F_01	ngram.anchorprobability.cascade_3.titlematch_1_1
UKP_E2K_A2F_02	ngram.anchorprobability.cascade_3.titlematch_3_3
UKP_E2K_A2F_03	ngram.anchorprobability.interlingualindex.titlematch_1_1
UKP_E2K_A2F_04	ngram.anchorprobability.interlingualindex.cascade_3_1
UKP_E2K_A2F_05	ngram.anchorprobability.cascade_3.incominglinksearch_1_1

8. REFERENCES

[1] R. Schenkel, *et al.*, "YAWN: A Semantically Annotated Wikipedia XML Corpus."

[2] INEX. (2010, *INEX 2010 Link-The-Wiki Task and Result Submission Specification* Available: http://www.inex.otago.ac.nz/tracks/wiki-link/runsubmission.asp?action=specification

[3] W. C. Huang, *et al.*, "The importance of manual assessment in link discovery," presented at the Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, Boston, MA, USA, 2009.

[4] D. Huang, *et al.*, "Overview of INEX 2007 Link the Wiki Track," in *Focused Access to XML Documents*, ed, 2008, pp. 373-387.

[5] W. Huang, *et al.*, "Overview of the INEX 2008 Link the Wiki Track," in *Advances in Focused Retrieval*. vol. 5631, S. Geva, *et al.*, Eds., ed: Springer Berlin / Heidelberg, 2009, pp. 314-325.

[6] W. Huang, *et al.*, "Overview of the INEX 2009 Link the Wiki Track," in *Focused Retrieval and Evaluation*. vol. 6203, S. Geva, *et al.*, Eds., ed: Springer Berlin / Heidelberg, 2010, pp. 312-323.

[7] W. C. Huang, *et al.*, "An Overview of INEX 2009 Link the Wiki Track," ed: http://www.inex.otago.ac.nz/data/proceedings/INEX2009-preproceedings.pdf, 2009.

[8] T. Sakai, *et al.*, "Overview of NTCIR-8 ACLIA IR4QA," in *Proceedings of NTCIR-8*, Tokyo, Japan, 2010, pp. 63-93.