The Effectiveness of Cross-lingual Link Discovery

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

[1]Faculty of Science and Technology,

Queensland University of Technology,
Brisbane, Australia

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

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

andrew@cs.otago.ac.nz

ABSTRACT

This paper describes the evaluation in benchmarking the effectiveness of cross-lingual link discovery (CLLD). Cross-lingual link discovery is a way of automatically finding prospective links between documents in different languages, which is particularly helpful for knowledge discovery of different language domains.

A CLLD evaluation framework is proposed for system performance benchmarking. The framework includes standard document collections, evaluation metrics, and link assessment and evaluation tools. The evaluation methods described in this paper have been utilised to quantify the system performance at NTCIR-9 Crosslink task. It is shown that using the manual assessment for generating gold standard can deliver a more reliable evaluation result.

Categories and Subject Descriptors

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

H.3.4 [**Information Systems**]: Systems and Software – performance evaluation (efficiency and effectiveness).

General Terms

Experimentation.

Keywords

Wikipedia, Cross-lingual Link Discovery, Assessment, Evaluation Framework, Assessment Tool, Evaluation Metrics.

1. INTRODUCTION

Wikipedia is an online multilingual encyclopaedia that contains a very large number of detailed articles covering most written languages. It is often considered to be a treasury of human knowledge. It includes extensive hypertext links between documents of same language for easy navigation.

However, the pages in different languages are rarely cross-linked except for direct equivalent pages (on the same subject) in different languages. This could pose serious difficulties to users seeking information or knowledge from different lingual sources, or where there is no equivalent page in one language or another.

Figure 1 shows several different language versions of the page on "Custard". Note that: 1) anchors are largely linked to articles in the source languages; 2) not all cross-language equivalent links exist – the English article "Custard" is not linked to the Italian custard article "Crema pasticcera", and *vice versa*; 3) some cross-language equivalent links are incorrect – the Chinese custard article "奶黄" is incorrectly linked to the Italian pudding article "Budino", and *vice versa*.

For English there are several mono-lingual link discovery tools. These help topic curators discover and maintain appropriate anchors and targets to add to a given document. No such tools yet exist, to support linking across multiple languages. The example in Figure 1 shows a need for this.

By contrast to monolingual link discovery, cross language link discovery (CLLD) algorithms actively recommend a set of meaningful anchors in a source document and establish links to documents in an alternative language. In other words, cross-lingual link discovery is a way of automatically finding hypertext links between documents in different languages.

The contribution herein is an evaluation framework for CLLD. This framework was put in place at NTCIR-9 and experiences from this are presented. The task was known as NTCIR-9 Crosslink [1]. Different from the overview of the Crosslink task paper, in this paper the evaluation methodology and metrics for cross-lingual link discovery are particularly examined, the evaluation framework is re-examined, and the effectiveness of cross-lingual link discovery is discussed.

The remainder of this paper is organized as follows: First the assessment challenges are discussed in Section 2. The evaluation framework is presented in Section 3. Manual assessment is discussed in Section 4. The effectiveness of CLLD methods is discussed in Section 5. We conclude and discuss future work in Section 6.

2. ASSESSMENT CHALLENGE

2.1 Link in Wikipedia

An anchor is a snippet of text that is relevant to the topic of the article and should be linked to a related article so that the reader can gather further information (or receive an explanation). Wikipedia anchors are often manually chosen and can target only one destination page.

But there are four types of links in Wikipedia:

- mono-lingual article-to- article (see also) links;
- mono-lingual anchor-to-article links;
- cross-lingual article-to-article (language) links;
- cross-lingual anchor-to-article links.

Wikipedia links are usually monolingual; the target page is in the same language as the source page and the anchor. Although article-to-article cross lingual links are not uncommon (they are listed on the left hand side of Wikipedia web pages as "languages" links), cross-lingual links from anchor to destination are rare.

HTML supports linking independent of language (indeed, it does not know anything about the language (or otherwise) of the target), but there are two fundamental problems that have inhibited the evolution of cross language linking. First, considerably greater effort is required to find link targets in a second (or subsequent) language (recall most links are inserted manually by humans). Second, none of the dominant web browsers directly support multiple links per anchor (although they can be programmed to do so), so each anchor is linked to a single target typically in the source language and there is no native way to add alternative (language) targets to the same anchor - monolingual linking is preferred on the assumption that the reader would prefer to stay in a single language. Currently, adding multiple language targets to a single anchor requires human support to find the links and browse support to display them – none of which exist.

Looking beyond these existing limitations, it is clear that multiple targets per link is beneficial – it can be seen on many social media websites such as Facebook where a user clicks an icon and is presented with a pop-up menu. It is also clear that cross-lingual hypertext linking is beneficial for (at the very least) language versions of Wikipedia that are sparse in coverage – although there are currently 3.8 million English Wikipedia articles, there are only 6,859 in Māori. It is unreasonable (even unethical) to restrict access to knowledge simply on a lingual basis. The study of *anchored cross-lingual linking* with multiple targets (one-to-many linking) is an essential addition to the Wikipedia.

2.2 Cross-lingual Link

Cross-lingual link discovery consists of two phases: 1) detecting prospective anchors in the source document; and 2) identifying relevant articles in the target language.

Although there is no hard limit to the number of anchors that may be inserted into a document, a user will become overwhelmed if every single term in an article is also an anchor – and so for evaluation purposes a limit of 250 anchors per document was imposed. Wikipedia currently supports 282 languages, but for evaluation purposes only up to 5 targets were allowed per anchor. In total this makes up-to 1250 outgoing links per article. Although the Wikipedia is a constantly evolving collection, for evaluation purposes a snapshot in a small number of languages (Chinese, Japanese, and Korean (CJK)) was taken. It is important to stress that these are the first experiments in CLLD and so such restrictions are not unreasonable.

The evaluated links list for a document can be symbolized as:

$$a_i \rightarrow (d_1, d_2, \ldots d_n) \text{ with } i \leq 250 \text{ and } n \leq 5$$

where a_i is the i^{th} anchor in the source document, and d_i is the cross-lingual CJK target document j for the anchor.

For the evaluation, the source language was chosen as English and the task was to identify the most relevant anchors and for each anchor the most relevant targets in CJK. Both anchors and targets were ordered on relevancy.

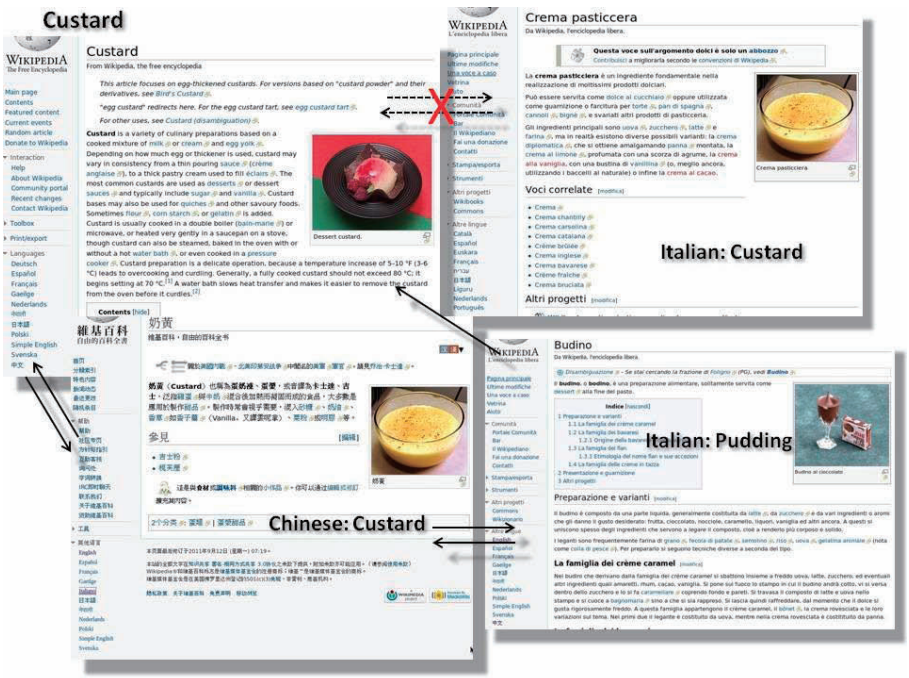

Figure 1: Lost in translation

2.3 Link Assessment

Evaluating Link Discovery is awkward because of the number of degrees of freedom. The algorithm must identify relevant anchors and relevant target articles.

Each anchor might occur several times within the source article, and in subtly different linguistic forms. It is unreasonable to score each instance in a single run, but also unreasonable not to score different linguistic variants in different runs. The best approach to measuring this imprecision is currently unclear but has been studied in the INEX Link Discovery Track [2-4] (where it changed from year to year).

However, what was discovered at INEX was that performance could be measured in two possible ways: automatically (using the Wikipedia itself as the ground truth); and manually (in the TREC paradigm)

2.3.1 Automatic Assessment

In automatic assessment the ground-truth (*qrel* set) is derived from links already in Wikipedia articles through triangulation. These come from two sources: First, all the mono-lingual links from the translation of the source article are considered relevant. Second, all the cross-lingual links from the mono-lingual links from the source article are considered relevant. For instance, if an English article is "Martial Art" then the relevant Chinese links are those links out of the Chinese "Martial Art" (武術) article, and the Chinese counterpart for all links out of the English article.

It is accepted that the ground truth may be incomplete. For example, it may contain links from the English version to which there is no appropriate anchor the CJK versions. It may also not contain the kinds of links that users click. However, it is reasonable to believe that this will not adversely affect the relative rank order of CLLD systems.

Evaluation using automatic assessment is likely biased towards links already in Wikipedia. Huang *et al* [5] suggest that manual assessment of monolingual link discovery could result in substantially different results.

2.3.2 Manual Assessment

An alternative approach to automatic assessment is manual assessment in the style of TREC. Human assessors are employed to examine each recommended link and to make a judgement call on the relevance (or otherwise) of the target. Human assessors can, of course, judge not only the link but can individually assess the relevance of the anchor and the target – a target might be relevant but the anchor not so (and *vice versa*).

In addition to the usual criticisms of the imperfection of evaluation using manual assessment in traditional information retrieval tasks (e.g. various tracks at INEX or TREC), there are new issues with cross-lingual link discovery. For example, it is difficult to find assessors skilled enough to read the source articles in one language (in this case English) while also having a good understanding of the target document and its alternate language. Note that the skill level (in both languages) required to do this is higher than that of an ordinary user reading both document.

For manual assessment the anchors and targets are shown to a human. For this reason it's important to include the offset and length of the anchor in the run. For the experiments herein the zero-based byte offset was used (lengths were also specified in bytes). As a validity check this was compared against the text that was also specified in the run (invalid matches were discarded). As can be expected, pooling was used.

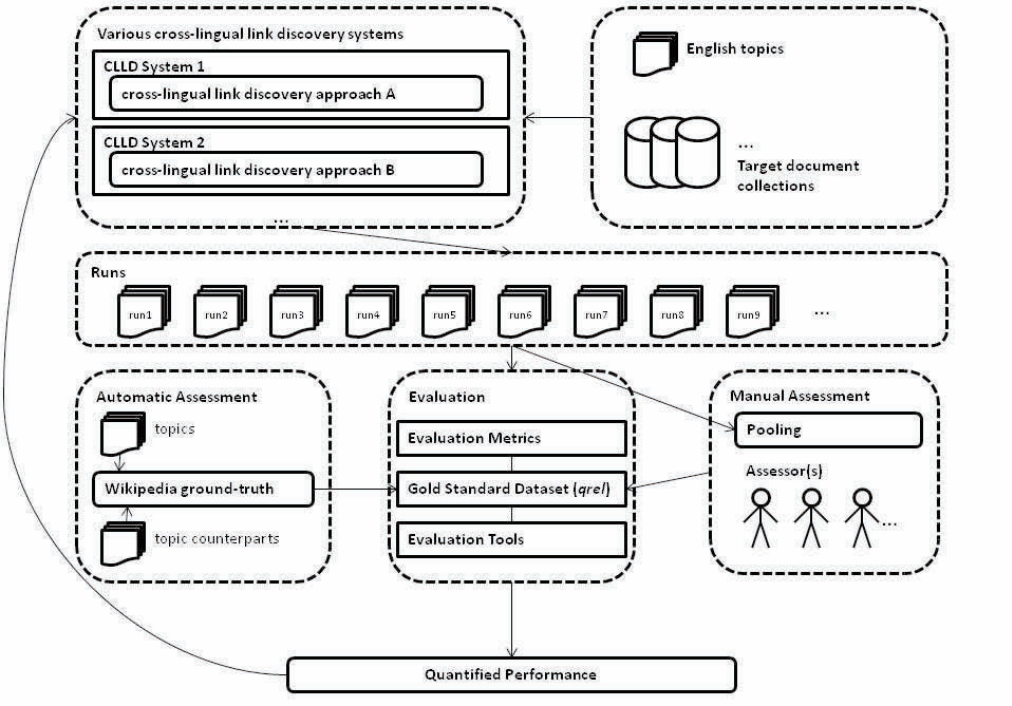

Figure 2: The Cross-lingual Link Discovery evaluation methodology

Table 1 Article statistics of Chinese and English Corpora

Corpus	Articles	Cross-lingual links
English	3,484,250	169,974 (en→zh, 4.9%) 292,548 (en→ja, 8.4%) 87,367 (en→ko, 2.5%)
Chinese	316,251	170,637 (zh→en, 54.0%)
Japanese	715,911	289579 (ja→en, 40.4%)
Korean	201,512	89230 (ko→en, 44.3%)
Total	4,717,924	200,825

3. THE EVALUATION FRAMEWORK

The CLLD evaluation framework [6] is a multi-lingual adaptation of the INEX Link-the-Wiki track [4] framework. The evaluation methodology is illustrated in Figure 2 and discussed in this section. The methodology was used in the NTCIR-9 Crosslink task [7] with source articles in English targeting links to Chinese, Japanese, and Korean (CJK).

3.1 Document Collections

The Wikipedia is an excellent collection to use for CLLD because it is a (mostly) closed hypertext collection and exists in several languages. Articles can be re-distributed for experiments under Creative Commons Attribution-Share-Alike License 3.0 [8], and so copyright issues are minimal.

June 2010 dumps of Wikipedia were downloaded and converted into XML using the YAWN system [9]. A multi-lingual adaptation of the Schenkel *et.al* [9] Java YAWN program was used to insert the XML structure.

The document collection is summarised in Table 1. Column 1 lists the language, column 2 the number of articles in that collection, and column 3 the number of links to English. For example, after conversion there were 316,251 Chinese articles of which 170,637 contained links to English articles.

3.2 Topics

A set of 25 articles were randomly chosen from the English Wikipedia and used as test topics for the evaluation. All test topics had their pre-existing links removed – a process known at INEX as orphaning.

3.3 Evaluation Methods

Evaluation was performed according to the already accepted INEX methods of file-to-file (F2F) and anchor-to-file (A2F).

In file-to-file evaluation the performance of the link discovery algorithm at finding articles that should be linked-to is measured regardless of the relevancy of anchors themselves. For instance, if a word "cooked" in the Custard article is marked irrelevant in assessment but the anchor is correctly linked to "pressure cooker" which is in *qrel*, so this link will be still considered relevant. F2F evaluation is perfect for automatic assessment of CLLD because appropriate anchors cannot necessarily be extracted from the corpus whereas appropriate target articles can.

In anchor-to-file evaluation, the correctness of anchors is also considered. Any given anchor can either be relevant or not-relevant to the article; if an anchor is not-relevant then under evaluation the link is considered non-relevant even if the target article is relevant. The term *non-relevant* in reference to anchor text means that a user will not see a need for the anchor text to be linked. Either because it does not describe an important concept in the context where it appears, or it is simply considered trivial. Manual assessment should be used to get a good ground truth for anchor-to-file evaluation, and this ground truth can also be used for file-to-file evaluation.

3.4 Metrics

For the experiments herein, *Precision-at-N, R-Prec,* and *Mean Average Precision* (*MAP*) were the metrics used to quantify performance. As with other traditional information retrieval evaluation evaluations, *precision* and *recall* are the two underlying key metrics to measure performance. But for CLLD precision and recall are computed subtly differently for the two evaluation methods (F2F and A2F).

3.4.1 Precision and Recall

File-to-File Evaluation

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

and,

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

The precision and recall are computed at each anchor (recall that 5 targets per anchor were permitted).

Anchor-to-File Evaluation

For anchor-to-file evaluation a similar precision definition to that used in INEX 2009 [10] was used. Both precision of anchor and precision of target are considered. The score of the anchor is defined as:

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

That is, if anchor i is relevant and it has at least one relevant target, then $f_{anchor}(i) = 1$. Otherwise the score is 0.

For each target of an anchor, if that target document, j, is relevant to the anchor then it receives a link score f_{link} of 1, otherwise 0 thus:

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

Finally, the anchor-to-file precision with respect to an article is:

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

$$Recall_{a2f} = \left(\sum_{i=1}^{n}\left(f_{anchor}(i)\right) \times \frac{\sum_{j=1}^{k} f_{link}(j)}{k_i}\right)/N \quad (9)$$

Where n is the number of identified anchors; N is the number of anchors in *qrel*; k is the number of returned targets for anchor i; and k_i is the number of targets recommended for this anchor.

3.4.2 System Evaluation Metrics

For both evaluation types, *MAP* is defined as:

$$MAP = (\sum_{t=1}^{n} \frac{\sum_{k=1}^{m} p_{kt}}{m})/n \quad (10)$$

where n is the number of topics (source articles used in evaluation); m is the number of identified items (articles for F2F or anchors in A2F); and P_{kt} is the precision of the top K items for topic t.

R-Prec is defined as:

$$R - Prec = \sum_{t=1}^{n} P_t @ R / n \quad (11)$$

where n is the number of topics; and $P_t@R$ is the precision at R where R is the number of unique items in the *qrel*s of topic t.

Similarly, *Precision-at-N* is computed using the average precision for all topics (source articles) at a pre-defined position N in the results list. Values of N were chosen as: 5, 10, 20, 30, 50, and 250.

4. MANUAL ASSESSMENT

4.1 Link Pooling for Assessment

In total 57 runs from 11 teams were submitted to the NTCIR-9 Crosslink task. General statistics of runs are broken down by language in Table 2. The first column shows the task, the second the number of submitted runs, the third shows the average number of links per topic. For example, in the English to Chinese run (En-2-Zh) there were 25 runs averaging 2969 links per topic. The other tasks were English-2-Japanese (En-2-Ja) and English to Korean (En-2-Ko).

Table 2 Average number of links in pooling

Task	Runs	Average links per topic
En-2-Zh	25	2969
En-2-Ja	11	666
En-2-Ko	21	924

All the prospective anchors and the corresponding targets were pooled. Unfortunately, anchors in some submitted runs did not pass the anchor validity check which is a hard requirement for all submissions; they were subsequently discarded from the pool. The pool was assessed to completion (there were no valid and unassessed links in any runs).

It took an assessor approximately one hour to finish assess one topic to completion. Each topic was assessed by a single assessor but assessors could complete multiple topic assessments.

During assessment the assessor could mark either the anchor or the target as relevant or non-relevant. If an anchor was assessed as non-relevant then that anchor's target articles were assessed as non-relevant.

4.2 Overlapping Anchors

Due to different methods used in different systems for anchor identification, pooled anchors might be overlapped. There is no hard specification with respect to the relevancy of overlapped anchors. All overlapped anchors are still judged one by one, and it is up to assessor(s) to decide if an overlapped anchor or entire overlapped anchors are relevant.

The decision of whether or not providing overlapping anchors in articles will be up to the applications that realise cross-lingual link discovery in a knowledge base according to user's own preference. And this is not the focus of this paper.

4.3 Assessment Tool

An assessment tool was developed for the task. It is shown in Figure 3 with an English-to-Chinese link. In the left pane the English source document (the assessment topic) is shown. In the right pane the Chinese target document is shown. The assessor clicked either the right or left mouse button to mark the link as relevant or non-relevant.

With the tool, assessors inspected each anchor and its corresponding links, accepting or rejecting each. This method of assessment is not dissimilar to the assessment approaches used in CLIR evaluations, and is similar to that used at INEX.

4.4 The Wikipedia Ground-Truth Run

In the INEX (mono-lingual) Link-the-Wiki track, links in the source (topic) article were added to the assessment pool and manually assessed [4]. Due to the way the automatic assessments were generated in the NTCIR-9 Crosslink task, it was not possible to do this in these experiments. For example, manually assessing targets extracted through the triangulation methods outlined in section 2.3.1 would require assessment without anchors – something of questionable utility since anchor selection is part and parcel of CLLD.

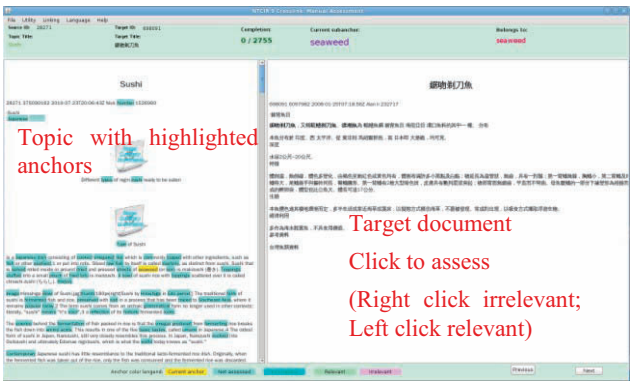

Figure 3: The NTCIR-9 Crosslink manual assessment tool

4.5 Human Assessors

Conveniently, there are many overseas students with differing backgrounds at Queensland University of Technology (QUT) each possessing (at least) bi-lingual skills. Overseas students make good assessors for two reasons: 1) they are highly educated and so can reasonably be expected to understand Wikipedia articles; and 2) they possess good language skills in their native language and English and so can be reasonably expected to be able to read articles in multiple languages. All assessors were compensated with movie tickets.

Table 3 Assessors information

Task	Assessors	Description
En-2-Zh	15	PhD students, some undergrads
En-2-Ja	1	Postdoc
En-2-Ko	5	Undergraduate students

Summary information on the assessors is given in Table 3. Column 1 lists the task; column 2 the number of assessors; and column 3 their level of education. For example, the 1 assessor for the English to Japanese task was a Postdoc.

It was initially thought that finding assessors for the English-to-Chinese task would be easy because of the relatively large proportion of Chinese students at QUT. However assessing all English-to-Chinese links was a challenge because:

- The English-to-Chinese task saw the largest average number of links per topic requiring more time to assess (approximately three hours per topic).
- Students considered themselves too busy to help.
- Motivation was low as compensation was low.

Due to the shortage of assessors three of the original topics were never assessed. Asking for runs on more topics than are finally assessed is a normal part of the INEX paradigm.

Finding assessors for the Japanese topics was also difficult, due to the lack of availability of English-Japanese bilingual speakers. We initially identified two additional volunteers but they eventually dropped out leaving only one assessor, an author of this paper.

Korean assessors were readily available in the form of undergraduate students, where they were given an instruction and assessed topics under the supervision of at least one author.

4.6 Links Found in Manual Assessment

A summary of the assessments is presented in Table 4. The first column lists the assessment type, the second column lists the number of unique links, and the third lists the overlap between the manual and automatic sets. For example, there were 1,681 links found through automatic triangulation of the 25 topics in the English to Korean Wikipedia, but 2,786 relevant links in the pool; the same pattern can be seen for English to Chinese. In the English to Japanese assessments the number of relevant link in the manual set is fewer than were automatically identified – this is most likely because the average number of links per topic was smaller because of the smaller pool size and the smaller number of submitted runs than is seen the other tasks.

Table 4 Links in the result sets of two assessments

Assessment set	Relevant links	Overlapping
En-2-Zh automatic	2116	1134
En-2-Zh manual	4309	
En-2-Ja automatic	2939	781
En-2-Ja manual	1118	
En-2-Ko automatic	1681	821
En-2-Ko manual	2786	

5. RUN PERFORMANCE

5.1 Evaluation Types and Measures

The English-to-Chinese subtask saw the most runs, the largest number of links per topic, and the largest number of relevant links; consequently the effectiveness of cross-lingual link discovery is discussed in that context herein (due to space limitation). For a full account of the experiments the reader is referred to the proceedings of NTCIR-9 [1].

File-to-file (F2F) evaluation was performed using both the automatic and manual assessment sets, but for reasons already given, anchor-to-file (A2F) evaluation could only be performed using the manual assessments. MAP was the preferred metric as it is well understood by the IR community.

5.2 Evaluation Results

Precision-recall curves for the English-to-Chinese runs are shown in Figure 4 and Figure 5. The scores of the top runs are shown in Table 5. From these figures and this table it can be seen that the top performing run under automatic file-to-file assessment was the HITS run [11], however HITS was outperformed by runs from UKP [12] and QUT when manual assessment was used.

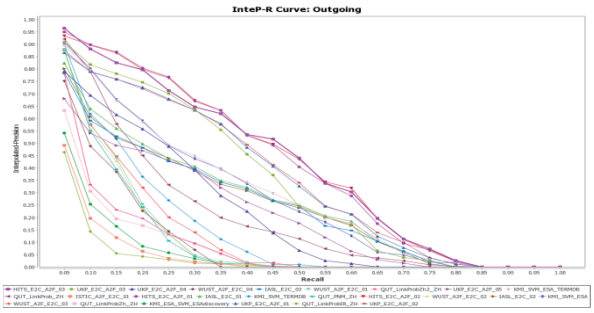

Figure 4: Interpolated precision-recall graph showing En-2-Zh F2F evaluation against the automatic assessments

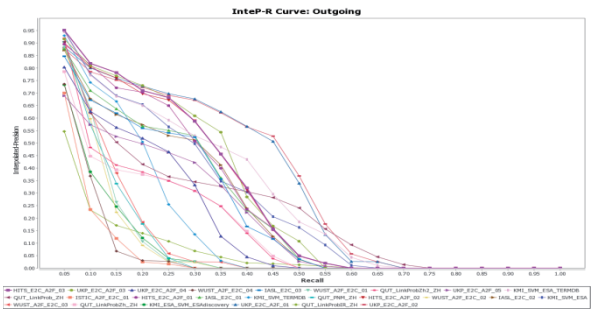

Figure 5: Interpolated precision-recall graph showing En-2-Zh A2F evaluation against the manual assessments

Table 5 MAPs of teams in two En-2-Zh evaluations

Automatic F2F evaluation		Manual A2F evaluation	
Run ID	**MAP**	**Run ID**	**MAP**
HITS	0.373	UKP	0.157
UKP	0.314	QUT	0.115
KMI	0.260	HITS	0.102
IASL	0.225	KMI	0.097
QUT	0.179	IASL	0.037
WUST	0.108	WUST	0.012
ISTIC	0.032	ISTIC	0.000

This does not necessarily mean that one run is better than another; it means that under different evaluation paradigms the runs exhibit different orderings. The HITS run is the best at identifying links similar to those already present in the Wikipedia, but the UKP run is better at identifying links with topical relevance to the source article.

However, the topical relevance of both the anchors and the targets are considered in anchor-to-file evaluation and so it is rea-

sonable to conclude that it is a more rigorous evaluation method (and hence *MAP* scores are lower). But, automatic evaluation does not require assessors and could be conducted on a large number of topics (source documents) making it more comprehensive.

5.3 Cross-Language Agreement

As discussed in Section 4.5, the manual assessment environment varies across different language topics (different number of assessors; different number of topics; and different skill base of assessors).

However, two groups, HITS and UKP, produce runs that consistently ranked higher than the others regardless of language subtask [1]. Both groups implemented their algorithms as language independent and then submitted runs for each subtask language. That is, HITS, for example, used the same algorithm for the Chinese, Japanese, and Korean subtasks. This is convenient as it makes it possible to study performance of the same algorithm across the three different languages.

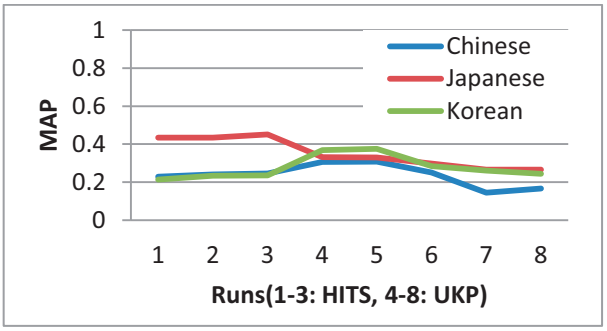

Figure 6: Performance of HITS and UKP, manual F2F

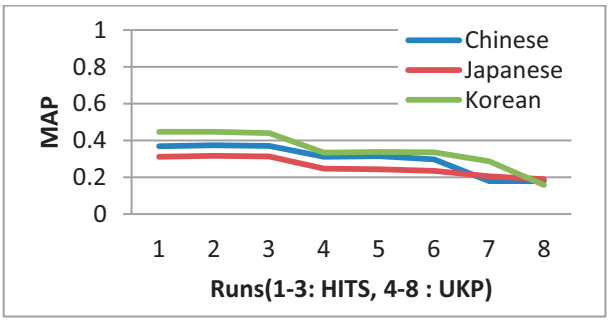

Figure 7: Performance of HITS and UKP, automatic F2F

The performance of the 3 runs from HITS and the 5 runs from UKP measured under manual file-to-file assessment is shown in Figure 6. Runs 1, 2, and 3 are the HITS runs and 4 to 8 are the UKP runs. The same runs in the same order are shown under automatic file-to-file assessment in Figure 7. From Figure 7 it can be seen that HITS runs perform best at Korean, then Chinese, then Japanese, and UKP runs similarly but not consistently. Under manual file-to-file assessment (Figure 6) the HITS runs perform better at Japanese, then Chinese, then Korean, and less consistency is seen in the UKP runs, however all algorithms get better scores for Japanese than for Chinese and score improvements seen in Korean are reflected with score improvements in Chinese.

A *t*-test between the Japanese runs scored automatically and manually shows a significant difference at 1%, but no significant difference (even at 5%) is seen for the Chinese and Korean runs. That is, for Japanese there is a significant difference between the performance depending on whether manual or automatic assessment is used, but no such difference is observed for Chinese of Korean.

5.4 Unique Relevant Links

The performance in terms of precision has been discussed in the previous sections. In this section the breadth of the algorithms is discussed. That is, an algorithm that correctly identifies links similar to those already in the collection is of interest, but so too are algorithms that identify novel (and relevant) links not present (even at the expense of some prevision). This is the traditional precision / recall trade off is well known in IR, but of particular interest in CLLD because it allows the hyperlink graph to grow in new unconstrained ways and (perhaps) in ways more useful to a user than simply taking then to the same familiar places.

Table 6 presents the statistics of unique relevant links discovered in the NTCIR-9 Crosslink runs. The first column lists the assessment set (automatic or manual), the second column lists the total number of unique links discovered across all runs, and the third column lists the most diverse group and the number of unique and relevant links they identified. For example, UKP found 97 of the 245 unique relevant links found in all the runs, which in turn amounts to 11.6% of the relevant links in the automatically extracted assessment set.

Of particular note is that in QUT runs [13] contribute 1103 unique relevant links to the manual assessment set, which is about 79% of total unique relevant links found. This suggests that the QUT runs is the most diverse, preferring its own links to those already present in Wikipedia. Without manual assessment no knowledge of the relevance of these links would have been revealed.

Table 6 Unique relevant English-to-Chinese links

Compared with	Total (%)	Team with highest (#)
Automatic	245 (11.6%)	UKP (97)
Manual	1397 (32.4%)	QUT (1103)

5.5 Discussion: CLLD in Action

In Section 5.2 and 5.3, the performance of the runs is quantified in various ways, but how good these systems? Can users satisfy their information needs?

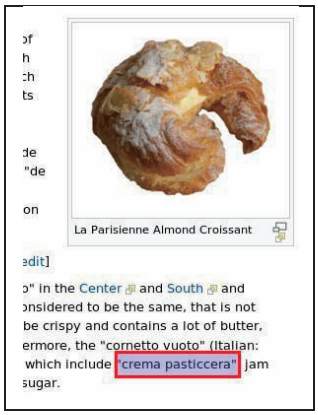

Figure 8: An anchor found for the topic "Croissant"

Figure 8 shows a snippet of the English Wikipedia article "Croissant". The boxed text reads “crema pasticcera” which is linked neither to the English “Custard” article nor the Italian “Crema pasticcera” article. When a multi-lingual user encounters across such a term they may ask:

- What is “crema pasticcera”?
- Is it in my own language or languages I read?

These questions cannot easily be answered without further manual searching of Wikipedia or translation using a translation service. Recall from Figure 1 (Section 1) that finding information for “crema pasticcera” in Wikipedia is very difficult because:

1. Users may not know this is Italian, even if they did,
2. The Italian article “Crema pasticcera” is not linked to the English article “Custard” (or the Chinese article “奶黄”).

Among all the submitted runs KMI’s runs [14] uniquely recommended “crema pasticcera” and correctly linked it to the Chinese article “奶黄”. KMI have developed an algorithm that solved the problem of our running example. In doing so they have also demonstrated the power (and importance) of CLLD in correctly expanding the Wikipedia to include cross lingual links.

However, it should also be noted that without manual assessment this link would not have been assessed as relevant – and so their run additionally demonstrates the importance of manual assessment. In fact, the *MAP* scores of the official evaluation results [1] show that their run KMI_SVM_ESA_TERMDB was not very effective when measured using automatic file-to-file assessments (scoring 7th) but was effective (pacing 3rd) measured using manual file-to-file assessments.

6. CONCLUSION AND FUTURE WORK

In this paper we describe the motivation for cross-lingual link discovery in a knowledge base such as Wikipedia. We also presented the assessment challenges, and the difficulties of assessor recruitment, along with our experiences and results from running the NTCIR-9 crosslink track.

We focused on effective system evaluation and explored automatic and manual evaluation. For automatic evaluation the ground-truth qrels were extracted from the Wikipedia through triangulation. Manual assessments were performed by multi-lingual assessors with a high level of education. Evaluation was both file-to-file and anchor-to-file. It is suggested that manual assessment could result in a more thorough evaluation but automatic assessment in a broader evaluation.

Evaluation of the runs shows that some of the algorithms used at NTCIR-9 were effective, finding links already in Wikipedia as well as previously unseen links, however no single algorithms was best at both. Some algorithms produced a disproportionally large number of unique relevant links suggesting that those teams focused on diversification in their result sets.

Further work is needed by these groups to produce algorithms capable of reliably identifying a large numbers of diverse and relevant cross-language links – and we expect to see such work at NTCIR and other evaluation forums in the near future.

7. REFERENCES

[1] L.-X. Tang, *et al.*, "Overview of the NTCIR-9 Crosslink Task: Cross-lingual Link Discovery," in *Proceedings of NTCIR-9*, to appear, 2011.

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

[3] 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.

[4] 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.

[5] 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.

[6] (2011, *Crosslink Project*. Available: http://code.google.com/p/crosslink/

[7] NTCIR. (2011, *Cross-Lingual Link Discovery Task*. Available: http://ntcir.nii.ac.jp/CrossLink/

[8] *Creative Commons Attribution-Share-Alike License 3.0*. Available: http://creativecommons.org/licenses/by-sa/3.0/

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

[10] 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.

[11] A. Fahrni and M. Strube, "HITS’ Graph-based System at the NTCIR-9 Cross-lingual Link Discovery Task," in *Proceedings of NTCIR-9*, to appear, 2011.

[12] J. Kim and I. Gurevych, "UKP at CrossLink: Anchor Text Translation for Cross-lingual Link Discovery," in *Proceedings of NTCIR-9*, to appear, 2011.

[13] L.-X. Tang, *et al.*, "Automated Cross-lingual Link Discovery in Wikipedia," in *Proceedings of NTCIR-9*, to appear, 2011.

[14] P. Knoth, "KMI, The Open University at NTCIR-9 CrossLink," in *Proceedings of NTCIR-9*, to appear, 2011.