An Evaluation Framework for Cross-lingual Link Discovery

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

†Faculty of Science and Technology, Queensland University of Technology, Brisbane, Australia

‡Department of Computer Science, University of Otago, Dunedin, New Zealand

* National Institute of Informatics, Japan

Abstract

Cross-Lingual Link Discovery (CLLD) is a new problem in Information Retrieval. The aim is to automatically identify meaningful and relevant hypertext links between documents in different languages. This is particularly helpful in knowledge discovery if a multi-lingual knowledge base is sparse in one language or another, or the topical coverage in each language is different; such is the case with the Wikipedia. Techniques for identifying new and topically relevant cross-lingual links are a current topic of interest at NTCIR where the CrossLink task has been running since the 2011 NTCIR-9. This paper presents the evaluation framework for benchmarking algorithms for cross-lingual link discovery evaluated in the context of NTCIR-9.

This framework includes topics, document collections, assessments, metrics, and a toolkit for pooling, assessment, and evaluation. The assessments are further divided into two separate sets: manual assessments performed by human assessors; and automatic assessments based on links extracted from the Wikipedia itself. Using this framework we show that manual assessment is more robust than automatic assessment in the context of cross-lingual link discovery.

Keywords: Wikipedia, Cross-lingual Link Discovery, Evaluation Framework, Validation, Assessment, Evaluation Metrics

1. Introduction

The Wikipedia is an online multilingual encyclopaedia containing a very large number of articles written in most modern written languages. It is extensively hypertext linked between related articles in the same language, and between equivalent (same-topic) articles in different languages. However, there are few links between related articles in different languages. This poses a problem to a multi-lingual knowledge-seeking user for several reasons:

First, a user may be trying to acquire language specific words for a domain in which they are familiar. As an example, a martial artist might be interested in acquiring the Chinese, Japanese, or Korean words for the arts in which they are trained. Figure 1 shows a snippet of an English Martial Arts Wikipedia article where it can be seen that the English anchors are linked only to related English articles; links to related Chinese, Japanese, and Korean articles are not present.

Second, a user may wish to understand colloquialisms from an acquired second language. As an example, 花蟹 (“flowery crab”) is in common use in Hong Kong to refer to the ten-dollar banknote. The use

[1] E-mail addresses: {l4.tang, s.geva, yue.xu}@qut.edu.au (Ling-Xiang Tang, Shlomo Geva and Yue Xu), andrew@cs.otago.ac.nz (Andrew Trotman), itakura@nii.ac.jp (Kelly Y. Itakura).

is non-obvious, and is given in the English Wikipedia article “Hong Kong dollar”; however neither this article nor the English article “Hong Kong ten-dollar note” are linked-to from the Chinese article “香港十元紙幣”. Figure 2 shows the English and Chinese articles about the Hong Kong ten-dollar note from which it can be seen that adding cross-lingual links would enhance the article.

Third, a user may simply require a translation. As an example, “crema pasticcera” might appear on a menu and a user might not know what this is. Figure 3 shows several different language versions of the Wikipedia “Custard” articles from where it can be seen that: anchors are typically linked to targets in the source language; not all cross-language equivalent-article links exist (e.g. English is not linked to Italian, and *vice versa*); Some cross-language equivalent-article links are incorrect (e.g., the Chinese article “奶黄” (custard) is linked to the Italian “Budino” (pudding) article).

This problem and its solution are exacerbated by the structure of the web. Web pages typically link from one source page to one target page. This might be acceptable for a mono-lingual web, but in the case of a multi-lingual web it is not. A multi-lingual user faced with reading content in a second (or subsequent) language might prefer to be directed to content in their first language if available. An example scenario is a user who starts reading the Wikipedia in Korean and follows a link to a Japanese article, not being directed back to Korean when content is available in that language. For multi-lingual link discovery it is necessary to extend the model of a link. Whereas before it was from one anchor to one target, it must now be from one anchor to multiple targets – each of which might be in a different language.

This leads to the task we are addressing: machine-assisted cross-language link-insertion into the Wikipedia. To do this a system for recommending links is needed. That system should take the Wikipedia and from it produce a *ranked* list of links for a human annotator to consider inserting (either manually or automatically). This machine task is known as cross language links discovery.

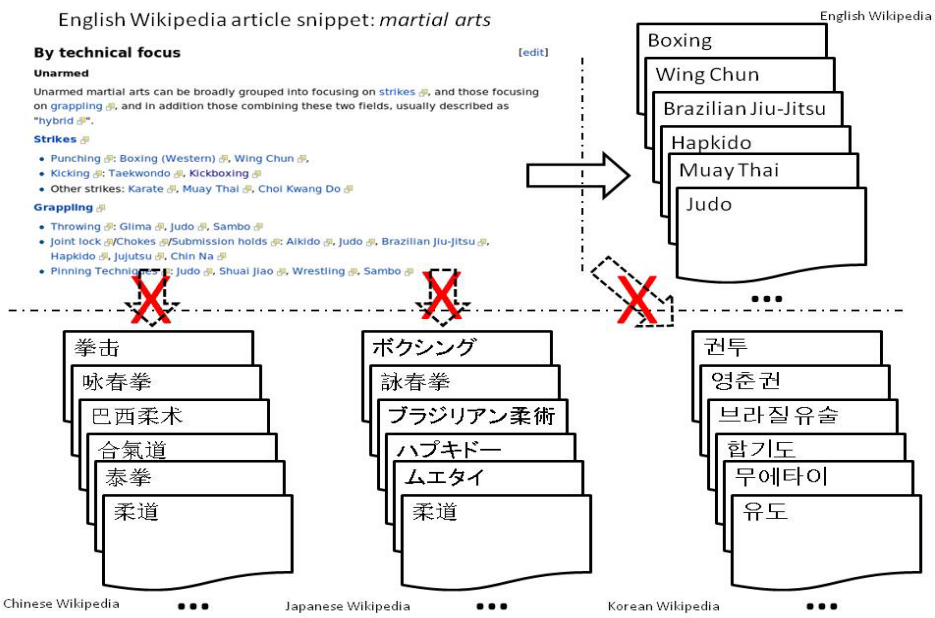

Figure 1: Cross-lingual links are often mising from Wikipedia articles on inherently multi-lingual topics

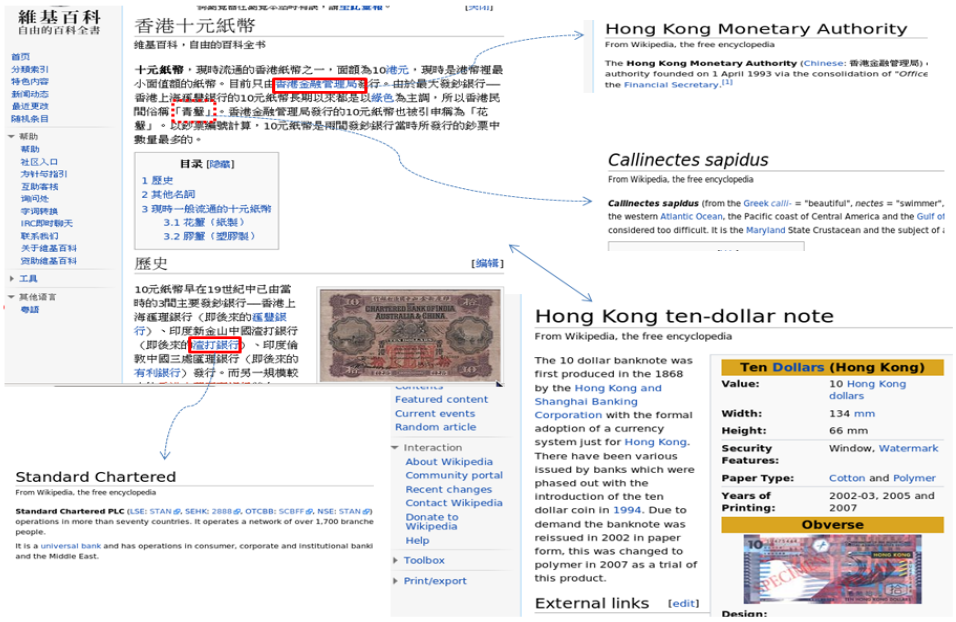

Figure 2: Colloquialisms are often not linked to multiple language explanations. In this case the Hong Kong "flower crab" is shown not to link to the English description

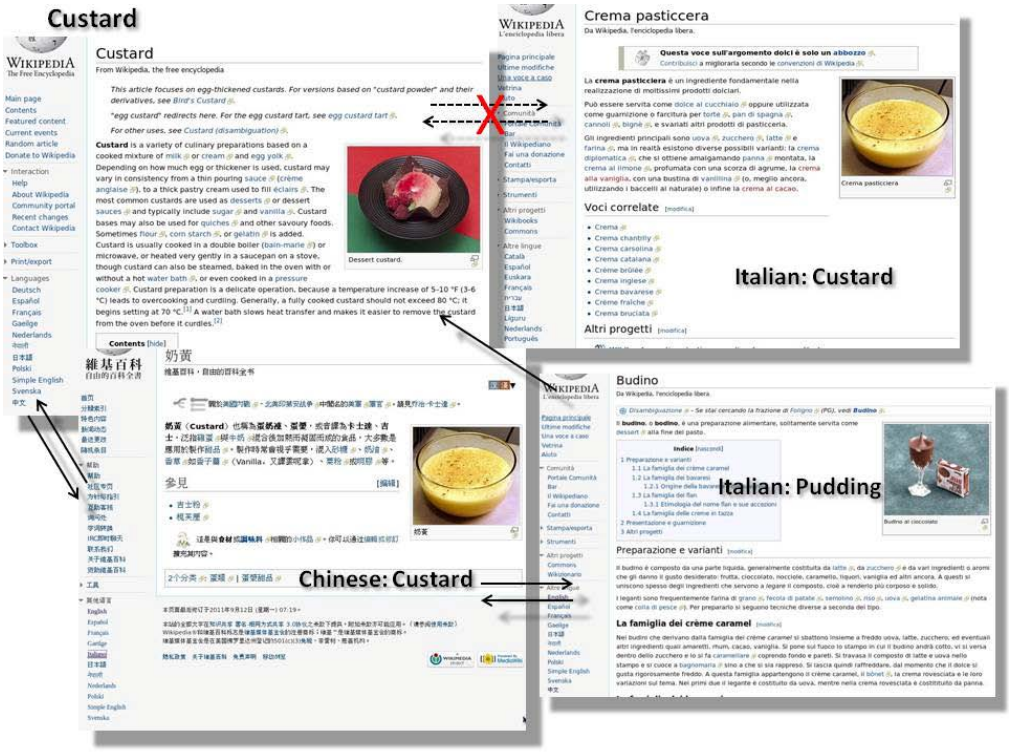

Figure 3: Some cross-language equivalent-articles links are incorrect (pudding is not always custard)

Herein we define cross language link discovery (CLLD) as recommending and ranking a set of meaningful anchors to place in a source document that point to topically related content in documents in a second or subsequent language.

Several English mono-lingual link discovery tools exist (e.g. Wikipedia Miner[2]). These tools help Wikipedia content creators and curators discover and maintain appropriate links for their articles. But none of these tools currently support cross-lingual link discovery. Prior work has examined the quality of mono-lingual link discovery, but none of these frameworks has examined the quality of cross-lingual link discovery. We set out to examine the problems associated with the performance measure of CLLD, and to build a re-usable test collection, by running the NTCIR-9 CrossLink track (Tang, Geva, Trotman, Xu, & Itakura, 2011). *In this paper we present an evaluation framework for CLLD and a detailed systematic examination of it in the context of NTCIR-9.* A preliminary examination was presented at EVIA 2011 (Tang, Itakura, Geva, Trotman, & Xu, 2011).

The remainder of this paper is organized as follows. Related work is outlined in section 2. The framework is given in section 3. Assessment is discussed in section 4. The toolset is discussed in section 5. The framework is applied in section 6. The effectiveness of CLLD methods is discussed in section 7. We conclude in section 8.

2. Related Work

Hypertext differs from text by the existence of navigational links between documents. Many methods of automatically generating hypertext from text have been proposed both academically and commercially and several patents exist (Malcolm, 2001; Sotomayor, 1998).

A common approach is to use information retrieval techniques to find links. Coombs (1990) discussed a system architecture for automatic document linking with the assistance of full-text search; and Allan (1996) presents methods for automatically linking documents based on information retrieval techniques.

Machine constructed links inserted for information retrieval purposes could improve the search and browse experience. Agosti & Crestani (1993) discuss a concept for building an automatic hypertext construction tool for the use of information retrieval; later they (Agosti, Crestani, & Melucci, 1996) present the design and implementation of such a tool named TACHIR. TACHIR implements term/document links through indexing, document/document links through citation graphs and statistical document/document similarity, concept/document links, and concept/concept links through the use of a thesaurus.

To build a hypertext for a specific knowledge domain, Kurohashi *et. al.* (1992) examine a method for automatic hypertext construction for the Encyclopaedic Dictionary of Computer Science. In their work (Kurohashi et al., 1992) natural language processing is used for locating meaningful conceptual phrases in the text corpus, which are used as anchors; links are then created for the discovered terms pointing to their descriptions. Crane, Smith, & Wulfman (2001) describe the creation of a humanities digital library with rich contents on London of pre-twentieth century. The inter collection “span-to-span” links, which are automatically generated by mining the historical “authority lists”, provide an easy way for information exploration. The authors (Crane, Smith, & Wulfman) highlight a feature of their system, and it can link Greek and Latin words to dictionary entries, which gained enormous popularity among students of the languages.

---

[2] http://wikipedia-miner.cms.waikato.ac.nz/

Many other methods (Allan, 1996; Salton, Allan, & Singhal, 1996; Smeaton & Morrissey, 1995), and those discussed above were proposed to tackle the automatic text structuring and link creation problem. With a detailed survey of these attempts, Agosti, Crestani, & Melucci (1997) suggest that without a proper evaluation methodology it is difficult to distinguish "good" or "bad" approaches.

More recently, Jihong & Bloniarz (2004) present a key-word based approach to generating intra-document and inter-document hyperlinks automatically for large collection of reference materials. Yang & Lee (2005) examine a text mining approach for hypertext construction with experiments on Chinese documents, but evaluation on this method is still relatively weak making it hard to compare their method with those previously proposed.

With the emergence of Wikipedia, automatic link discovery again became a current research topic, this time for the purpose of efficient knowledge discovery and link maintenance. Many methods have been proposed, such as that of Adafre & De Rijke (2005). Mihalcea & Csomai (2007) propose the *Wikify* system for the Wikipedia. Milne & Witten (2008) tested a learning-to-link method designed to link any document to the Wikipedia. Link mining approaches have also been proposed, including the work of He & de Rijke (2010), and Sorg & Cimiano (2008) use Support Vector Machines (SVM). Erbs, Zesch, & Gurevych (2011) provide a comprehensive analysis of state-of-the-art mono-lingual link discovery approaches.

Little work has been done on cross-lingual link discovery. Sorg & Cimiano (2008) discovered that only a small proportion of Wikipedia articles present in both German and English were linked to each other. They set out to identify the missing links using a classification-based approach. Melo & Weikum (2010) examine incorrect Wikipedia language-links between articles on the same topic. Yeung *et. al*. (2011) propose a framework for assisting cross-lingual editing in Wikipedia.

As far as we know no tools have been developed or reported to suggest cross-lingual links and no systematic evaluation of algorithms has been performed or reported – we study these in this contribution. INEX ran the mono-lingual Link-the-Wiki track for several years and while doing so explored various evaluation frameworks. The performance of link discovery systems under two different of assessment criteria (automatic and manual) was examined (D. Huang, Xu, Trotman, & Geva, 2008; W. Huang, Geva, & Trotman, 2009, 2010). Under automatic assessment the suggested links were compared to the Wikipedia itself, but under manual assessment a Cranfield paradigm experiment was conducted with human assessors. Substantial performance differences depending on the assessment method were identified (W. C. Huang, Trotman, & Geva, 2009). The Link-the-Wiki track of INEX 2010 focused on linking the Te Ara[3] encyclopaedia (Trotman, Alexander, & Geva, 2011), an encyclopaedia of people's stories rather than entities. Te Ara proved difficult to automatically link.

At NTCIR-9 (the evaluation forum focusing on East-Asian cross-lingual information retrieval) we ran the CrossLink track[4] to examine the performance of CLLD algorithms from English source documents to Chinese, Japanese, and Korean target documents. The best performing approaches there were from HITS (Fahrni, Nastase, & Strube, 2011) and UKP (Kim & Gurevych, 2011). Their algorithms consistently performed well across the different language pairs and under different evaluation methods (Tang, Geva, et al., 2011).

[3] http://www.teara.govt.nz

[4] http://ntcir.nii.ac.jp/CrossLink/

3. The Framework

3.1 The Evaluation Methodology

The effectiveness and the efficiency of various CLLD approaches can only be measured in rigorously designed experiments. It this paper we are interested only in the effectiveness because (currently) we consider the task to be off-line and therefore efficiency is less important.

The evaluation methodology presented is illustrated in Figure 4. Similar to the Cranfield methodology (Voorhees, 2007), there are four parts: *Inputs*, *Systems*, *Outputs*, and *Evaluation*.

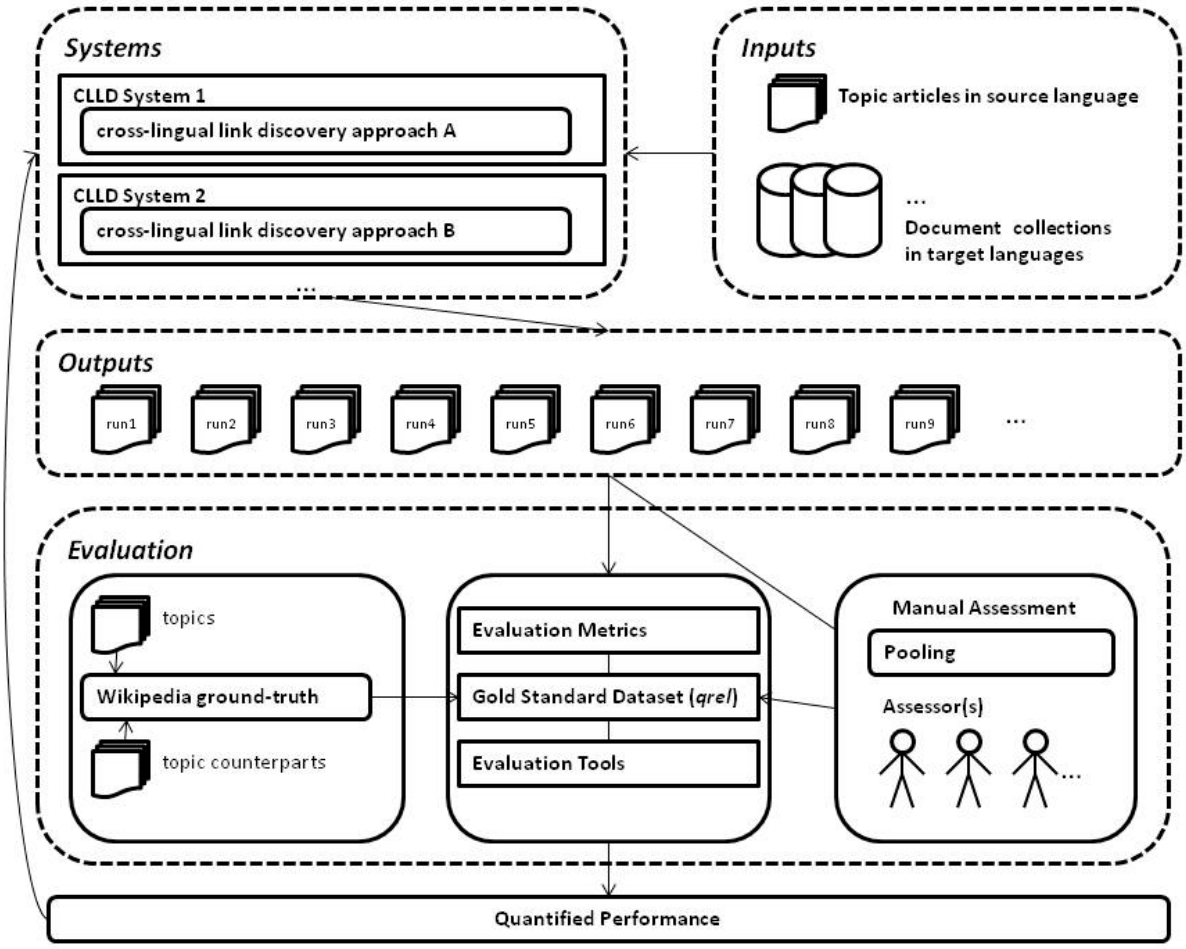

Figure 4: The Cross-lingual Link Discovery evaluation methodology

Inputs is the test collection used by the different systems. The test collection includes *topics* and *target document sets*. To create the topics we took a small set of the Wikipedia articles in their source language (English) and removed all the exiting links from those documents (a process known as orphaning). It is these articles that are processed by the CLLD systems and for which links are recommended. To create the target document set we took Wikipedia articles in Chinese, Japanese, and Korean. The CLLD system task was to identify relevant targets for this set of topics, and to rank these targets as well. It is well un-

derstood that the target document sets contain prior link knowledge (pre-existing links) that can be used for machine learning.

Systems is a set of CLLD systems (i.e., algorithms) to be assessed by the framework. In the case of NTCIR-9 CrossLink, these systems were produced by different research groups who implemented different algorithms of their own design.

Outputs is the output of the CLLD systems. Each output is called a run, each run contains the recommendations for each topic, and for each topic the run consists of a set of ranked anchors and links for that topic.

Evaluation includes the assessment method and evaluation metrics. In our framework there are two kinds of assessments: automatic assessments and manual assessments. Under the automatic assessment the links already present in the Wikipedia (before orphaning) are considered to be the ground-truth gold standard. Under the manual assessment all links in all runs are pooled and assessed by a human judge in the manner of the Cranfield methodology. Our framework includes several metrics to measure the performance of runs using these assessment sets, one such metric is LMAP, a variant of mean average precision for link discovery.

3.2 Toolkits in the Framework

With an evaluation framework in place it becomes possible to measure the performance of CLLD system in a reproducible manner, and consequently to systematically refine and improve the performance of such systems. But the complete evaluation framework includes not only the collections (topics, documents, and assessments), evaluation metrics, but also tools for submission pooling, link assessment, and system evaluation.

We also include tools for run validation. Validation is a key part of the evaluation framework because it ensures those who are developing algorithms that they are able to correctly generate runs for performance comparison purposes. Only valid runs can be pooled and assessed, and typically in Information Retrieval the greater the variety of runs the more reusable the test collection.

The tool set developed for the CLLD framework discussed in this paper (including validation tool, assessment tool, and evaluation tool) is a multi-lingual adaptation of the tools previously used in the INEX Link-the-Wiki track (W. Huang et al., 2010). These open source[5] tools were refactored and enhanced to support cross-language link discovery at NTCIR-9.

4. The Cross-lingual Link Discovery Task

Evaluating cross-lingual link discovery methods is awkward because of the number of degrees of freedom in judging the relevance of suggested anchors and targets. Different methods of assessing links result in different evaluation results. This section describes how the link structure already present in the Wikipedia can contribute to the assessment, that evaluation is more effective if manual assessment is included, and metrics for measuring the performance of CLLD systems.

4.1 Links in the Wikipedia

A link is a navigation entity within a single document that consists of two parts, an anchor text and a target. The target can either be within the same document or any other document. Links are provided for

[5] http://code.google.com/p/crosslink/

easy access to related information and are powerful elements for information navigation. Links are followed by a simple click.

An anchor is a snippet of text associated with the link and that is relevant to the topic of the target article. Anchors should be linked to related articles so that the reader can gather further information (or receive explanations). Wikipedia anchors are often manually added and can currently target only one destination article. There are four types of links in the 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 article is in the same language as the source article and 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 article are rare.

HTML supports linking independent of language (indeed, it does not know anything about the language 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 manually identify link targets in a second or subsequent language. Second, none of the dominant web browsers directly support multiple links per anchor (but they can be programmed to do so). Each anchor is typically linked to a single target in the source language and there is no native way to add alternative (language) targets to the same anchor. It is reasonable to assume that monolingual linking is preferred on the assumption that the reader would prefer to stay in a single language. Adding multiple language targets to a single anchor requires human support to find them and browse support to display them – neither of which is available in quantity.

Looking beyond these limitations, it is clear that multiple targets per link are 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 4 million English Wikipedia articles, there are only 7,040 in Māori. It is unreasonable (even unethical) to restrict access to knowledge simply on a lingual basis. Anchored cross-lingual links with multiple targets (one-to-many linking) is an essential addition to the Wikipedia.

4.2 Link Assessment

For automatic assessment of CLLD the ground-truth assessment set is derived from links already present in Wikipedia articles through triangulation. These come from two sources: First, all the mono-lingual links from the target language version 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 the 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 counterparts of each link out of the English article.

This ground truth may be incomplete. For example, it may contain links from the English version to which there is no an appropriate anchor in the target language version, it may not contain the kinds of links that users click, or there may be no version in the target language. 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 to be biased towards links already in the Wikipedia. Huang *et al* (W. C. Huang et al., 2009) suggest that manual assessment of monolingual link discovery could result in substantially different results. This is likely to be true in the multi-lingual case and is tested for the framework herein.

An alternative approach to automatic assessment is manual assessment following the Cranfield paradigm used at TREC and other Information Retrieval evaluation forums. For manual assessment human assessors are employed to examine each recommended link and to judge the relevance (or otherwise) of the link. A human assessor can judge more fine-grained than simply document topic relevance, for example they can individually assess the relevance of the anchor and of the target – a target might be relevant to the anchor, but the anchor not relevant from the user perspective (or *vice versa*).

In addition to the usual criticisms of evaluation using manual assessment seen in traditional information retrieval tasks, there are new issues with manual assessment of cross-lingual link discovery. For example, it is difficult to find assessors skilled enough to read the source articles in one language while also having a good understanding of the target documents in another language. Note that the skill level (in both languages) required to do this is higher than that of an ordinary user reading both documents because the assessor must thoroughly comprehend both.

4.3 Task Definition

Cross-lingual link discovery consists of two phases (in no particular order): 1) detecting prospective anchors in the source article; and 2) identifying relevant articles in the target language. The performance of both must be measured.

The Wikipedia is a constantly evolving collection. To alleviate this problem a snapshot in a small number of languages (English, Chinese, Japanese, and Korean) was taken. The task was to identify, from within this snapshot, a set of anchors for a source article and for each anchor a set of target documents. For each source article the links can be symbolized as:

$$\boldsymbol{a_i} \rightarrow \left(\boldsymbol{d_1}, \boldsymbol{d_2}, \cdots \boldsymbol{d_j}\right); \boldsymbol{i} < M; j < N \quad (1)$$

where a_i is the i^{th} anchor in the source document; d_j is target document j for the anchor; M is the number of anchors that are identified; and N is the number of target links identified for the anchor. For the purpose of manual assessment M and N should be set at values that make assessment feasible. For the purpose of automatic assessment this is not necessary.

4.4 Link Evaluation

4.4.1 Evaluation Types

Evaluation of mono-lingual link discovery has been studied in the INEX Link Discovery Track (D. Huang et al., 2008; W. Huang et al., 2009, 2010; Trotman et al., 2011), however the INEX evaluation metrics changed from year to year. The INEX experiments identified two different paradigms of assessment: file-to-file and anchor-to-file and these are adopted and incorporated in the new setting of the cross-lingual link discover task.

File-to-file ("See Also") links are present in many Wikipedia pages where the titles of related articles are linked to the associated article in a section of their own. In file-to-file evaluation the performance of the

link discovery algorithm at finding related articles is measured with disregard to any possible anchor identified by the CLLD algorithm. For example, if a CLLD algorithm were to identify the anchor "bowl" in an article on "Custard" and to link it to an article on "Bavarian Cream", and the assessor were to identify the target article as relevant but the anchor as not relevant, then under file-to-file evaluation the link would be considered relevant. File-to-file evaluation is ideally suited to automatic CLLD assessment because appropriate anchors cannot necessarily be extracted from the corpus, whereas appropriate target articles can.

Anchor-to-file evaluation also considers the anchor text. There are many possible anchors that might pertain to a given relevant target article, and it is important for a CLLD system to identify an appropriate one. Under anchor-to-file evaluation both the anchor and target must be relevant for a link to be considered relevant. In the example above, a link from "Custard" to the relevant "Bavarian Cream" article placed on the non-relevant "bowl" anchor would be a non-relevant link. Manual assessment is necessary for anchor-to-file evaluation, but the assessments can subsequently be used for file-to-file evaluation.

In this setting, relevant anchor to irrelevant target document assessment and the INEX anchor-to-BEP (best entry point) evaluation are not considered.

4.4.2 Link Precision and Recall

Precision and recall are the two underlying key metrics used to measure system performance in Information Retrieval (R. A. Baeza-Yates & Ribeiro-Neto, 1999). For CLLD the traditional definitions are extended to account for both anchors and targets, and to account for multiple targets per anchor. In this section two metrics are derived, one for file-to-file evaluation and the other for anchor-to-file evaluation.

4.4.2.1 File-to-File Evaluation

Precision at some point in the results list is defined in the usual way:

$$\boldsymbol{Precision_{f2f}} = \frac{\boldsymbol{found\ and\ relevant}}{\boldsymbol{found}} \tag{2}$$

as is recall:

$$\boldsymbol{Recall_{f2f}} = \frac{\boldsymbol{found\ and\ relevant}}{\boldsymbol{relevant}} \tag{3}$$

When file-to-file evaluation is used for "see also" links no anchor is specified and so the results list is analogous to a search engine results list. In this case precision and recall can be computed at each point in the results list and used in system metrics such as Link Mean Average Precision (LMAP, see Section 4.4.3). In the case of file-to-file evaluation of an anchor-to-file run, precision and recall are computed at each anchor and then plugged into a system metric.

4.4.2.2 Anchor-to-File Evaluation

Anchor-to-file evaluation also takes anchors into consideration and so precision and recall are further extended. Considering initially the anchor (similarly to INEX 2009 (W. Huang et al., 2010)):

$$f_{anchor} = \begin{cases} \mathbf{1}, \text{if } \boldsymbol{anchor} \text{ is relevant} \\ \mathbf{0}, \quad \text{otherwise} \end{cases} \tag{4}$$

That is, if an assessor has identified at least one relevant target for an anchor and the anchor is in the run then the anchor component of the score is 1. The alternate case is that no relevant targets are known for the anchor and so it receives a score of 0.

Likewise, if a given target for a given anchor is known to be relevant then it scores a target component score of 1, otherwise it scores 0:

$$f_{target} = \begin{cases} \mathbf{1}, \text{if } \boldsymbol{target} \text{ is relevant} \\ \mathbf{0}, \quad \text{otherwise} \end{cases} \tag{5}$$

Given the score for the anchor and the score for each target of that anchor it is possible to define precision for a single-anchor multiple-target link. For a specific anchor *anchor*, its precision is defined as:

$$Precision_{link}(anchor) = f_{anchor} \times \frac{\sum_{j=1}^{k} f_{target_j}}{k} \tag{6}$$

That is; the precision of the link for *anchor*, $Precision_{link}(anchor)$, is 0 if the anchor score, f_{anchor}, is 0; otherwise it is the set-wise precision of the k targets, j=1...k, specified for that anchor (which might also be 0). Likewise for recall:

$$Recall_{link}(anchor) = f_{anchor} \times \frac{\sum_{j=1}^{N} f_{target_j}}{N} \tag{7}$$

where N is the number of known relevant targets for the anchor. The precision for at some point, after n anchors, in the results lists is given by:

$$Precision_{a2f} = \frac{\sum_{i=1}^{n} Precision_{link}(i)}{n} \tag{8}$$

And likewise for recall when there are M known relevant anchors

$$Recall_{a2f} = \frac{\sum_{i=1}^{n} Recall_{link}(i)}{M} \tag{9}$$

4.4.2.3 Example

Consider an article, A_{topic}, which must be linked, and consider a results list, L_{topic}, of links from anchors, a_i, to articles, d_j, in the target collection:

Consider an article A_{topic} which contains 12 anchors denoted as a_1, …, a_{12}, assuming that the anchors have been linked to 32 target documents denoted as $d_{1, \ldots,} d_{32}$, the results list of the links, L_{topic}, is given below:

L_{topic} = {
$a_1 \rightarrow (d_{131})$,
$a_2 \rightarrow (d_{13}, d_{234}, d_{350})$,
$a_3 \rightarrow (d_{323})$,
$a_4 \rightarrow (d_{123}, d_{315})$,
$a_5 \rightarrow (d_1, d_{33}, d_{235})$,
$a_6 \rightarrow (d_{13}, d_{23}, d_{35})$,
$a_7 \rightarrow (d_{12}, d_{24}, d_{36}, d_{231}, d_{389})$,
$a_8 \rightarrow (d_3)$,
$a_9 \rightarrow (d_{19}, d_{99}, d_{101}, d_{203}, d_{450})$,
$a_{10} \rightarrow (d_{13}, d_{23})$,
$a_{11} \rightarrow (d_4, d_{39}, d_{375}, d_{399})$,
$a_{12} \rightarrow (d_{88}, d_{293})$
}

After manual assessment the assessment set, S_{topic}, (the ideal vector) contains 7 anchors pointing to 12 target articles as shown below:

S_{topic} = {
$a_2 \rightarrow (d_{131}, d_{234})$,
$a_3 \rightarrow (d_{314})$,
$a_5 \rightarrow (d_1, d_{33}, d_{352})$,
$a_8 \rightarrow (d_3)$,
$a_{10} \rightarrow (d_{13}, d_{23})$,
$a_{11} \rightarrow (d_{41}, d_{389})$,
$a_{12} \rightarrow (d_{88})$
}

In L_{topic} there are a total of 7 relevant anchors comprising 11 relevant links (some anchors point to more than one relevant target), so under file-to-file evaluation the precision (equation (2)) and recall (equation (3)) are:

$$Precision_{f2f} = \frac{11}{32} = 0.344$$

$$Recall_{f2f} = \frac{11}{12} = 0.917$$

Under anchor-to-file evaluation, the precision (equation (8)) and recall (equation (9)) are:

$$Precision_{a2f} = \left(0 \times \frac{1}{1} + 1 \times \frac{1}{3} + 1 \times \frac{0}{1} + 0 \times \frac{0}{2} + 1 \times \frac{2}{3} + 0 \times \frac{2}{3} + 0 \times \frac{1}{5} + 1 \times \frac{1}{1} + 0 \times \frac{0}{5} + 1 \times \frac{2}{2} + 1 \times \frac{0}{4} + 1 \times \frac{1}{2}\right) / 12$$

$$= (0.333 + 0.667 + 1 + 1 + 0.5)/12 = 0.292$$

$$Recall_{a2f} = \left(0 \times \frac{1}{1} + 1 \times \frac{1}{3} + 1 \times \frac{0}{1} + 0 \times \frac{0}{2} + 1 \times \frac{2}{3} + 0 \times \frac{2}{3} + 0 \times \frac{1}{5} + 1 \times \frac{1}{1} + 0 \times \frac{0}{5} + 1 \times \frac{2}{2} + 1 \times \frac{0}{4} + 1 \times \frac{1}{2}\right) / 7$$

$$= (0.333 + 0.667 + 1 + 1 + 0.5)/7 = 0.50$$

4.4.3 System Evaluation Metrics

In this section traditional Information Retrieval system-metrics are extended to incorporate the CLLD definitions of precision and recall. In particular, *Precision-at-N*, *R-Prec*, and *Link Mean Average Precision* (LMAP) are defined for cross-lingual link discovery.

R-Prec is defined as:

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

where *n* is the number of topics (articles used in evaluation); and $p_t@R$ is the precision at *R* where *R* is the number of known relevant items for topic *t*. Similarly, ***Precision-at-N*** is computed as the mean over *n* topics of the precision at some cutoff, *N*, in the results list.

Link Mean Average Precision (LMAP) is defined as:

$$LMAP = \frac{\sum_{t=1}^{n} \frac{\sum_{k=1}^{m} p_{kt}}{m}}{n} \quad (11)$$

where *n* is the number of topics; *m* is the number of identified items in the results list (articles for file-to-file evaluation or anchors for article-to-file evaluation); and p_{kt} is the precision of the top *K* items for topic *t*.

LMAP is analogous to mean average precision (MAP) used for measuring the performance in document retrieval, but uses a graded score for the precision of an individual result in the results list (a one-anchor multi-target set). That score is the precision at the top *K* items according to either equation (2) or equation (8). For the case of file-to-file evaluation of an anchor-to-file results list, each target is considered to take one slot place in the results list, for example,

$a_1 \rightarrow (d_{13}, d_{234}, d_{350})$

is equivalent to:

$a_1 \rightarrow d_{13,}$

$a_1 \rightarrow d_{234,}$

$a_1 \rightarrow d_{350}$

When examining the precision at increasing points down the results list, or plotting recall against precision, the precision can increase or decrease giving a saw-tooth shape (Manning, Raghavan, & Schütze, 2008). To smooth the curve for a results list from a search engine it is not unusual to use interpolated precision. The same technique can be used in CLLD.

The interpolated link precision, *Interp-P* with *r* recall points is defined as:

$$\boldsymbol{Interp - P = \frac{\sum_{t=1}^{R} \max(p_t@r)}{R}, 0 \leq r \leq 1} \quad (12)$$

where *R* is the number of predefined equally spaced recall points and *r* is one of those recall points, $p_r@r$ is the precision at recall point *r* for topic *t*. The *max*() guarantees that the interpolated precision is the highest precision found in a given recall range.

This section presented a formal definition of CLLD and suggested two different tasks: file-to-file linking and anchor-to-file linking. It then defined precision and recall for each task and these lead to a variant of MAP called LMAP, of R-Prec, and finally to interpolated precision. Each of these three metrics can be used to evaluate the performance of either file-to-file or anchor-to-file systems when the appropriate version of precision is used.

5. The Framework Tool Set

5.1 Validation Tool

For anchor-to-file links both the anchor and the target must be specified. Specification of a target document for CLLD is not different from that used in Information Retrieval ranking experiments such as those conducted at TREC. Similarly, a document ID can be used as target. The specification of the anchor, however, is different.

It is not possible to simply specify the text of a link because the given anchor text may occur in the source document (the topic) multiple times; and so this would be ambiguous. The offset and length in the topic file could be used; however this leads to a counting problem. Counting in bytes is impractical because many documents are encoded using more than one byte per character (e.g. UTF-8), and therefore byte-offsets are sub-character. Specifying in characters is awkward if the file is encoded in a mark-up language (e.g. XML) because positions within the mark-up itself could be specified.

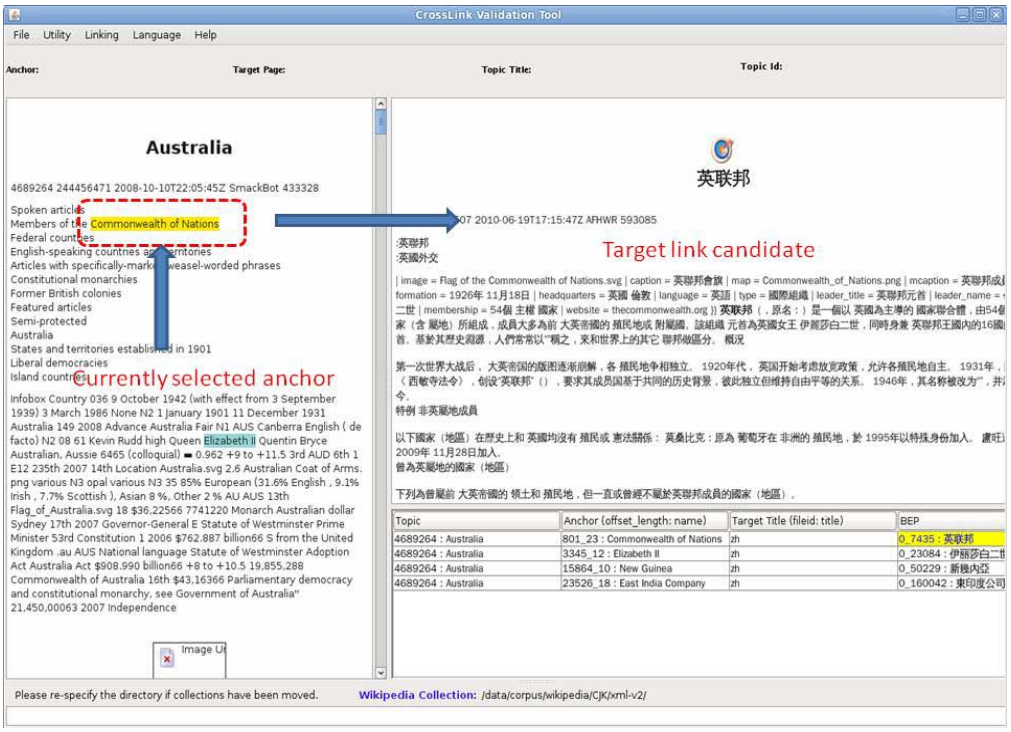

Figure 5: Crosslink run validation tool showing English-to-Chinese links

For the framework character offsets after mark-up parsing are used to specify the start position of an anchor and the length of the anchor text. This is non-trivial to compute, but the correct specification in a run is essential for robust evaluation. Consequently, the framework includes a run validation tool.

A screenshot of the tool is given in Figure 5. In the left pane the anchors specified in a run are shown in place in the source document (in this case the Wikipedia article on Australia). By clicking on a link the list of targets is displayed in the bottom-right pane. By clicking in that pane the target document is loaded into the top-right pane (in this case the 英联邦 Wikipedia article). The tool allows experimenters to not only validate their anchors, but to also examine the targets specified in their runs. Any miss-positioned anchors are obvious because they are displayed as such and many incorrectly specified targets are also immediately obvious.

The validation is not the primary topic of this contribution therefore a thorough investigation of its use is left for further work. However, of the 57 runs submitted to NTCIR-9, 11 contained a large number of invalid links –it is assumed that the producers of those runs had not used the validation tool.

5.2 Assessment Tool

An assessment tool was developed so that assessors could assess links graphically. The assessor inspected each anchor and its corresponding target articles, accepting or rejecting the relevance one by one. This is not dissimilar to the assessment in mono-lingual or cross-lingual Information Retrieval: given an anchor and its context, the assessor judges the relevance of the target document.

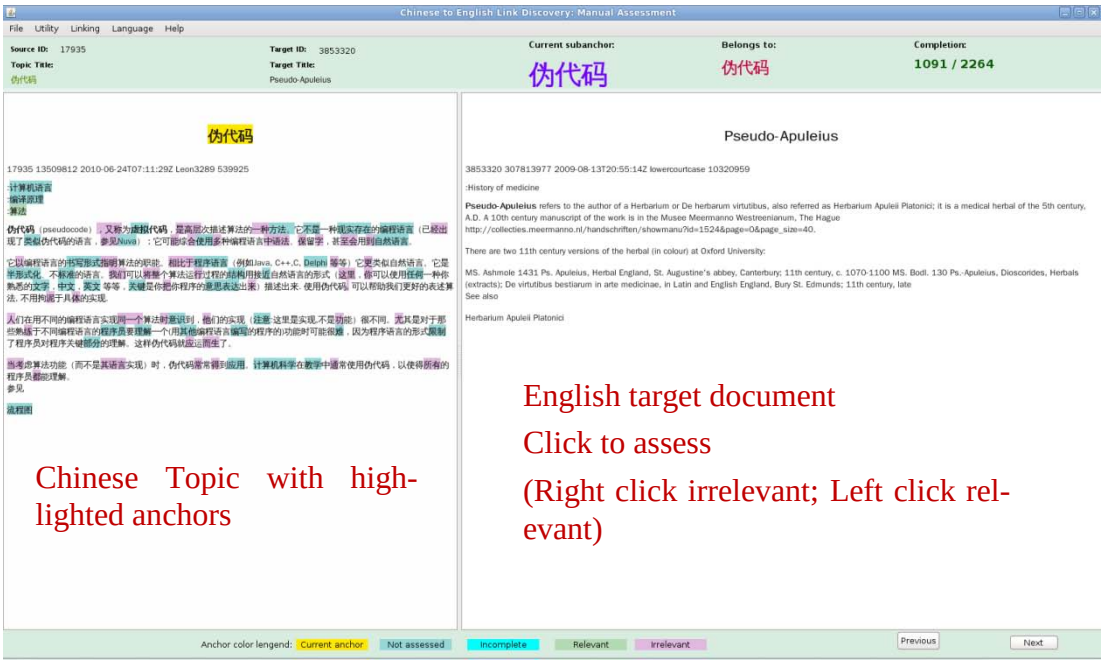

Figure 6: Crosslink manual assessment tool showing Chinese-to-English links

The assessment tool is similar to the validation tool. Figure 6 shows a Chinese to English assessment in progress. In the left pane the source document (the topic) is shown with assessed relevant anchors and assessed non-relevant anchors shown in different colours (green and red respectively (when onscreen or colour print)); unassessed in a third colour (cyan) and the current anchor being assessed in a fourth colour (yellow) . In the right pane the target document for the current anchor is shown. The assessor assesses by either clicking the right mouse button (for non-relevant) or the left mouse button (for relevant). There is no functional difference between the assessment tool for any pair of languages and the tool has been used for English, Chinese, Japanese, and Korean.

It is premature to thoroughly evaluate the assessment tool because (in its current form) it has not been used sufficiently to draw conclusions. Prior work at INEX has examined each of the progressive changes that lead to the current tool, that work is discussed in the Link the Wiki track overview papers (Geva, Kamps, Trotman, & Huang, 2009; W. Huang et al., 2009). However, section 6.5 cursorily presents rough assessment times.

5.3 Evaluation Tool

Given a valid run and a set of assessments a researcher is faced with the question of quantizing the performance (precision) of their runs. Section 4 discussed some of the metrics that might be used and the reasons for different kinds of assessment. This section discusses a tool for visualizing performance.

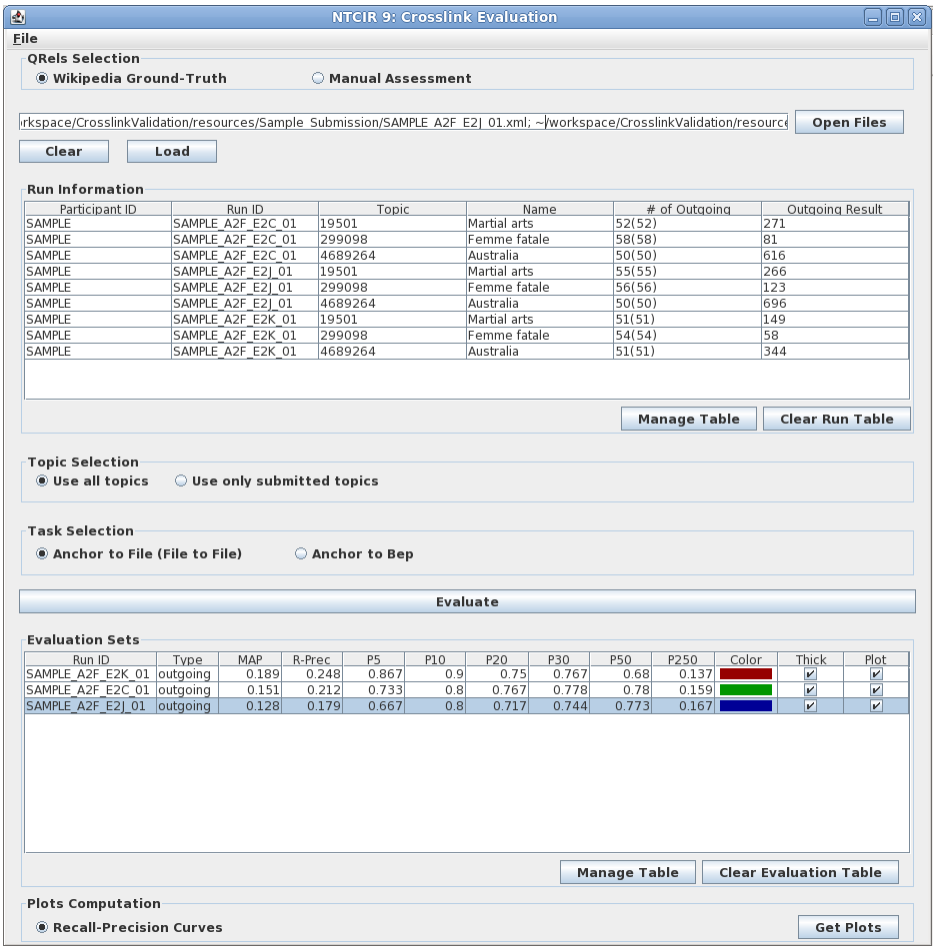

Figure 7: The evaluation tool comparing three runs

Figure 7 shows a screenshot of the evaluation tool. At the top the user can specify which assessment set (manual or automatic) they prefer. Below that they can specify the name of a run-file to load. Below that a summary of the run is provided by the system. The user can then specify which topics to measure (or all of them), and the kind of assessment (only file-to-file and anchor-to-file are supported for CLLD, recall that the tool is derived from the INEX tool that allows the assessment of anchor-to-BEP runs). The bottom table lists a set of runs to compare along with the LMAP, R-Prec and Precion-at-N for various *N*. The user can select runs and ask for a precision / recall graph of interpolated precision to be drawn; precision is interpolated at 20 points: 0.05, 0.10, 0.15, 0.20, .., 1.00.

6. Application of the Framework

The previous sections discussed the Cross Language Link Discovery problem and outline a framework and tools to help assess the quality of CLLD algorithms. This section discusses the implementation of the framework in the NTCIR-9 CrossLink task.

6.1 Document Collection

The Wikipedia is an excellent collection for CLLD because it exists in several languages and is a (mostly) closed hypertext with low levels of cross-lingual links. Wikipedia articles can be re-distributed under the Creative Commons Attribution-Share-Alike License 3.0 ("Creative Commons Attribution-Share-Alike License 3.0,"), and so copyright issues are minimal.

As the experiments were conducted at NTCIR-9, English, Chinese, Japanese, and Korean were chosen as the languages for experimentation, and the task was to link from English to the other languages (CJK). The June 2010 snapshot of the Wikipedia was chosen and downloaded then converted into XML using the YAWN system (Schenkel, Suchanek, & Kasneci, 2007). This system had already been used at INEX, but for NTCIR it was adapted to support new languages.

The document collection is summarised in Table 1. Column 1 lists the language, column 2 the number of articles in that collection. In column 3 the number of links pointing from the CJK articles to English and *vice versa*. For example, there are 316,251 Chinese articles of which 170,637 (54%) contained links to English articles. It is evident that there are already cross language links in this collection – the task is, therefore, to add more.

Table 1: Document collection statistics

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

6.2 Topics

A set of 25 articles were selected from the English Wikipedia and used as the test topics for evaluation based on the following three topic selection criteria: 1) they have to have corresponding counterparts in all three CJK languages; 2) they were of appropriate length that participants could generate sufficient anchors evaluation; and 3) they contained a good number of pre-existing links useful for the purpose of automatic evaluation.

The test topics had their pre-existing links removed – a process known at INEX as orphaning. Table 2 lists the topics. Column 1 and 4 are the topic number, columns 2 and 5 give the article titles and columns 3 and 6 give the document identifiers. For example, topic 8 is the English Wikipedia article "Source code" with doc ID 27661.

6.3 Run Specification

At NTCIR (and INEX) link recommendation is considered to be a recommendation task, consequently runs are ranked lists of links. Although there is no practical limit to the number of anchors that might be

inserted into an individual document, a user will become overwhelmed if every term in every article were also an anchor. The question of where to draw the line is a research direction left for future work. For practical (manual assessment) purposes a limit of 250 anchors per article was imposed and only 5 targets per language per anchor were allowed. This sums to at most 1,250 outgoing links per article per target language.

Table 2: List of topics used at NTCIR-9

#	Title	ID	#	Title	ID
1	Zhu Xi	244468	14	Cretaceous	5615
2	Kimchi	178952	15	Croissant	164372
3	Kim Dae-jung	152967	16	Crown prince	236210
4	Fiat money	22156522	17	Cuirassier	504031
5	Boot	191574	18	Dew point	54912
6	Asian Games	39205	19	Oracle bone script	480299
7	Spam (food)	28367	20	Ivory	15165
8	Source code	27661	21	Jade	191395
9	Puzzle	86368	22	Kiwifruit	17363
10	Pasta	23871	23	African Wild Ass	3696045
11	Abdominal pain	593703	24	Mohism	21032
12	Cuttlefish	20976520	25	Sushi	28271
13	Credit risk	372603			

6.4 Link Pooling

In total 57 runs from 11 teams were submitted. Table 3 presents statistics of those runs. Column 1 lists the task, column 2 lists the number of runs submitted, and column 3 gives the average number of links found per topic (averaged over all runs). For example, in the English-to-Chinese task there were 25 runs averaging 2,969 unique links per topic. The remaining tasks were English-to-Japanese, and English-to-Korean.

Table 3: NTCIR-9 CrossLink run statistics

Task	Runs	Average links per topic
English-to-Chinese	25	2969
English-to-Japanese	11	666
English-to-Korean	21	924

All the links from all runs were pooled and de-duplicated. Despite the availability of the validation tool, some anchors in some runs did not pass validation checks – those links were discarded from the pool. It was possible for some runs to specify overlapping anchors. For example, had one run specified “world

war" it did not (could not) preclude another from specifying "world" and a third from specifying "war". As anchors were judged one by one this did not cause pooling or assessment problems.

The pool was assessed to completion by human assessors. After assessment all valid links in all runs had been assessed, there were no valid unassessed links in any runs.

6.5 Human Assessors

Recruiting enough volunteers with different language backgrounds is challenging. Conveniently, there are many students with differing backgrounds at Queensland University of Technology (QUT) many possessing (at least) bi-lingual skills. These students make good assessors for two reasons: 1) they are well educated, and can reasonably be expected to understand Wikipedia articles; and 2) they possess good language skills in their native language and English, and can be reasonably expected to be able to read and comprehend articles in these languages. Assessors were compensated for their time with movie tickets.

Table 4: Assessor information

Task	Assessors	Description
English-to-Chinese	15	PhD students, Undergrad students
English-to-Japanese	1	Postdoc
English-to-Korean	5	Undergrad students

Table 4 lists summary information on the assessors. Column 1 lists the task, column 2 the number of assessors, and column 3 their positions. For example, the 1 assessor for the English to Japanese task was a Postdoc researcher at QUT.

Finding assessors for the English-to-Chinese task was initially considered to be straightforward because of the abundance of Chinese students at QUT. However, assessing the English-to-Chinese links was challenging for several reasons. First, this task saw a large number of unique links and so considerable time was required for assessment (about 3 hours per topic). Second, Chinese students at QUT were conscientious studiers who considered themselves too busy to help. Third, and related, compensation was insufficient to overcome this hurdle. Due to the shortage of assessors whom could be found locally i.e. in QUT, three of the topics were assessed by two participating teams of the English-to-Chinese subtask.

Finding assessors for the Japanese-to-English task was also difficult, but this time due to the lack of availability of English-Japanese bilingual speakers. We initially identified three volunteers; however two were eventually unable to participate for personal reasons leaving only one assess (an author of this paper).

Finding assessors for the Korean-to-English task was not problematic. Bilingual English / Korean speakers were readily available in Brisbane in the form of undergraduate students.

All assessors were given instruction and assessed topics under the supervision of at least one author of this contribution.

For the Japanese and Korean tasks it took about 1 hour for one assessor to assess one topic to completion. Each topic was assessed by a single assessor (no cross-assessor agreement experiments were conducted due to scarcity of assessors, and so this is left for future work). Some assessors assessed more than one topic.

6.6 Links Found in Manual Assessment

A summary of the resulting assessments is presented in Table 5. Column 1 lists the assessment task. Column 2 lists the number of relevant links found and the percentage of those links seen in intersection of the automatic and manual sets. Column 3 lists the absolute number of links seen in the intersection of the manual and automatic sets. For example, there were 1,681 links seen in the automatic set (found through triangulation of the 25 topics) in the English-to-Korean Wikipedia, 2,786 relevant links in the manual set (assessed by a human), 821 links were seen in the intersection which is 49% of 1,681; a similar pattern can be seen for the English-to-Chinese topics. In the English-to-Japanese assessments the number of relevant links 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 5: Relevant links in two assessment sets

Assessment Set	Relevant links (% of overlapping)	Overlapping
English-to-Chinese automatic	2,116 (54%)	1,134
English-to-Chinese manual	4,309 (26%)	
English-to-Japanese automatic	2,939 (27%)	781
English-to-Japanese manual	1,118 (70%)	
English-to-Korean automatic	1,681 (49%)	821
English-to-Korean manual	2,786 (29%)	

6.7 The Validity of Automatic Assessment

It is reasonable to ask whether manual assessment or automatic assessment results in the most robust evaluation outcome. From Table 5, it can be seen that a substantial number of the Wikipedia-extracted automatic assessments overlap with those manually assessed from the pool. It is, therefore, reasonable to expect that the two different forms of assessment will correlate.

To validate this hypothesis the automatically extracted assessments were manually assessed. One assessor was chosen for each language and the assessments were assessed to completion for topical relevance. Recall that no anchors exist in these assessments and so the assessment was for file-to-file ("see also") relevance.

Figure 8 shows the interpolated precision recall curve for each of the different language combinations. There is a substantial difference in quality between the three different versions of the Wikipedia. The Chinese assessor considered most links obtained through triangulation to be relevant, however both the Japanese and Korean assessors did not. One possible reason for this discrepancy might be the assessment technique itself. Links were presented without anchors (and therefore without context) making assessment difficult. Further investigation is required.

Nonetheless, this result suggests that automatic assessment alone may not be sufficient – a result also seen at INEX in English link discovery. We leave for further work the determination of the error in automatic assessment, but we anticipate it will be small because Information Retrieval evaluation is

concerned with the relative rank order of systems rather than the exact performance score (which changes from topic set to topic set).

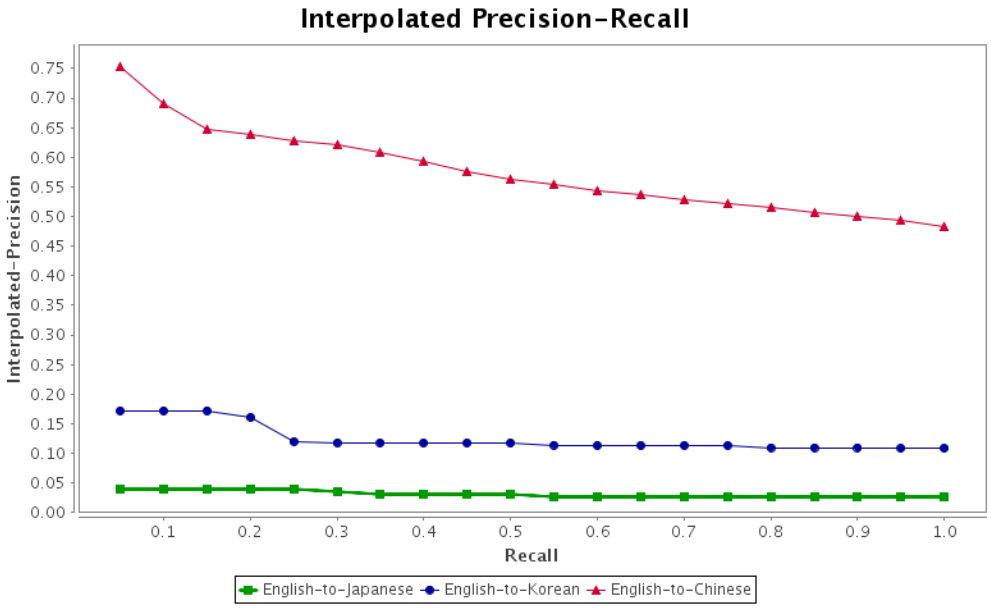

Figure 8: Interpolated precision-recall graph showing file-to-file evaluation of the automatic assessments once manually assessed. Links were taken in order of appearance in their respective articles.

7. Effectiveness of Cross-lingual Link Discovery

The framework presented herein was used to assess runs submitted to the NTCIR-9 Crosslink task. As discussed in Section 6.4 the English-to-Chinese task received the most runs, the largest number of unique links per topic, and the largest number of relevant links. As discussed in Section 6.7, it also appears to have the most robust assessment sets. The effectiveness of cross-lingual link discovery is, consequently, discussed in this section only with respect to that task. For a full account of the English-to-Japanese and English-to-Korean result see the overview of the NTCIR-9 CrossLink track (Tang, Geva, et al., 2011).

File-to-file evaluation was performed using both the automatic and manual assessment sets. Anchor-to-file evaluation can only be, and was, only performed using the manual assessments. LMAP, the MAP like metric discussed in Section 4.4.3, was chosen as the preferred metric because MAP is well understood by the Information Retrieval community; however results using the other metrics presented in Section 4.4.3 are given for completeness.

7.1 Results

The performance of the best (by LMAP) runs submitted by each of the top 7 participating groups is given in Table 6. The left of the table shows the performance against the automatic assessments and the right against the manual assessments, both are sorted in decreasing score order. The first column gives the name of the research group and the second column gives the precision score. For example, HITS (Fahrni et al., 2011) scored an LMAP score of 0.373 and placed first under automatic assessment but only scored an LMAP score of 0.102 under manual assessment placing third (behind UKP (Kim &

Gurevych, 2011) and QUT (Tang, Cavanagh, et al., 2011)). Similar tables are shown for R-Prec (Table 7) and Precision-at-5 (Table 8). The team with the highest early precision under manual assessment was KMI (Knoth, Zilka, & Zdrahal, 2011). Precision / recall graphs for all English-to-Chinese runs are given in Figure 9 and Figure 10 from where it can be seen that scores range wildly between different runs and the different evaluation methods.

It is difficult to identify one group as overall best because performance order varies from metric to metric and from automatic to manual assessment; however some groups consistently performed well including HITS who consistently outperformed other groups at identifying links like those already in the Wikipedia (i.e. automatic assessment).

It is reasonable to conclude that manual assessment is a more rigorous evaluation method because the LMAP scores are consistently lower. Automatic evaluation, on the other hand, is inexpensive and quick because it does not require manual assessment. Automatic assessment can also be scaled to the entire collection whereas manual assessment is constrained to the number of topics that can reasonably be assessed in the available time.

Table 6: LMAP of the top teams in two English-to-Chinese evaluations

Automatic F2F evaluation		**Manual A2F evaluation**	
Group	**LMAP**	**Group**	**LMAP**
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

Table 7: R-Prec of the top teams in two English-to-Chinese evaluations

Automatic F2F evaluation		**Manual A2F evaluation**	
Group	**R-Prec**	**Group**	**R-Prec**
HITS	0.471	UKP	0.171
UKP	0.417	QUT	0.133
KMI	0.345	KMI	0.114
IASL	0.347	HITS	0.105
QUT	0.244	IASL	0.036
WUST	0.207	WUST	0.010
ISTIC	0.101	ISTIC	0.000

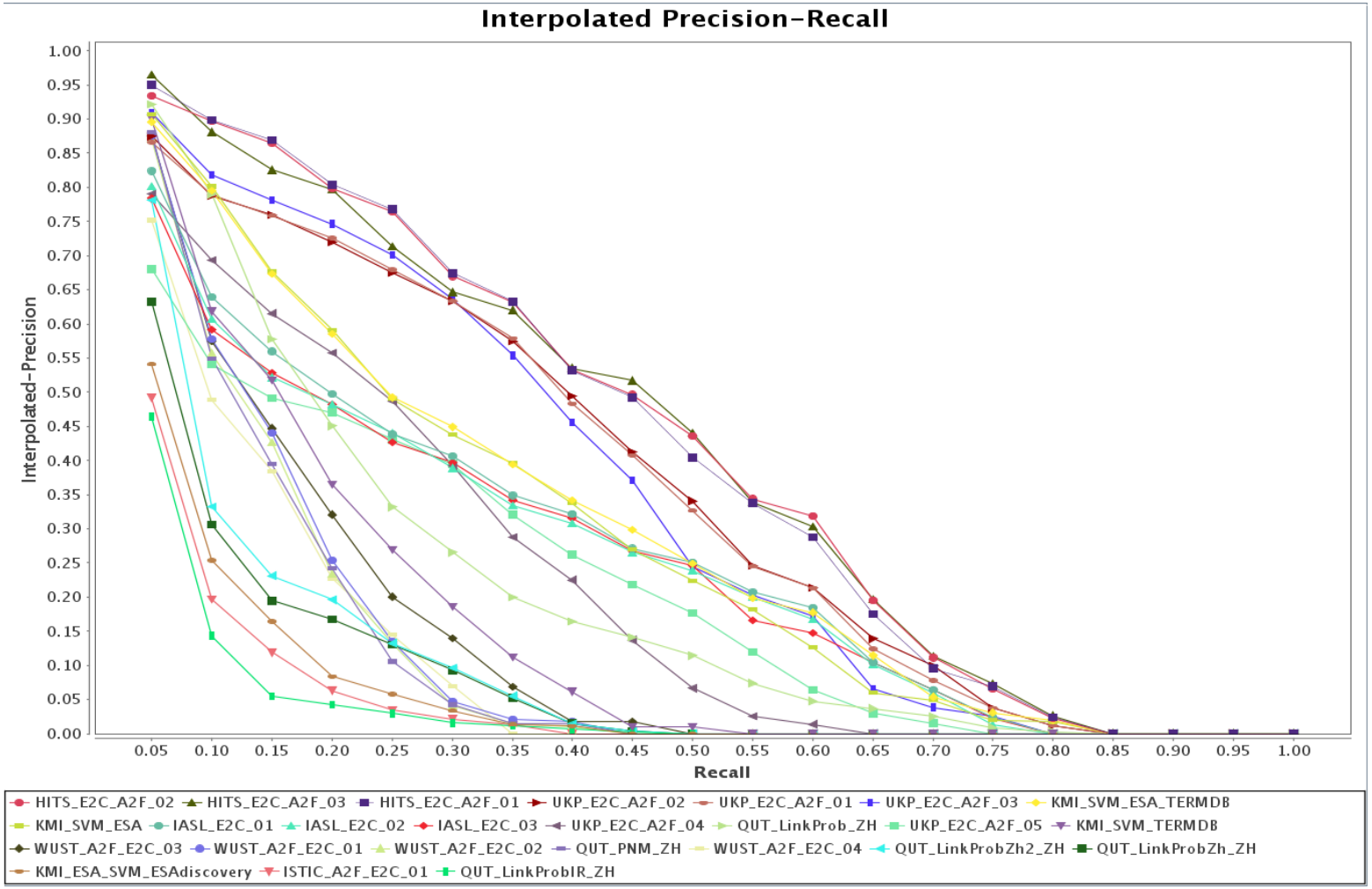

Figure 9: Interpolated precision-recall graph showing English-to-Chinese file-to-file evaluation against the automatic assessments

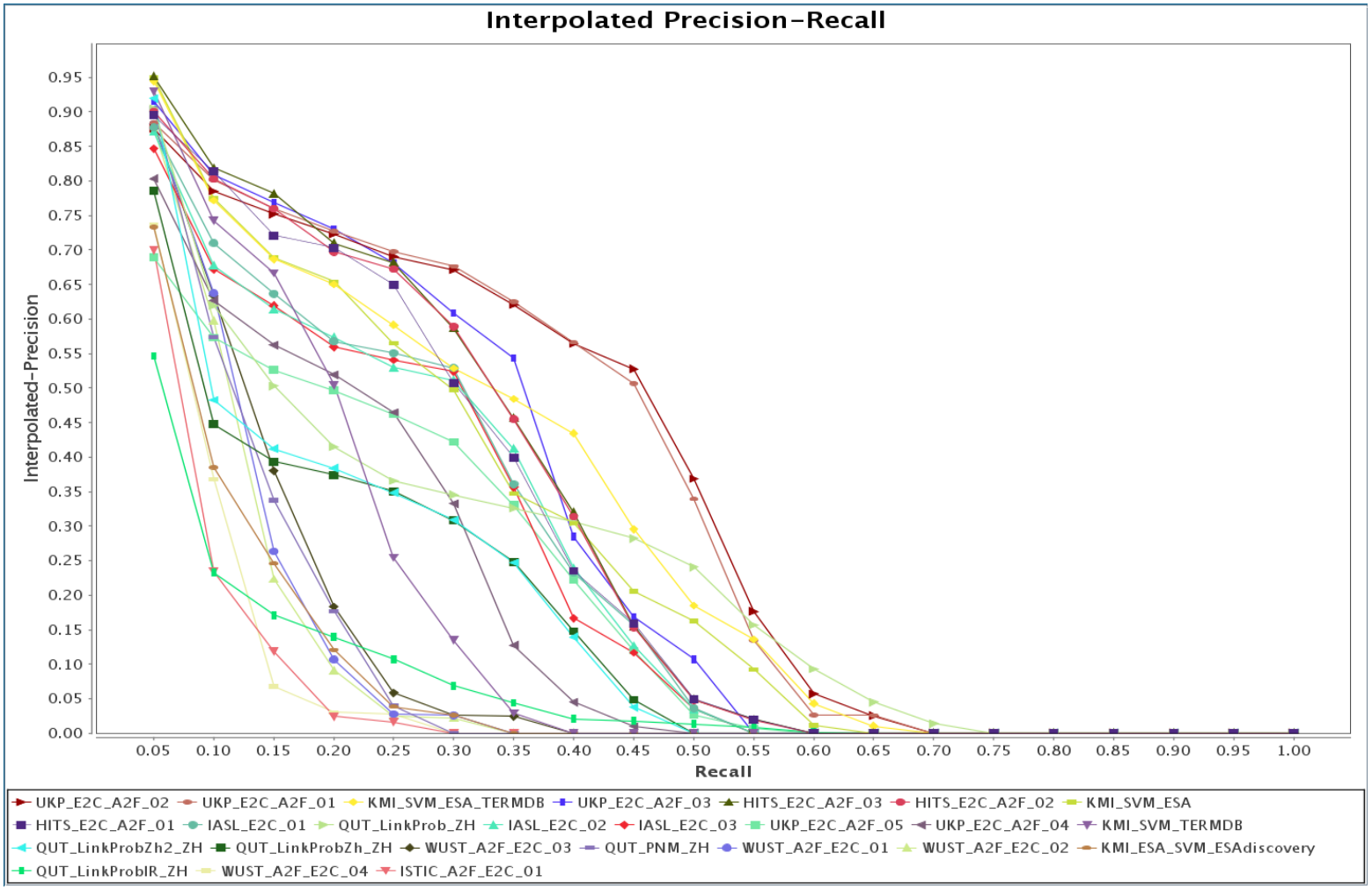

Figure 10: Interpolated precision-recall graph showing English-to-Chinese anchor-to-file evaluation against the manual assessments

Table 8: Precision-at-5 of the top teams in two English-to-Chinese evaluations

Automatic F2F evaluation		Manual A2F evaluation	
Group	**P@5**	**Group**	**P@5**
HITS	0.832	KMI	0.376
QUT	0.776	QUT	0.336
KMI	0.728	HITS	0.240
UKP	0.688	UKP	0.240
WUST	0.576	IASL	0.096
IASL	0.544	WUST	0.056
ISTIC	0.168	ISTIC	0.000

7.2 Statistical Analysis

To substantiate any claim that one group performed better than any other we applied the paired two-tailed *t*-test on the best runs from each group and for each of the three metrics: LMAP, R-PREC, and P@5. As we performed 6 tests per metric with Bonferroni correction, the threshold *p*-value was set to 0.00833 (5% by 6 tests). Table 9 shows, for each metric, the significance test results for the English-to-Chinese task.

In file-to-file evaluation with the Wikipedia ground-truth (automatic assessment), measured with either LMAP or R-PREC, HITS's best run is significantly better than any other run; in the anchor-to-file evaluation with manual assessment, measured with either LMAP or R-PREC team UKP's system is significantly better than the others.

Overall there is no obvious difference in significance tests results using scores computed with either LMAP or R-PREC. Respectively, with LMAP there are 20 (95%) significant differences in the file-to-file automatic evaluation, and 18 (86%) significant differences in the anchor-to-file manual evaluation, out of the total 21 paired runs. For example, it is shown that in manual anchor-to-file evaluation, there are no significant differences of system performances only between teams ranked 2 to 4.

However, in significance tests for runs measured with R-PREC, all systems perform significantly better than those with lower rankings in all types of evaluation. It is difficult to justify this test result with the 100% difference rate when comparing it with the significant test results computed with LMAP scores. Although using the R-PREC metric for system performance evaluation produces similar results to LMAP, attaining significant improvements appears to be more difficult than doing so with LMAP.

The tests also show that when measured with P@5, no team performed better than any other!

7.3 Comparison of CLLD Algorithms

7.3.1 The Top Performing CLLD Algorithms

This section briefly outlines the approaches taken by the four top groups at NTCIR-9 CrossLink. The interested reader is referred to the works of each respective group for further information.

Table 9: Statistical significance (t-test, paired two-tailed, $p < (0.05 / 6 = 0.00833)$)

F2F Evaluation with Wikipedia Ground-Truth

(a-1) LMAP

	1	2	3	4	5	6	7
HITS		#	#	#	#	#	#
UKP			#	#	#	#	#
KMI				=	#	#	#
IASL					#	#	#
QUT						#	#
WUST							#
ISTIC							

(b-1) R-PREC

	1	2	3	4	5	6	7
HITS		#	#	#	#	#	#
UKP			#	#	#	#	#
KMI				#	#	#	#
IASL					#	#	#
QUT						#	#
WUST							#
ISTIC							

(c-1) P@5

	1	2	3	4	5	6	7
HITS		=	=	=	#	#	#
QUT			=	=	#	#	#
KMI				=	=	=	#
UKP					=	=	#
WUST						=	#
IASL							#
ISTIC							

A2F Evaluation with Manual Assessment

(a-2) LMAP

	1	2	3	4	5	6	7
UKP		#	#	#	#	#	#
QUT			=	=	#	#	#
HITS				=	#	#	#
KMI					#	#	#
IASL						#	#
WUST							#
ISTIC							

(b-2) R-PREC

	1	2	3	4	5	6	7
UKP		#	#	#	#	#	#
QUT			#	#	#	#	#
HITS				#	#	#	#
KMI					#	#	#
IASL						#	#
WUST							#
ISTIC							

(c-2) P@5

	1	2	3	4	5	6	7
KMI		=	=	=	#	#	#
QUT			=	=	#	#	#
HITS				=	#	#	#
UKP					#	#	#
IASL						=	#
WUST							=
ISTIC							

Note: the number row (1, 2, 3, 4, 5, 6, 7) of each significance test table is the rank of teams measured with a metric; we have 6 tests per metric for each evaluation type. '#' indicates statistical significance at 5% (with Bonferroni), '=' indicates no statistical significant difference.

HITS (Fahrni et al., 2011) implemented a CLLD system with strong disambiguation. Candidate anchors were recognised by matching all possible n-grams in the topic document with phrases in a pre-constructed lexicon. A threshold was set to filter out terms with low key phraseness. Next, a graph-based disambiguation process was performed to remove those anchor candidates that might have caused ambiguity. Finally, remaining anchors were sorted based on the expected relevance produced by a trained binary classifier. Uniquely, their lexicon was seeded with sources other than the document collection (Wikipedia dumps from different dates, different language corpora, etc).

UKP (Kim & Gurevych, 2011) followed the common practice of linking documents in four steps: 1) anchor selection; 2) anchor ranking; 3) anchor translation; 4) target discovery. Their best runs used a strategy combining word N-grams for anchor selection, anchor probability for anchor ranking, and Wikipedia page triangulation for translation.

QUT (Tang, Cavanagh, et al., 2011) used a link mining approach that mines the existing link structure for anchor probabilities, and relies on the cross-lingual article title triangulation for anchor translation. Their method contributed the largest number of unique relevant links to the pool (Tang, Geva, et al., 2011).

KMI (Knoth et al., 2011) grouped articles with similar concepts as the source document (the topics) and designate these the target documents. To do this they used various methods including explicit semantic analysis (ESA). They discarded links if no anchors could be found through triangulation to the English Wikipedia. Links were then ranked by a Support Vector Machine (SVM).

Table 10: Comparison of different implementations of the four CLLD systems

Team	Anchor Source	Anchor Ranking	Disambiguation	Translation	Key Features
HITS	Wiki-Titles	Prior probability	Maximum edge weighted clique	Triangulation	Multilingual concept repository Anchor disambiguation
UKP	Noun phrases, Named entities, Wiki-Anchors, Wiki-Titles, N-Grams	BM25, Anchor probability, Anchor strength	N/A	Triangulation, Dictionary, Machine translation	Several anchor selection methods, Several anchor text translation methods
QUT	Wiki-Links	Link Probability, Page Name Matching, CLIR	N/A	Triangulation, Machine translation	Mostly link prior probability
KMI	Wiki-Titles	Link Ranking	N/A	Similar concept clustering	Target documents first Ranking with SVM

A comparison of these four approaches is given in Table 10. The team is listed in column 1, the source of the anchors is listed in column 2, the anchor ranking method is given in column 3 with the disambiguation method listed in column 4, column 5 gives the method used to translate anchors. The key features of the algorithms are given in column 6. For example, QUT used the text of already-present Wikipedia links which they ranked on prior-probability as seen already in the collection. QUT did not disambiguate possible targets; for translation of the anchors they used triangulation.

Three of the four approaches discover links by finding candidate anchors in source articles, then finding target documents for them. The fourth, KMI, works in reverse, identifying target documents first, and then looking for anchors for them. Only HITS performed disambiguation.

Many of the algorithms used at NTCIR-9 CrossLink built on prior work in mono-lingual link discovery. There the link prior-probability has proven to be a successful indicator of link relevance (Itakura & Clarke, 2008; Mihalcea & Csomai, 2007; Milne & Witten, 2008). The most successful approaches at INEX were base on the simple approach:

$$\boldsymbol{p_{a \to d} = \frac{lf_{a \to d}}{df_a}} \tag{13}$$

Where $p_{a \to d}$ is the probability that an anchor text, a, previously seen in the collection pointing to document d will again appear as anchor text pointing to document d. $lf_{a \to d}$, the link frequency, is the number of documents containing the given anchor as anchor text pointing to d, and df_a is the number of documents containing anchor text a.

7.3.2 Cross-Language Agreement

Two groups, HITS and UKP, produced runs that consistently ranked higher than the others regardless of language subtask (Tang, Geva, et al., 2011). Both groups implemented language independent algorithms and then submitted runs using those algorithms trained on each of the three languages (Chinese, Japanese, and Korean). It is, consequently, possible to study the performance of the same algorithm across three different languages.

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 11. 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 12. From Figure 12 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 11) 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.

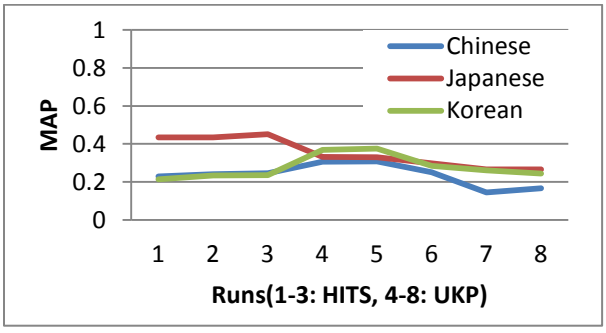

Figure 11: Performance of HITS (1-3) and UKP (4-8), manual file-to-file assessment

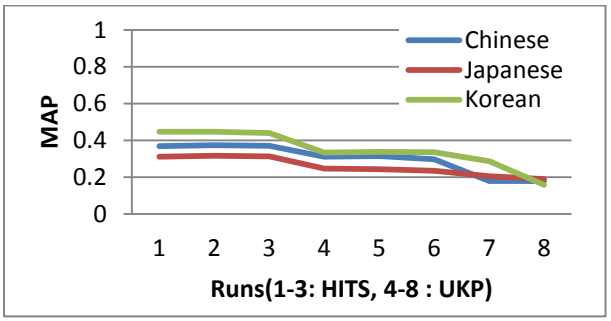

Figure 12: Performance of HITS (1-3) and UKP (4-8), automatic file-to-file assessment

A *t-test* between the Japanese runs scored automatically versus 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.

7.4 Unique Relevant Links

Algorithms that correctly suggest links similar to those already present in the collection are of interest, but so too are algorithms that identify novel relevant links not present (even at the expense of some precision). This is the precision / recall trade off well known in Information Retrieval, but it is of particular interest in CLLD because it allows the hyperlink graph to grow in new unconstrained ways. That is, algorithms that link mine (such as that of QUT) are important for consistency, but algorithms that suggest novel links (such as that of KMI) allow the graph to grow in new directions. Table 11 presents the statistics of unique relevant links discovered in the NTCIR-9 Crosslink runs. Column 1 lists the assessment set (automatic or manual), the column 2 lists the total number of unique links discovered across all runs, and column 3 lists the most diverse group and the number of unique and relevant links they identified, followed by the second and third most diverse group. For example, UKP found 97 of the 245 unique relevant links found in all the runs, which is about 11.6% of the relevant links in the automatically extracted assessment set.

Of particular note is that for the English-to-Chinese subtask QUT (Tang, Cavanagh, et al., 2011) contributes 1,103 unique relevant links to the manual assessment set, which is approximately 80% of total unique relevant links found. Comparatively, for the English-to-Japanese task, HITS runs (Fahrni et al., 2011) discovered 138 unique relevant links roughly 52% of total unique relevant links; Unique relevant links found in KSLP runs (Kang & Marigomen, 2011) consist of 47% (362 links) of the total. This suggests that these systems are the most diverse, preferring their own links to those already present in Wikipedia. Without manual assessment no knowledge of the relevance of these links would have been revealed.

Table 11 suggests that the assessments may not be exhaustive and we leave for further work determining the number of relevant links that might be present in the topic articles and that have not yet been identified. Caution must, consequently, be exercised when re-using the topics and assessments. Metrics such as *BPREF* (R. Baeza-Yates & Ribeiro-Neto, 2011) have been developed for document retrieval in the presence of incomplete assessments and we leave for further work the derivation of such metrics for CLLD. The assessments are likely to be sufficient for algorithms similar from those used by the partici-

pants of NTCIR-9 CrossLink, but we leave for further work the determination of the error of such an approach.

Table 11: Unique relevant English-to-Chinese links

English-to-Chinese Links

Compared with	Total (% of relevant links)	Team with highest (#)	2nd	3rd
Automatic	245 (11.6%)	UKP (97)	QUT (95)	KMI (27)
Manual	1397 (32.4%)	QUT (1103)	KMI (152)	UKP (88)

English-to-Japanese Links

Compared with	Total (% of relevant links)	Team with highest (#)	2nd	3rd
Automatic	603 (20.3%)	QUT (172)	KSLAB (159)	HITS (141)
Manual	266 (23.8%)	HITS (138)	UKP (122)	QUT (6)

English-to-Korean Links

Compared with	Total (% of relevant links)	Team with highest (#)	2nd	3rd
Automatic	129 (7.7%)	HITS (62)	KSLP (32)	UKP (24)
Manual	817 (29.3%)	KSLP (362)	KMI (305)	HIPS (105)

We also leave for further work the analysis of the nature of the new links identified by the algorithms. As the links are identified algorithmically there is undoubtedly patterns in those links – and for some reason that pattern is not seen in the Wikipedia. We also leave for further work the question of why there are not in the Wikipedia and any method that might be used to automatically add them to the Wikipedia. We note that Wikipedia policy is that any author of a bot changes Wikipedia is responsible for undoing any damage caused by such a bot.

8. Conclusions and Future Work

This paper presents an evaluation framework for benchmarking cross-lingual link discovery (CLLD) algorithms. It includes detailed discussions on the evaluation methodology and metrics used as well as some of the assessment challenges. The framework was applied to runs submitted to the NTCIR-9 crosslink track. The task of that track was to recommend anchors and target documents (links) from English Wikipedia articles into Chinese, Japanese, and Korean Wikipedia articles. The framework presented herein was instantiated as part of NTCIR-9.

Both automatic evaluation and manual evaluation were examined. For automatic evaluation the ground-truth assessments were extracted from direct links already present in the articles and from indirect links identified through triangulation. For manual evaluation assessment was performed by bi-lingual volunteers with a high level of education. Evaluation with these assessments was performed on two kinds of links, file-to-file ("see also") links and anchor-to-file ("in-text") links.

A substantial number of topically relevant links not already in Wikipedia were identified in submitted runs. This suggests that automatic assessment may be an inadequate method of measuring the performance of CLLD systems and that manual assessment may be necessary.

The top four CLLD approaches submitted to NTCIR-9 CrossLink were discussed. They differ in methods of identifying target documents, of disambiguating text, and of ranking prospective links. Evaluation showed that those algorithms were more effective at finding links already in Wikipedia than previously

unseen links, and no single algorithm was best at both. Some algorithms produced a disproportionally large number of unique relevant links suggesting that those groups focused on diversification in their result sets.

The additional anchors and links identified in the runs submitted to the NTCIR-9 CrossLink task may be useful additions to the Wikipedia. However the Wikipedia does not currently support multiple targets per anchor and so changes to the code-base would be necessary before addition of the links. If that were to be done then in future work it would be prudent to assess the quality of the links in situ before mass change. The topical relevance of the new links has been determined by the manual assessors. The similarity to already present links has been determined through automatic assessment. But the usefulness of the links remains unknown – that is, we don't know whether or not a user would click on the links and whether they would find the target documents beneficial to their work task. We have, throughout, outlined many areas for further work, and we also leave for further work the analysis of user behaviour and log files.

The Wikipedia does not appear to store a list of preferred languages for each user. Doing so would make it possible to automatically filter a set of cross-language links to the most appropriate set for the user. Indeed user-specific link sets were seen in early hypermedia systems where a domain expert might want to see a different link set (e.g. from Wikipedia to PubMed) to a novice (e.g from Wikipedia to a dictionary). We leave for further work an investigation into user-specific links in the Wikipedia and form the Wikipedia to sources external to the Wikipedia.

One of our initial motivations was that of expanding the Wikipedia's coverage in languages that were underrepresented in topical coverage. Linking from English to Chinese, Japanese, and Korean was appropriate in the context of NTCIR-9 CrossLink, but involves linking from an overrepresented language to others. Consequently at NTCIR-10 CrossLink-2 we are currently studying linking from Chinese, Japanese and Korean to English. Together CrossLink and CrossLink-2 provides a real opportunity to enhance Wikipedia cross language linking and we look forward to reporting on that work in the future.

9. Acknowledgements

We thank Wei Song, Yan Shen, Bin Liu, Tony Wang, the Korean Student Association of QUT (particularly Yong Jae Lee, Tina Son, and Hye Jung Yang) and others who helped assess the most submissions for the NTCIR-9 Crosslink task. We also thank Yi-Li Hsun and Maofu Liu for helping us finish manual assessment of three remaining topics for the English-to-Chinese subtask

References

Adafre, S. F., & De Rijke, M. (2005). *Discovering missing links in Wikipedia.* Paper presented at the Proceedings of the 3rd international workshop on Link discovery, Chicago, Illinois.

Agosti, M., & Crestani, F. (1993). *A methodology for the automatic construction of a hypertext for information retrieval*. Paper presented at the Proceedings of the 1993 ACM/SIGAPP symposium on Applied computing: states of the art and practice, Indianapolis, Indiana, USA.

Agosti, M., Crestani, F., & Melucci, M. (1996). Design and implementation of a tool for the automatic construction of hypertexts for information retrieval. *Information Processing & Management, 32*(4), 459-476. doi: http://dx.doi.org/10.1016/0306-4573(95)00075-5

Agosti, M., Crestani, F., & Melucci, M. (1997). On the use of information retrieval techniques for the automatic construction of hypertext. *Information Processing & Management, 33*(2), 133-144.

Allan, J. (1996). *Automatic hypertext link typing*. Paper presented at the Proceedings of the the seventh ACM conference on Hypertext, Bethesda, Maryland, USA.

Baeza-Yates, R., & Ribeiro-Neto, B. (2011). *Modern Information Retrieval: The Concepts and Technology Behind Search* (2 ed.): Addison Wesley.

Baeza-Yates, R. A., & Ribeiro-Neto, B. A. (1999). *Modern Information Retrieval*: ACM Press / Addison-Wesley.

Coombs, J. H. (1990). *Hypertext, full text, and automatic linking*. Paper presented at the Proceedings of the 13th annual international ACM SIGIR conference on Research and development in information retrieval, Brussels, Belgium.

Crane, G., Smith, D. A., & Wulfman, C. E. (2001). *Building a hypertextual digital library in the humanities: a case study on London*. Paper presented at the Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries, Roanoke, Virginia, USA.

Creative Commons Attribution-Share-Alike License 3.0. from http://creativecommons.org/licenses/by-sa/3.0/

Erbs, N., Zesch, T., & Gurevych, I. (2011, 18-21 Sept. 2011). *Link Discovery: A Comprehensive Analysis.* Paper presented at the The fifth IEEE International Conference on Semantic Computing (ICSC 2011)

Fahrni, A., Nastase, V., & Strube, M. (2011). *HITS' Graph-based System at the NTCIR-9 Cross-lingual Link Discovery Task.* Paper presented at the Proceedings of NTCIR-9, Tokyo, Japan.

Geva, S., Kamps, J., Trotman, A., & Huang, W. (2009). *Overview of the INEX 2009 Link the Wiki Track.* Paper presented at the Proceedings of INEX 2009.

He, J., & De Rijke, M. (2010). *An Exploration of Learning to Link with Wikipedia: Features, Methods and Training Collection.* Paper presented at the Proceedings of INEX 2009.

Huang, D., Xu, Y., Trotman, A., & Geva, S. (2008). *Overview of INEX 2007 Link the Wiki Track.* Paper presented at the Proceedings of INEX 2007.

Huang, W., Geva, S., & Trotman, A. (2009). Overview of the INEX 2008 Link the Wiki Track. In S. Geva, J. Kamps & A. Trotman (Eds.), *Proceedings of INEX 2008* (Vol. 5631, pp. 314-325): Springer Berlin / Heidelberg.

Huang, W., Geva, S., & Trotman, A. (2010). *Overview of the INEX 2009 Link the Wiki Track.* Paper presented at the Proceedings of INEX 2009.

Huang, W. C., Trotman, A., & Geva, S. (2009). *The importance of manual assessment in link discovery.* Paper presented at the Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, Boston, MA, USA.

Itakura, K., & Clarke, C. (2008). *University of Waterloo at INEX2007: Adhoc and Link-the-Wiki Tracks.* Paper presented at the Proceedings of INEX 2007.

Jihong, Z., & Bloniarz, P. A. (2004, 5-7 April 2004). *From keywords to links: an automatic approach.* Paper presented at the Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004. International Conference on.

Kang, I.-S., & Marigomen, R. (2011). *English-to-Korean Cross-linking of Wikipedia Articles at KSLP*. Paper presented at the Proceedings of NTCIR-9, Tokyo, Japan.

Kim, J., & Gurevych, I. (2011). *UKP at CrossLink: Anchor Text Translation for Cross-lingual Link Discovery.* Paper presented at the Proceedings of NTCIR-9, Tokyo, Japan.

Knoth, P., Zilka, L., & Zdrahal, Z. (2011). *KMI, The Open University at NTCIR-9 CrossLink.* Paper presented at the Proceedings of NTCIR-9, Tokyo, Japan.

Kurohashi, S., Nagao, M., Sato, S., & Murakami, M. (1992). *A method of automatic hypertext construction from an encyclopedic dictionary of a specific field*. Paper presented at the Proceedings of the third conference on Applied natural language processing, Trento, Italy.

Malcolm, J. W. (2001). Automatic creation of hyperlinks: Google Patents.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*: Cambridge University Press.

Melo, G. d., & Weikum, G. (2010). *Untangling the cross-lingual link structure of Wikipedia.* Paper presented at the Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden.

Mihalcea, R., & Csomai, A. (2007). *Wikify!: linking documents to encyclopedic knowledge*. Paper presented at the CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management.

Milne, D., & Witten, I. H. (2008). *Learning to link with wikipedia.* Paper presented at the CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management.

Salton, G., Allan, J., & Singhal, A. (1996). Automatic text decomposition and structuring. *Information Processing & Management, 32*(2), 127-138.

Schenkel, R., Suchanek, F., & Kasneci, G. (2007). *YAWN: A Semantically Annotated Wikipedia XML Corpus.* Paper presented at the 12. GI- Fachtagung für Datenbanksysteme in Business, Technologie und Web (BTW 2007), Aachen, Germany.

Smeaton, A. F., & Morrissey, P. J. (1995). Experiments on the automatic construction of hypertext from texts. *New Review of Hypermedia and Multimedia, 1*(1), 23-39.

Sorg, P., & Cimiano, P. (2008, 2008). *Enriching the Crosslingual Link Structure of Wikipedia - A Classification-Based Approach.* Paper presented at the AAAI 2008 Workshop on Wikipedia and Artifical Intelligence.

Sotomayor, B. R. (1998). Automatic summary page creation and hyperlink generation: Google Patents.

Tang, L.-X., Cavanagh, D., Trotman, A., Geva, S., Xu, Y., & Sitbon, L. (2011). *Automated Cross-lingual Link Discovery in Wikipedia.* Paper presented at the Proceedings of NTCIR-9, Tokyo, Japan.

Tang, L.-X., Geva, S., Trotman, A., Xu, Y., & Itakura, K. Y. (2011). *Overview of the NTCIR-9 Crosslink Task: Cross-lingual Link Discovery.* Paper presented at the Proceedings of NTCIR-9, Tokyo, Japan.

Tang, L.-X., Itakura, K. Y., Geva, S., Trotman, A., & Xu, Y. (2011). *The Effectiveness of Cross-lingual Link Discovery.* Paper presented at the Proceedings of The Fourth International Workshop on Evaluating Information Access (EVIA), Tokyo, Japan.

Trotman, A., Alexander, D., & Geva, S. (2011). *Overview of the INEX 2010 Link the Wiki Track* Paper presented at the Proceedings of INEX 2010.

Voorhees, E. M. (2007). TREC: Continuing information retrieval's tradition of experimentation. *Commun. ACM, 50*(11), 51-54. doi: 10.1145/1297797.1297822

Yang, H.-C., & Lee, C.-H. (2005). A text mining approach for automatic construction of hypertexts. *Expert Systems with Applications, 29*(4), 723-734. doi: http://dx.doi.org/10.1016/j.eswa.2005.05.003

Yeung, C.-m. A., Duh, K., & Nagata, M. (2011). Assisting cross-lingual editing in collaborative writing. *SIGWEB Newsl.*(Spring), 1-5. doi: 10.1145/1942800.1942804