Automated Cross-lingual Link Discovery in Wikipedia

Ling-Xiang Tang[1], Daniel Cavanagh[1], Andrew Trotman[2], Shlomo Geva[1], Yue Xu[1], Laurianne Sitbon[1]

[1]Faculty of Science and Technology,
Queensland University of Technology,
Brisbane, Australia

{ l4.tang, s.geva, yue.xu, laurianne.sitbon}@qut.edu.au, danielcavanagh85@gmail.com

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

andrew@cs.otago.ac.nz

ABSTRACT

At NTCIR-9, we participated in the cross-lingual link discovery (Crosslink) task. In this paper we describe our approaches to discovering Chinese, Japanese, and Korean (CJK) cross-lingual links for English documents in Wikipedia. Our experimental results show that a link mining approach that mines the existing link structure for anchor probabilities and relies on the "translation" using cross-lingual document name triangulation performs very well. The evaluation shows encouraging results for our system.

Categories and Subject Descriptors

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

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

General Terms

Algorithms, Experimentation.

Keywords

NTCIR, Crosslink, Wikipedia, Link Probability, Page Name Matching, Transliteration.

1. INTRODUCTION

At NTCIR-9, we participate in all subtasks of cross-lingual link discovery task. They are, namely, English-to-Chinese, English-to-Japanese, and English-to-Korean subtasks.

Among all language sub-sets of Wikipedia, English Wikipedia contains the largest number of articles. However, the links in the current English Wikipedia are mainly pointed at articles of the same language. Without direct links to articles in other languages, it may cause difficulties when viewing cross-lingual materials for people who are bi-lingual readers or knowledge contributors, or second language acquisition students (e.g. English learners of Chinese).

In this paper, we propose our approaches to alleviate this problem. In the current Web 2.0 era, access of cross-lingual information should be simple and easy.

The remainder of this paper is organized as follows: First, we discuss English to CJK cross-lingual document linking approaches in Section 2. The experimental runs and results are discussed in Section 3. We then conclude in Section 4.

2. CJK CROSS-LINGUAL LINKING

To locate CJK cross-lingual links for English Wikipedia articles, we separate the link discovery into two phases: 1) detecting prospective anchors in the source document; and 2) for each anchor, identifying relevant documents in the target language corpus. Once the anchor is identified, a link, *a→d*, is created (where *a* is the anchor, *d* is the target document).

Inspired by the monolingual link discovery approaches used at INEX, we are interested in testing the effectiveness of these methods in the cross-lingual link environment. The methods adopted for the Crosslink task experiments are: *link mining* method[1] and *page name matching* method[2]. With the same anchor identification strategy as that of link mining method, a cross-lingual information retrieval approach is also experimented. Furthermore, for English-to-Japanese document linking, we use a name entity identification method with English-to-Japanese transliteration to look up cross-lingual links on Wikipedia.

2.1 Finding Links with Link Mining

2.1.1 Mono-lingual Link Probability

Wikipedia contains a rich set of existing anchored links. These links contain pairs of specified anchor texts and the associated target documents. Hence, the existing link information can be used to recommend new links.

Itakura & Clarke[1] calculate the *anchor weight* (or link probability), γ, using:

$$\gamma = \frac{number\ of\ pages\ that\ have\ link(a \rightarrow d)}{number\ of\ pages\ that\ have\ text\ of\ anchor(a)}\ . \tag{1}$$

where the numerator is the link frequency, *lf*, of anchor *a* pointing to document *d*; and the denominator is the document frequency (*df*) of anchor *a* in the corpus. The computed γ score indicates the probability that a given phrase is an anchor and linked to a specific target document. So with this method, when an anchor is identified, the target document is also determined.

Mihalcea & Csomai[3] and Milne & Witten[4] also use a similar method to weight phrases. With computed link probabilities of anchor candidates, better links can be created for Wikipedia, or any documents in wild can be linked with Wikipedia.

Generally, to link a document of the same language: First, compute all possible n-gram substrings in the source document. Next, look-up its γ score for each n-gram text. Then, these anchor candidates are sorted on the γ score. Last, an arbitrary number (based on a threshold, or alternatively a density) of highly ranked links are then chosen. In the case of overlapping anchors, the longest anchor is chosen.

All γ scores of existing anchored links can be pre-calculated and stored in a *link table*, T_{link}. This table of mono-lingual anchor-to-target ($a \rightarrow d$) pairs can be created by mining the existing link structure of Wikipedia.

2.1.2 Cross-lingual Link Probability

To make the link mining method work with cross-lingual linking, a bridge needs to be built between English anchors and prospective Chinese documents. One way to use the link mining approach discussed previously is:

- First, mine in one language to create a list of candidate anchors;
- Second, “translate” those anchors into the second language;
- Then with the translations target documents can be identified using mono-lingual link recommendation methods.

To build such a language bridge, a table of documents existing in both Chinese and English could be used. Such a table, T_{lang}, can be generated from the page-to-page language links present in Wikipedia.

This is a form of cross-lingual document name triangulation in Wikipedia. A CJK page is a good target for an English anchor if there exists a link from the anchor to the English document and from the English document to the CJK document. The relationship of the triangulation is illustrated in Figure 1.

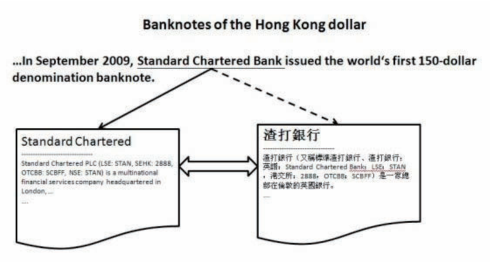
Banknotes of the Hong Kong dollar
...In September 2009, Standard Chartered Bank issued the world's first 150-dollar denomination banknote.
Standard Chartered
Standard Chartered PLC (LSE: STAN, SEHK: 2888, OTCBB: SCBFF, NSE: STAN) is a multinational financial services company headquartered in London, ...
....
渣打銀行
渣打銀行（又稱標準渣打銀行、渣打銀行；英語：Standard Chartered Bank；LSE：STAN，港交所：2888，OTCBB：SCBFF）是一家總部在倫敦的英國銀行。
...

Figure 1: Cross-lingual triangulation

2.1.3 Realisation of English-2-Chinese Linking

To implement the English-2-Chinese cross-lingual link discovery approaches two tables are needed. The first is the table of anchor-to-target pairs from link mining, $T_{link\text{-}english}$. The second is the table of corresponding document titles in both languages extracted from the language links T_{lang}.

To generate the table $T_{link\text{-}english}$, Link mining technique of Itakura & Clarke is utilised: first, trawl English Wikipedia and extract all anchor target pairs; then re-trawl the collection looking for the frequency of the anchor phrases used either as a link or in plain text.

Note that the same anchor text may be linked to different destinations in different instances where it appears and so it is necessary to identify the most likely link.

We only use this link mining method in English-2-Chinese subtask. Three different implementations of English-2-Chinese document linking are given below.

1) Linking with triangulation

First, build a mono-lingual English link table, $T_{link\text{-}english}$ by mining the English Wikipedia for all English anchor-target pairs. Since standard data set of Crosslink task doesn’t have an English Wikipedia corpus, so we obtain the English document collection from evaluation forum INEX 2009[5].

Remove from this table all rows for which there is no English document corresponding to the Chinese target. This information comes from T_{lang}. Several entries from T_{lang} are given in 1.

Table 1: Extracts from T_{lang}

English	Chinese
Citibank	花旗银行
Coconut	椰子
Scent of a Woman	女人香 (1992 年電影)
Michael Jordan	米高佐敦
Enya	恩雅

Then compute the link frequency and document frequency of all n-gram anchors in $T_{link\text{-}english}$. The document frequency of an n-gram anchor is the number of documents that contain that n-gram regardless of whether or not it is seen as an anchor.

Several entries from $T_{link\text{-}english}$ are given in Table 22. Each document has a unique id, a link frequency, and a document frequency.

Table 2: Extracts from $T_{link\text{-}english}$

Anchor	ID	*lf*	*df*
Citybank	231026	431	485
Enya	9482	287	420
Coconut Crab	810400	5	5
Audrey Tautou	342753	69	90
Wombat	33864	35	35

Finally, a candidate list of links is produced from this.

2) Linking with machine translation

In this implementation two link mining tables are generated; one for Chinese, $T_{link\text{-}chinese}$, and one for English, $T_{link\text{-}english}$. The anchors are then translated into English using Google Translate[1].

[1] http://code.google.com/apis/language/translate/overview.html

This method is similar to *Implementation 1)*, but the candidate target link of an anchor is not from $T_{lang,}$, but from $T_{link\text{-}chinese}$. Only those anchors that have translations in $T_{link\text{-}chinese}$ are used.

The link probability for the English anchors is taken from $T_{link\text{-}english.}$ That is, the Chinese link probabilities are used for calculating gamma scores for the English anchor texts.

3) Linking with machine translation (2)

Implementation 2 is used; however, anchors are sorted with the γ scores computed by using the English link probability table, $T_{link\text{-}english}$,

2.2 Finding Links with Page Name Matching

2.2.1 Page Name Matching

An alternative approach for cross-lingual link discovery is title matching (also known as name-matching, and entity matching). For mono-lingual link discovery Geva [2] builds a *page title table,* a list of titles of all documents in Wikipedia. For a new document he builds a list of all possible n-gram substrings and from that chooses the longest that are also in the page title table as the anchors. The targets are the documents with the given title.

To use this in English to other languages link discovery, it is necessary to first construct a table of corresponding English and target language documents. Then, for a new English document, identify all substrings that match document titles in other languages as the anchors. The targets are the corresponding documents of target language.

The name matching method is simple, but has proven to be effective. It is particularly useful if no pre-existing links exist in the document collection (a scenario in which link mining cannot be used).

If the link-graph contains no islands, each document contains only one incoming link, and all anchors are document titles; then the link mining *link table* covers all document titles and all weights are equal. That is, it is the *page title table* from title matching algorithm. So title matching is a special case of link mining.

2.2.2 Realisation of English-to-CJK Linking

Thank to the simplicity of this method, we can easily adapt it to discover Chinese, Japanese, and Korean documents for English documents. With this method only the cross-lingual page name mapping table, T_{lang}, is needed. The table contains a list of all document titles for which there are both English and target language articles (linked to each other). So Table T_{lang} can be easily extended by including the mapping of other languages. Table 3 shows the extracts of page title mapping for English to CJK Wikipedia documents.

The candidate list of links is produced by searching for n-grams in the source document that are also in the English column of T_{lang}, then linking to the corresponding English document.

Separately, we produced three runs using this method for English-to-Chinese, English-to-Japanese, and English-to-Korean subtasks respectively.

2.3 Finding Links with Cross-Lingual Information Retrieval

2.3.1 Cross-lingual information retrieval

The cross-lingual information retrieval approach to cross-lingual link discovery involves identifying anchors in one language, translating them into the target language, and then using them as search terms in a ranking search engine. The top ranked documents are chosen as targets to the anchors.

To identify the anchors, the same anchor detection strategies discussed previously might be used. This includes anchor mining and document titles. Alternatively a dictionary might be used.

2.3.2 Realisation of English-to-Chinese Linking

Anchors are identified using the link mining approach. The English anchors are translated into Chinese using Google Translate. Then an information retrieval system is used to identify candidate target documents from the Chinese Wikipedia.

We used a slightly modified BM25 ranking function for document ordering. In that function:

$$IDF(q_i) = log\frac{N}{n} \tag{2}$$

Where N is the number of documents in the corpus, and n is the document frequency of query term q_i. The retrieval status value of a document d with respect to query $q(q_1,...,q_m)$ is calculated as:

$$rsv(q,d) = \sum_{i=0}^{m} \frac{tf(q_i,d)*(k_1+1)}{tf(q_i,d)+k_1*\left(1-b+b*\frac{len(d)}{avgdl}\right)} * IDF(q_i) \tag{3}$$

Where $tf(q_i,d)$ is the term frequency of term q_i in document d; $len(d)$ is the length of document d and $avgdl$ is the mean document length. Parameters k_1 and b were 0.7 and 0.3 respectively (values previously shown to be effective).

Table 3: Extracts from T_{lang} after including title mapping of all CJK languages

English	Chinese	Japanese	Koran
Citibank	花旗银行	シティバンク、エヌ・エイ	한국씨티은행
Coconut	椰子	ココナッツ	코코넛
Scent of a Woman	女人香 (1992 年電影)	セント・オブ・ウーマン/夢の香り	여인의 향기
Michael Jordan	米高佐敦	マイケル・ジョーダン	마이클 조던
Enya	恩雅	エンヤ	엔야

2.4 Finding Links with Transliteration

2.4.1 Transliteration of Anchors

The use of transliteration of anchors was explored to determine its usefulness in translation-based link discovery techniques, specifically as a second attempt at finding links for anchors for which no translation is available (for instance, if translation services aren't currently available; or the anchor is not found in the chosen dictionary or is otherwise untranslatable; or the translation is not reasonable; etc.).

Because the focus of this run was the transliteration, only basic anchor discovery and link discovery were implemented around the transliterator to allow it to be used. This basic process is similar to the CLIR technique above, but with the following differences:

- The Stanford named entity recogniser was used instead of link mining to determine suitable anchors
- The search functionality of the live (online) Wikipedia was used instead of a local index of the provided 2009 Wikipedia corpus to determine suitable foreign-language documents
- If Google Translate failed to provide a suitable translation, the transliterator was used to calculate possible transliterations.

2.4.2 Realisation of English-to-Japanese Linking

The run was restricted to the English-to-Japanese task, because a transliterator was written only for Japanese. The transliteration itself is broken into 3 steps:

1) A normaliser that transforms irregular letter clusters into regular clusters based on their sound. This tries to keep in mind how the clusters would generally be transliterated into Japanese. For example, 'ough' can sound like 'or' and so these sounds ought to be normalised to the same thing
2) A simpler sound mapper that directly maps syllables or letter clusters into their possible equivalents in Japanese based on their sound, keeping in mind past transliterations that would generally no longer be applied to words entering the Japanese language ('archaic' transliterations).

   Each 'syllable' can have more than one valid transliteration, and so the transliterator keeps track of all individual syllable transliterations and outputs an enumeration of all possible full transliterations.

   For example, the ‘s’ in 'as' could theoretically be transliterated as either ス [su] or ズ [zu] (pronounced as a 'z' but written 's') and the mappings could thus be:

   [a] [su]

   [zu]

   , which would result in the set [“asu”, “azu”].
3) A post-normaliser that irons out any inconsistencies caused by the simpler sound mapper.

   This means that multiple possible transliterations are output from the system for one term and each one is tried in turn until either a link is discovered for one or there are none left to process.

Anchor discovery utilises the Stanford named entity recogniser[2], trained with the 4-class 'CoNLL' model that is provided with the SNER library, to determine named entities. Each named entity found is considered a suitable anchor.

Before doing link discovery, Google Translate is used to translate all anchors, and if any of them don't have translations then the transliterator is used to calculate possible transliterations. Links are then attempted to be discovered using the process below, and if nothing is found for any of the translations/transliterations then discovery is done on the original English anchor.

Link discovery utilises the search functionality of the online Japanese Wikipedia to identify candidates for linking. The top result from a search for a given anchor is considered the best candidate, and then a further 4 links are gotten by extracting the first 4 links from the best candidate itself, starting from the opening paragraph and minus anything within parentheses (as these were determined to generally be irrelevant things like pronunciation aids).

2.5 Comparison of CLLD Methods

A comparison of the above two approaches is presented Table 4. The cross-lingual information retrieval approach can find links never seen before. The link mining method produces more accurate results. The page name matching is particularly helpful when there is no pre-existing links existing, but the available links for recommendation are very limited.

Table 4: Pros and cons of the link discovery approaches

Method	Pros	Cons
ML	More accurate, less noisy	Only finds links already in the corpus
PNM	Simple, effective	Only finds links matched with the page title
CLIR	Finds links not seen elsewhere in the corpus	May be noisy
TRANSLITERATION	Simple	May not be very accurate

3. EXPERIMENT

3.1 Experimental Runs

From the different implementations for three cross-lingual link discovery methods discussed in Section 2, we generated 8 runs. The run names and system descriptions are outlined in Table 5. Run names for English-to-Chinese subtask are in *_ZH pattern; similarly, *_JA runs are Japanese runs; and QUT_PNM_KO is the run using page name matching algorithm for English-to-Korean task.

All English-to-Chinese runs except for *QUT_PNM_ZH* use the same anchor identification strategy. So, the difference in the performance of those runs *(*LinkProb*_ZH*) can be attributed to different anchor ranking, translation, and lookup (IR or link probability) methods.

[2] http://nlp.stanford.edu/software/tagger.shtml

3.2 Results and Discussion

The Crosslink task uses MAP, R-Prec, and P@N as the main evaluation metrics[6], so we employ the same measures to evaluate our runs' performance. The scores of eight different runs computed using the evaluation tool with official qrel are given in Table 6. Runs are sorted on MAP in two groups (file-to-file and anchor-to-file evaluations). Precision and recall curves are given in Figure 2.

3.2.1 Evaluation of Link Mining Runs

The ranking of the English-to-Chinese runs in two types of evaluations (F2F and A2F) is different. Even so, it can be seen from both Table 6 and Figure 2 that run *QUT_LinkProb_ZH* performed the best in both evaluations. It indicates this run has the best combination of strategies of anchor ranking, translation and link recommendation.

For other runs the Wikipedia ground-truth evaluation prefers cross-lingual page name matching method for automatic link discovery, but the evaluation with the manual assessment results finds the link mining methods using either Chinese or English source as link predictor can contribute more relevant links.

There is no obvious performance difference between run *QUT_LinkProbZh2_Zh* and *QUT_LinkProbZh_ZH* in both F2F and A2F evaluations. And also the different ranking of these two runs in both evaluations suggests that either English corpus or Chinese corpus can be used as a good source of link predictor.

However, the relatively low ranking of the above discussed two runs indicates the machine translation adopted for connecting the identified anchors and the cross-lingual target documents results in a worse performance of link finding if compared with that of the best run *QUT_LinkProb_ZH* which utilises the translation using Wikipedia cross-lingual page name triangulation.

3.2.2 Evaluation of Page Name Matching Runs

Given the limited number of page-to-page cross-lingual links existing in Wikipedia and resulting relatively small size of T_{lang} used by cross-lingual page name matching algorithm, the reasonable performance of all *PNM* runs (*PNM_ZH*, *PNM_JA* and *PNM_KO*) for all three language subtasks is surprising but encouraging.

3.2.3 Evaluation of CLIR Runs

The cross lingual information retrieval approach (run: *QUT_LinkProbIR_ZH*) has the lowest performance scores of all metrics in both evaluations. This is because the search engine is good at identifying relevant documents and not entities (document titles).

However, it is interesting to see that the *QUT_LinkProbIR_ZH* runs, even with the worst performance, contribute the highest number of unique relevant documents in English-to-Chinese subtask when evaluated with the *qrels* from manual assessment according to the official assessment results of NTCIR Crosslink task [6]. This result is encouraging but not surprising. As it is expected, cross-lingual information retrieval approach may not be able to accurately locate the exact match of target links with the suggested anchors, but it provides an opportunity for other also interesting and relevant links being seen by the information seekers.

Table 5. System information of QUT runs

Run ID	**Description**
English-to-Chinese	
QUT_PNM_ZH	This is a run using the PNM algorithm, and the cross-lingual title-to-target table is generated from the NTCIR 9- Crosslink: Chinese Wikipedia Corpus
QUT_LinkProbIR_ZH	Use the anchors recommended by link probability, and retrieve relevant links using a search engine with anchors as query terms
QUT_LinkProbZh2_ZH	Same as QUT_LinkProbZh_ZH , except for that anchors are sorted based on Chinese link probability table.
QUT_LinkProbZh_ZH	Use two set of link probability tables (one Chinese; one English mining from English Wikipedia corpus from INEX), and tables are connected by translation. Anchors are sorted based on English link probability table.
QUT_LinkProb_ZH	Use link probability for anchor sorting and link recommendation
English-to-Japanese	
QUT_PNM_JA	This is a run using the PNM algorithm, and the cross-lingual title-to-target table is generated from the NTCIR 9- Crosslink: Chinese Wikipedia Corpus
QUT_TRANSLITERATION_JA	The Stanford Named Entity Recogniser is used with the included 4-class CoNLL 2003 Shared Task model to identify named entities. After extracting these from the text, each named entity is translated using Google Translate, and if this fails then a potential list of transliterations is calculated (using a custom-written transliteration module). All terms are then passed to Wikipedia's Japanese search engine to identify suitable pages to link to. The top result for each is considered the best link, and this link is followed and a further 4 links gathered in ascending order. If for some reason 5 links
English-to-Korean	
QUT_PNM_KO	This is a run using the PNM algorithm, and the cross-lingual title-to-target table is generated from the NTCIR 9- Crosslink: Chinese Wikipedia Corpus

Table 6: Performance of experimental runs in both f2f and a2f evaluation

Run ID		MAP	R-Prec	P@5	P@10	P@20	P@30	P@50	P@250
metric scores computed with *qrel* from Wikipedia ground-truth									
f2f	LinkProb_ZH	0.179	0.244	0.776	0.588	0.480	0.404	0.319	0.132
	PNM_KO	0.122	0.208	0.552	0.460	0.384	0.321	0.244	0.062
	PNM_ZH	0.088	0.166	0.592	0.472	0.362	0.307	0.242	0.064
	PNM_JA	0.076	0.143	0.624	0.504	0.394	0.333	0.262	0.079
	LinkProbZh2_ZH	0.069	0.154	0.360	0.284	0.248	0.221	0.187	0.082
	LinkProbZh_ZH	0.059	0.148	0.304	0.208	0.168	0.161	0.156	0.082
	TRANSLITERATION_JA	0.047	0.145	0.160	0.136	0.126	0.139	0.152	0.099
	LinkProbIR_ZH	0.023	0.067	0.184	0.160	0.118	0.109	0.084	0.044
metric scores computed with *qrel* from manual assessment									
a2f	LinkProb_ZH	0.115	0.133	0.336	0.308	0.294	0.288	0.277	0.172
	LinkProbZh_ZH	0.094	0.119	0.320	0.244	0.260	0.273	0.269	0.158
	LinkProbZh2_ZH	0.090	0.117	0.312	0.312	0.304	0.299	0.271	0.155
	PNM_JA	0.087	0.016	0.128	0.124	0.108	0.096	0.077	0.020
	PNM_KO	0.043	0.043	0.136	0.200	0.220	0.217	0.193	0.047
	PNM_ZH	0.030	0.033	0.208	0.204	0.214	0.220	0.187	0.045
	LinkProbIR_ZH	0.008	0.026	0.104	0.104	0.072	0.073	0.070	0.033
	TRANSLITERATION_JA	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

3.2.4 Evaluation of Transliteration Run

The transliteration approach was not overly successful, given it had the second lowest performance of the runs, but a run focused on transliteration is never going to be a top performer when the majority of anchors don't need to pass through the transliterator. The problems inherent in the CLIR runs are generally applicable to this run, but as the effectiveness of transliteration was the focus of this run, these problems won't be discussed.

Determining the effectiveness of the transliteration was hampered by a couple of things:

- The named entity recogniser performed poorly (paucity of anchors, nonsensical anchors, etc.), so the transliterator didn't have a full range of anchors to be tested against. Leveraging the link mining method used in the other runs would have been more sensible
- The online search engine only returns results with the exact search term in them, not for anything with similar terms or for partial matches, which is particularly harsh for a transliteration process that outputs best guesses that may not quite match the correct transliteration

The worst problem is that there is no a true baseline run that the transliteration run can be compared against, so the effectiveness of the transliteration is not directly calculable. The similarity of the CLIR run to this run affords a grainy view, but there are enough differences between them for the effects of the transliteration to be swamped, and so there is no conclusion to the useful of transliteration. Still, the slightly better results of the transliteration run compared to the CLIR run is mildly encouraging, and further investigation with a proper baseline is easily doable.

3.2.5 Comparison with Other Teams

File-to-File Evaluation with Wikipedia Ground-Truth

Our runs didn't score well in the file-to-file evaluation with Wikipedia ground-truth. Run *QUT_LinkProb_ZH* is only ranked fourth when sorted on the Precision-at-5 metric in the English-to-Chinese task.

File-to-File Evaluation with Manual Assessment Results

In the file-to-file evaluation with manual assessment results, run *QUT_LinkProb_ZH* has the number on ranking when measured using Precision-at-5 metric in the English-to-Chinese task.

Anchor-to-File Evaluation with Manual Assessment Results

When the relevancy of anchors is taken into consideration, run *QUT_LinkProb_ZH* achieved the fourth position in ranking on all metrics (*MAP, R-Prec, Precision-at-N*) in the English-to-Chinese task. Our team is in second when ranked in team.

Overall, our runs, especially those submitted for English-to-Japanese and English-to-Korean tasks, have medium performance when compared to the other good runs submitted to the task. But we contribute largest number of unique relevant links that users might think deserve further reading.

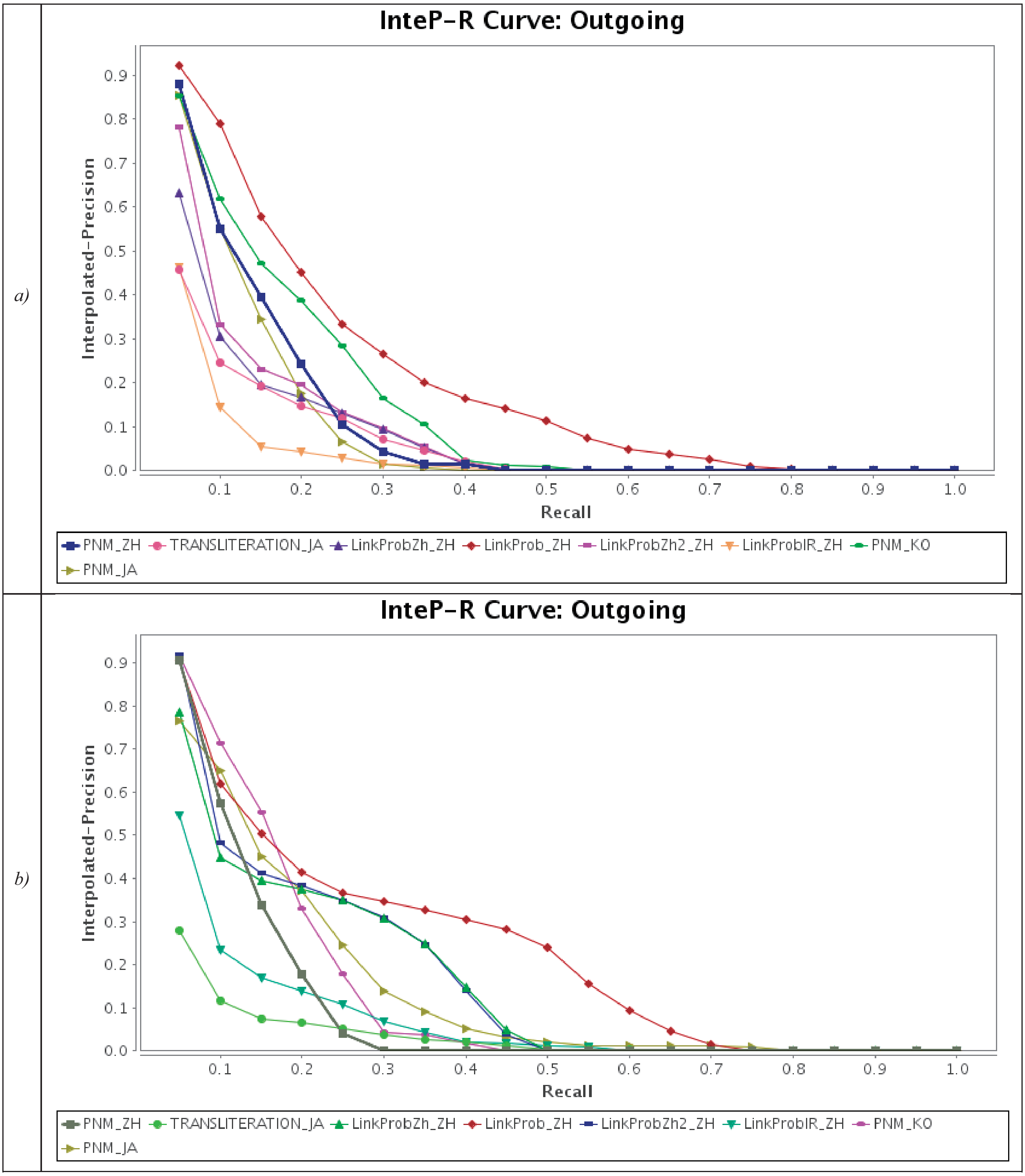

Figure 2. The interpolated precision/recall curves of runs. Plot *a)* is the f2f evaluation using Wikipedia ground-truth; plot *b)* is the a2f evaluation using manual assessment result.

4. CONCLUSION

In this paper we present our approaches to realising cross-lingual linking from English documents to CJK documents. Several automatic linking methods were tested. The methods employed include: link mining, page name matching, cross-lingual information retrieval and transliteration with online Wikipedia search service.

Link mining method with Wikipedia cross-lingual document name triangulation (run: *QUT_LinkProb_ZH*) performed the best among all implementations, and also achieved encouraging results in the overall evaluations of Crosslink task. This method requires pre-mining on the existing link structure of Wikipedia. In order to compute a list of English anchor / target probabilities, additional English Wikipedia corpus from INEX[5] was employed for this English link mining.

Since our submissions contribute the most unique links in the overall evaluation, in future the performance of our system could be further improved if links of different implementations are properly combined and re-ranked with a better anchor weighting strategy.

5. REFERENCES

[1] K. Itakura and C. Clarke, "University of Waterloo at INEX2007: Adhoc and Link-the-Wiki Tracks," in *Focused Access to XML Documents*, ed, 2008, pp. 417-425.

[2] S. Geva, "GPX: Ad-Hoc Queries and Automated Link Discovery in the Wikipedia," in *Focused Access to XML Documents*, ed, 2008, pp. 404-416.

[3] R. Mihalcea and A. Csomai, "Wikify!: linking documents to encyclopedic knowledge," in *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pp. 233-242.

[4] D. Milne and I. H. Witten, "Learning to link with wikipedia," presented at the Proceeding of the 17th ACM conference on Information and knowledge management, Napa Valley, California, USA, 2008.

[5] (2009, *INitiative for the Evaluation of XML-Retrieval (INEX)*. Available: http://www.inex.otago.ac.nz/

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