Simple Yet Effective Methods for Cross-Lingual Link Discovery (CLLD) – KMI @ NTCIR-10 CrossLink-2

Towards a better understanding of the link discovery task

Petr Knoth
KMI, The Open University
Walton Hall, Milton Keynes
United Kingdom
petr.knoth@open.ac.uk

Drahomira Herrmannova
KMI, The Open University
Walton Hall, Milton Keynes
United Kingdom
d.herrmannova@open.ac.uk

ABSTRACT

Cross-Lingual Link Discovery (CLLD) aims to automatically find links between documents written in different languages. In this paper, we first present a relatively simple yet effective methods for CLLD in Wiki collections, explaining the findings that motivated their design. Our methods (team KMI) achieved in the NTCIR-10 CrossLink-2 evaluation the best overall results in the English to Chinese, Japanese and Korean (E2CJK) task and were the top performers in the Chinese, Japanese, Korean to English task (CJK2E)[1] [Tang et al.,2013]. Though tested on these language combinations, the methods are language agnostic and can be easily applied to any other language combination with sufficient corpora and available pre-processing tools. In the second part of the paper, we provide an in-depth analysis of the nature of the task, the evaluation metrics and the impact of the system components on the overall CLLD performance. We believe a good understanding of these aspects is the key to improving CLLD systems in the future.

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, Languages

Keywords

Cross-lingual Link Discovery, Link Discovery, Semantic Similarity, Explicit Semantic Analysis, NTCIR, Wikipedia

1. INTRODUCTION

While document cross-referencing is an essential part of organising textual information on the Web, manual discovery and maintenance of appropriate links in large quickly growing collections is overwhelmingly time-consuming. In multilingual document collections, interlinking semantically related information in a timely manner becomes even more challenging. Therefore, suitable software tools that could facilitate the link discovery process by automatically analysing the multilingual content are needed. The NTCIR-10: CrossLink-2 task provides an evaluation forum for Cross-Lingual Link Discovery (CLLD) systems. In CrossLink-2, the performance of different CLLD methods is assessed on the Wikipedia corpus, which has some suitable properties for evaluating CLLD systems: a) It is a very large multilingual text collection, b) the articles are well-interlinked and the interlinking has been approved by a large community of users and c) a large proportion of articles contains explicit mapping between different language version.

Our paper is organised as follows. In Section 2, we present CLLD methods designed by team KMI that can be used to suggest a set of cross-lingual links from an English Wikipedia article to articles in Chinese, Japanese and Korean (English to CJK) or from an article in Chinese, Japanese and Korean to English (CJK to English). Though the CrossLink tasks focus only on these language combinations, our methods are generaly applicable to any language combination. We report the performance of the designed methods in Section 3. Section 4 is dedicated to the analysis of the CLLD task, the evaluation metrics and the impact of CLLD components on the overall performance. Finally, we discuss related work in Section 5 and summarise our findings.

2. LINK DISCOVERY METHODS

2.1 Preliminaries

In the following text, we will often use the terms *anchor, concept, term, link, sense, target, outlink and Wikipedia version* in the following way. By *term* we understand any textual fragment (typically a noun phrase) that can be potentially used as the (clickable) body of a hypertext *link*. By *anchor*, we understand an actual instance of a term used as the body of a hypertext link. We will refer to instances of the Wikipedia collection written in different languages as *Wikipedia versions*. Every Wikipedia page describes a *concept*. The name of the described concept is usually provided as the title of the Wikipedia page. Though concepts are, in principle, language independent, we will refer to the page an ordinary monolingual link points to only as *concept* and to an equivalent page in another language as the *equivalent concept*. A *link* is consequently defined by an anchor-concept pair and a *cross-language link* by an anchor-equivalent concept pair. Alternatively, the CrossLink terminology uses the term *target* to refer to the concept linked by an anchor and the term *outlink* to refer to a link from a particular concept. We can say that every anchor in a Wikipedia version links to a concept in the same Wikipedia version. A concept in a Wikipedia version can have an equivalent concept in another Wikipedia version. A concept can be linked to from many (synonymous) anchors. Different anchors can use the same term to link to different concepts (we say the term can refer to multiple *senses*).

The CrossLink task can be described as follows: given a

[1]Our most successful methods in the English to CJK task were not evaluated in the CJK to English task (see Section 3.1).

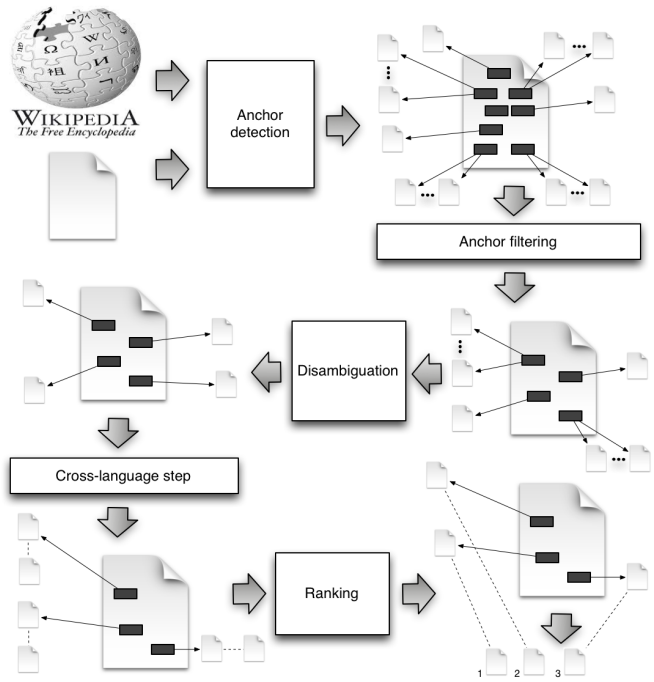


Figure 1: Cross-Lingual Link Discovery process

new concept *(orphan document)*[2] in the source language, the goal is to identify a ranked list of suitable anchors in the orphan document and link them to relevant concepts (*targets*) in the destination language version of Wikipedia.

Our methods solve the task in the steps illustrated in Figure 1. Each step is now described in detail.

2.2 Anchor detection

For the purposes of anchor detection, we compiled dictionaries of Wikipedia candidate anchors and concepts for each language. Each anchor corresponds to at least one concept. For example, the English dictionary contains about 14 million terms corresponding to about 4.2 million concepts. We then look up all occurrences of the dictionary terms in the orphan document. To make the anchor detection process quick, we first load the dictionary content into memory using the *trie* data structure and then perform in one pass through the orphan document the identification of dictionary terms in the text of the orphan document.

2.3 Anchor filtering

The anchor detection step produces many candidate anchors with a very low frequency of occurrence in a general corpus. We measure the prior probability of a term appearing as an anchor to asses how likely a term represents a good anchor. We define this probability as:

$$p(a) = \frac{N_a}{N_t}, \tag{1}$$

where N_a is the number of terms t appearing as an anchor a and N_t is the number of terms t in the collection. To make this probability technically easier to calculate, we estimate it at the granularity of documents. In this case N_a refers to the number of documents where term t appears as an anchor (i.e. term t occurs in the document at least once as an anchor). N_t refers to the number of documents where term t appears. We use an index of the appropriate Wikipedia version to obtain the values of N_a and N_t. Anchors detected in the previous step not satisfying the condition $p(anchor) > \theta$ where θ is a threshold are discarded from further processing. In our runs, we experimentally set this threshold to 0.001.

[2]The term orphan document is used to by the task organisers to refer to a new Wikipedia page without any link markup.

2.4 Disambiguation

In the disambiguation step, we select one out of n possible concepts for the detected anchor. The mappings from anchors to concepts is part of the dictionary extracted from Wikipedia we used in the anchor detection step. Using this mapping, given an anchor in one language, we can look up all n possible senses (concepts) of that anchor in that language. This gives us the set of Wikipedia pages the anchor can link to. We calculate a score for each of the available concepts and choose the concept with the highest score.

The scoring measure $s(c, a)$ makes use of two components: (a) the conditional probability of concept c given anchor a and (b) the similarity of anchor's context ctx_a with the text describing concept ctx_c in the source language.

$$s_{c,a} = \alpha p(c|a) + \beta sim(ctx_a, ctx_c), \tag{2}$$

where α and β are parameters. While these parameters can be estimated using machine learning techniques to achieve optimal performance, in our runs, we experimentally set $\alpha, \beta = 0.5$ and generally found the system to perform well.

2.4.1 The probability component

We define the conditional probability of a concept c given an anchor a using the Bayes' rule as follows:

$$p(c|a) = \frac{p(c)p(a|c)}{p(a)}, \tag{3}$$

We can estimate $p(a)$ as $p(a) = \frac{N_a}{N_{|A|}}$, where N_a corresponds to the number of occurrences of anchor a and $N_{|A|}$ the number of occurrences of all anchors. We can calculate $p(c)$ as $p(c) = \frac{N_c}{N_{|A|}}$, where N_c is the number of occurrences of (any) anchor representing concept c divided by the total number of occurrences of all anchors $N_{|A|}$. We further estimate $p(a|c)$ as:

$$p(a|c) = \frac{N_{a \cap c}}{N_c}, \tag{4}$$

where $N_{a \cap c}$ denotes the number of occurrences anchor a represents concept c. We can then rewrite equation (3) as

$$p(c|a) = \frac{\frac{N_c}{N_{|A|}} \cdot \frac{N_{a \cap c}}{N_c}}{\frac{N_a}{N_{|A|}}} = \frac{N_{a \cap c}}{N_{|A|}} \cdot \frac{N_{|A|}}{N_a} = \frac{N_{a \cap c}}{N_a}. \tag{5}$$

2.4.2 Context similarity component

In our submission, we tested two similarity methods for the purposes of concept disambiguation:

Explicit Semantic Analysis (ESA) – is a method that calculates semantic relatedness of two texts by mapping their term vectors to a high dimensional space (typically, but not necessarily, the space of Wikipedia concepts) and calculates cosine similarity between these high dimensional vectors. Measuring semantic similarity using ESA has been previously found to produce better results than calculating similarity directly on document vectors using cosine and other similarity measures and it has also been found to outperform the results that can be obtained by measuring similarity on vectors produced by Latent Semantic Analysis (LSA) [Gabrilovich and Markovitch,2007]. We have previously explored

the use of ESA in the context of link and cross-lingual link discovery in [Knoth et al.,2011b]. Since ESA is a method for calculating the similarity of two textual fragments, we apply it to measure the similarity of the context of the anchor being disambiguated with the textual fragments defining the concepts the anchor can be referring to. We define the context of the anchor as the sentence in which the anchor occurs. The context of the concept is defined as the first paragraph of the article describing the concept.[3]

Link Co-occurrence Similarity – calculates the proportion of Wikipedia pages where there occurs both (a) an anchor linking the concept being investigated and (b) an anchor that matches the title of the orphan document. Let t denote the title of the orphan document, c an anchor text referring to the concept investigated and P the set of all Wikipedia pages. We then define the link co-occurrence similarity $lcs(t, c, P)$ as

$$lcs(t, c, P) = \frac{|p \in P : t \in p \wedge c \in p|}{|p \in P : t \in p \vee c \in p|}. \quad (6)$$

The *lcs* similarity follows the idea that the similarity of two concepts (one representing the orphan document and the second one a Wikipedia page) is expressed by the proportion of Wikipedia pages where both concepts occur together.

2.5 Cross-language step

The goal of the cross-language step is to find an equivalent concept in the target Wikipedia version to the concept selected in the disambiguation step. In many cases, Wikipedia contains links between pages in different language versions referring to the same concept. In those cases, the cross-language step is straightforward. If a cross-language link is missing for the concept we need to translate, we can make use of the fact that the *same-as* relation is transitive. Therefore, we can try to find the cross-language link using other Wikipedia language versions. For example, there might be no direct cross-language link for translating a concept represented by an English page to Korean, but there might be a link from English to Vietnamese and from Vietnamese to Korean for that concept.

Our implementation uses the following language versions of Wikipedia in this order: English, German, French, Italian, Dutch, Japanese, Chinese, Korean, Vietnamese. We look for transitive relationships using breadth first search. If a translation for a concept is not found, the concept is discarded from further processing.

While we have observed that having more Wikipedia versions allows us to translate a higher proportion of concepts for the CrossLink language combinations, we believe that a much higher improvement could be seen if the transitivity assumption is applied to language combinations not involving the most resourced Wikipedia language – English.

2.6 Ranking

In the ranking step, each discovered *(source language) anchor – (target language) concept* pair (link) is assigned a rank. All pairs are then sorted in a descending order according to the their rank and returned in the specified output format (*result list*). Our results show that the ranking phase has a substantial impact on the overall results (see Section 4). We have experimented with three ranking methods receiving unexpected, but interesting results:

[3]If the first paragraph is shorter than five sentences we include two or more paragraphs.

Anchor probability ranking – is a method which assigns as a rank the anchor probability $p(a)$ used in the anchor filtering step (Section 2.3). Despite its simplicity, this ranking strategy yielded surprisingly good results.

Machine learned ranking – learning optimal ranking from data is a common strategy in information retrieval [Liu,2009]. To test this approach in the context of CLLD, we have extracted a set of features that can be useful for the ranker. We have then trained a ranking Support Vector Machine (SVM-rank [Joachims,2006]) to learn the optimal ranking model using the pointwise approach. We have then tested all different combinations of the features and also each feature independently. The tested features included:

- *Generality* – the depth of the concept page in the Wikipedia category graph.
- *Category distance* – the shortest path from the orphan document to the concept's page in the category graph normalised by two times the maximum depth.
- *Tfidf* – the term frequency of the term used as an anchor in the orphan document times the inverse document frequency of the concept.
- *Anchor probability* – the anchor probability described in Section 2.4.1.
- *Similarity* – The ESA or link similarity described in Section 2.4.2.
- *Relative position* – four features corresponding to the normalised first, last and average position and the position distance of the first and the last occurrence of the anchor in the orphan document.

Surprisingly, we have not seen any combination of these features to outperform in terms of MAP our single best feature - the anchor probability - on its own. Therefore, we decided for simplicity to drop the use of SVM model in our ranking completely. We think this is an interesting negative result. It remains to be determined whether better results can be achieved with these features if the ranking model is trained using the pairwise or listwise approach [Liu,2009] instead of the pointwise approach.

Oracle ranking – is a non-deterministic approach in which we produce random ranks and test the generated result list against the evaluation tool in the F2F Wikipedia ground truth (GT) setting. The ranking of the best performing result list is then used.

In the experimentation process, we discovered that our methods often generate a low number (significantly less than the allowed 250) but high quality links. Since this can still lead to a decrease in performance, in some of our runs, we top up the result list with additional links until all allowed link slots are used. One strategy is to add alternative disambiguations (i.e. to take the second best, third best, etc. disambiguated concepts for an anchor). We will further discuss this strategy in Section 4.

3. EXPERIMENTS

3.1 KMI runs

KMI submitted two runs for each E2CJK combination and three runs for each CJK2E combination (15 runs altogether). All of the KMI runs follow the pattern described in Figure 1 (i.e., 1. Anchor detection, 2. Anchor filtering, 3. Disambiguation 4. Cross-language step, 5. Ranking). The names of the runs code the choices we made in the disambiguation (step 3) and ranking phases (step 5) as described in Table 1. The column SIM indicates whether ESA or link

Run suffix	SIM	ADD	RANK
E2CJK runs			
01-ESA	ESA	Y	APR
02-ORC	ESA	Y	ORC
CJK2E runs			
01-LIS	LIS	Y	APR
02-ORC	LIS	Y	ORC
03-LIS	LIS	N	APR

Table 1: KMI runs description

similarity (LIS) was used in the disambiguation phase. The column ADD indicates (Y/N) if additional low scoring disambiguations were added in the result set. RANK indicates if oracle ranking (ORC) or anchor probability (APR) were used in ranking. Our CJK2E runs differed from the E2CJK runs in the disambiguation phase. While we used ESA in E2CJK, at the time of submission, we did not have a running instance of ESA for Chinese, Japanese and Korean and therefore used the link similarity approach only.

3.2 Evaluation

The methods have been evaluated at different granularity levels anchor-to-file (A2F) and file-to-file (F2F). There were two evaluation modes: a) GT is derived automatically from the existing link structure of Wikipedia (Wikipedia GT) and b) all anchors and targets are pooled and the evaluation is carried out by a human assessor (Manual assessment). *Precision-at-N (P@N)*, *R-Prec*, and *Mean Average Precision (MAP)* were used as the main performance metrics. More information about GT, the evaluation setup and a detailed description of the evaluation measures can be found in the overview paper [Tang et al.,2013] and Section 4.2.

The results for all experiments, including a theoretical boundary for F2F Wiki GT explained in Section 4.2, are summarised in Table 2. Graphs 2, 3, 4, 5, 6, 7 show the performance of the designed methods for different language combinations and assessment strategies.

3.3 Performance comparison with other teams

The NTCIR-10 CrossLink-2 organisers reported [Tang et al.,2013] that overall, our methods achieved the highest scores in multiple evaluation scenarios (measured with different metrics: *LMAP, R-Prec, Precision-at-N* in different evaluation levels against different GTs) for E2CJK. KMI methods are consistently the top (mostly among the top three) performers in the CJK2E task.

3.4 How can the performance be improved?

There is a number of ways in which our methods could be improved and optimised for better performance. We see the main possibilities in:

The use of ESA for disambiguation in CJK2E – Our methods utilised ESA only in E2CJK tasks where it performed consistently better than link similarity, which was used in all CJK2E experiments. Yet, ESA can be in a straightforward way adapted for Asian languages.

Anchor detection - We have compared our results with the runs of other teams and discovered that our system did not detect anchors that were only part of a term and we also did not use stemming. For example we did not detect anchor *plaque* in term *plaque-reducing* and anchor *Korea* in term *Korean king* (while links to *Dental plaque* and *Korea* were in GT). In English Wikipedia, anchors that are not composed of whole tokens do not exist, but it remains to be determined whether generating them can be useful. In addition, we discovered that in the distributed orphan documents end of line characters were often missing, which resulted in the concatenation of some words, such as *poultry* into *poultryand*, and also not all markup was removed, which is why we did not detect anchor *Peking duck* in string *Peking duck.JPG*. Consequently, the fact that our anchor detection algorithm assumed anchors to be composed only of whole terms had a significant negative impact on the performance of our runs.

Tuning parameters in the disambiguation step – In our submission, we have set the parameters α and β used as weights for the similarity and probability components in the disambiguation stage as equal, however we expect it would be possible to tune (or machine learn) these parameters to achieve better performance. Such approach would be similar to the one reported in [Milne and Witten,2008].

Considering more than one disambiguation per anchor in the first step – There are many situations when it makes sense for an anchor to link to multiple targets. For example, in the context of an article about American War, it can be relevant for the anchor *president* to link to the page explaining the general concept, the page about the *President of the United States of America* as well as the page about the 16th president of the United States *Abraham Lincoln*. While the Web (HTML) does not support by default multiple links per anchor, such approach can be easily put into practise and has been encouraged by the task organisers. Our implementation of the methods currently selects the best disambiguation in the first round and the second best, third best, etc. in the following rounds after the best, second best, etc. disambiguated concept is selected for each anchor. It might be possible to achieve better performance in manual assessment if more than one disambiguation is used in the first round. However, it is likely this would decrease the performance of the system in the Wiki evaluation as there is by definition only a single link per anchor in Wiki GT.

4. DISCUSSION

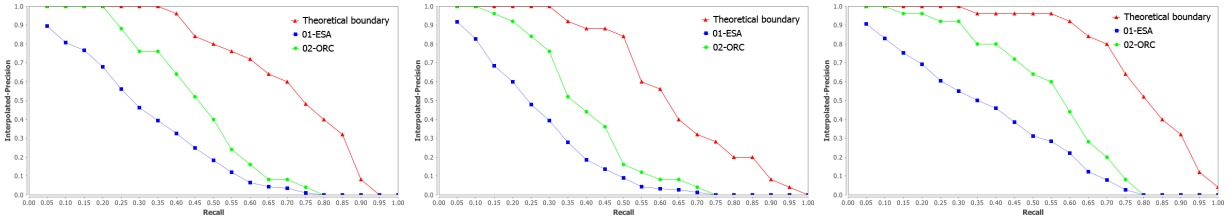
4.1 What have we learned?

ESA vs link similarity disambiguation – Our experiments show that ESA outperforms link similarity.

Ranking strategy – Ranking is a subtasks of CLLD with perhaps the highest influence on the final results. It is interesting that in our case, a trivial ranking technique produced better results than the SVM machine learned model using the pointwise approach. Regardless of whether better ranking features can be found and whether a better model can be trained using the pointwise or listwise approach, we believe that in order to develop more optimal ranking strategies, it is crucial to better understand the nature of the methods (where does the system make mistakes) and the task itself (what is exactly in GT). Our results demonstrate that while the optimal ranking techniques (ORC runs) with one GT (for which they were optimised) achieve substantially higher performance than our anchor probability ranking runs, the ESA runs perform equally well when applied to a different GT. This suggest the following: (a) the quite simple anchor probability ranking is almost as good as the oracle ranking leaving little room for improvement of ranking methods unless we want to over-fit them to achieve high performance on one particular GT, (b) it confirms how subjective the CLLD task is [Knoth et al.,2011b] and largely explains the high variability of the results of different systems under different evaluation settings.

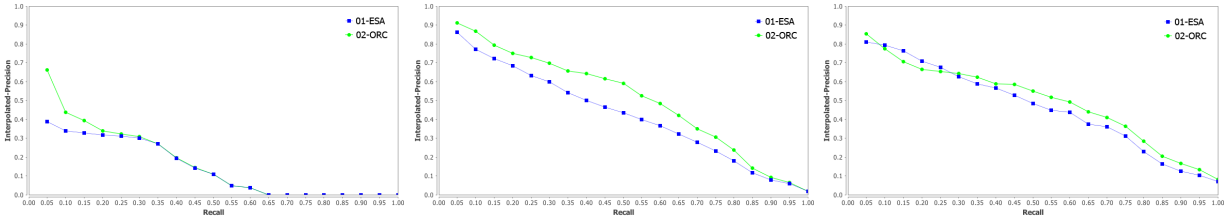
Run ID	LMAP	R-Prec	Run ID	LMAP	R-Prec
English-to-Chinese			**Chinese-to-English**		
F2F, Wikipedia ground truth			F2F, Wikipedia ground truth		
Theoretical boundary	0.652	0.652	Theoretical boundary	0.579	0.579
KMI-E2C-A2F-02-ORC	0.404	0.404	KMI-C2E-A2F-02-ORC	0.221	0.337
KMI-E2C-A2F-01-ESA	0.249	0.335	KMI-C2E-A2F-01-LIS	0.221	0.336
			KMI-C2E-A2F-03-LIS	0.219	0.336
F2F, manual assessment			F2F, manual assessment		
KMI-E2C-A2F-02-ORC	0.133	0.273	KMI-C2E-A2F-02-ORC	0.067	0.180
KMI-E2C-A2F-01-ESA	0.112	0.275	KMI-C2E-A2F-01-LIS	0.067	0.180
			KMI-C2E-A2F-03-LIS	0.064	0.180
A2F, manual assessment			A2F, manual assessment		
KMI-E2C-A2F-01-ESA	0.174	0.201	KMI-C2E-A2F-01-LIS	0.077	0.060
KMI-E2C-A2F-02-ORC	0.168	0.210	KMI-C2E-A2F-02-ORC	0.077	0.060
			KMI-C2E-A2F-03-LIS	0.076	0.060
English-to-Japanese			**Japanese-to-English**		
F2F, Wikipedia ground truth			F2F, Wikipedia ground truth		
Theoretical boundary	0.587	0.587	Theoretical boundary	0.641	0.641
KMI-E2J-A2F-02-ORC	0.341	0.341	KMI-J2E-A2F-02-ORC	0.224	0.224
KMI-E2J-A2F-01-ESA	0.206	0.285	KMI-J2E-A2F-01-LIS	0.114	0.176
			KMI-J2E-A2F-03-LIS	0.113	0.176
F2F, manual assessment			F2F, manual assessment		
KMI-E2J-A2F-02-ORC	0.450	0.513	KMI-J2E-A2F-02-ORC	0.171	0.271
KMI-E2J-A2F-01-ESA	0.383	0.424	KMI-J2E-A2F-01-LIS	0.138	0.202
			KMI-J2E-A2F-03-LIS	0.137	0.202
A2F, manual assessment			A2F, manual assessment		
KMI-E2J-A2F-02-ORC	0.452	0.337	KMI-J2E-A2F-02-ORC	0.072	0.058
KMI-E2J-A2F-01-ESA	0.440	0.279	KMI-J2E-A2F-03-LIS	0.062	0.042
			KMI-J2E-A2F-01-LIS	0.062	0.042
English-to-Korean			**Korean-to-English**		
F2F, Wikipedia ground truth			F2F, Wikipedia ground truth		
Theoretical boundary	0.744	0.744	Theoretical boundary	0.409	0.409
KMI-E2K-A2F-02-ORC	0.492	0.492	KMI-K2E-A2F-01-ORC	0.144	0.240
KMI-E2K-A2F-01-ESA	0.302	0.384	KMI-K2E-A2F-03-LIS	0.143	0.240
			KMI-K2E-A2F-01-LIS	0.143	0.239
F2F, manual assessment			F2F, manual assessment		
KMI-E2K-A2F-02-ORC	0.433	0.493	KMI-K2E-A2F-01-ORC	0.264	0.284
KMI-E2K-A2F-01-ESA	0.424	0.457	KMI-K2E-A2F-01-LIS	0.262	0.284
			KMI-K2E-A2F-03-LIS	0.260	0.284
A2F, manual assessment			A2F, manual assessment		
KMI-E2K-A2F-01-ESA	0.537	0.311	KMI-K2E-A2F-01-LIS	0.184	0.073
KMI-E2K-A2F-02-ORC	0.533	0.293	KMI-K2E-A2F-01-ORC	0.184	0.073
			KMI-K2E-A2F-03-LIS	0.180	0.073

Table 2: The summary of the KMI runs results.



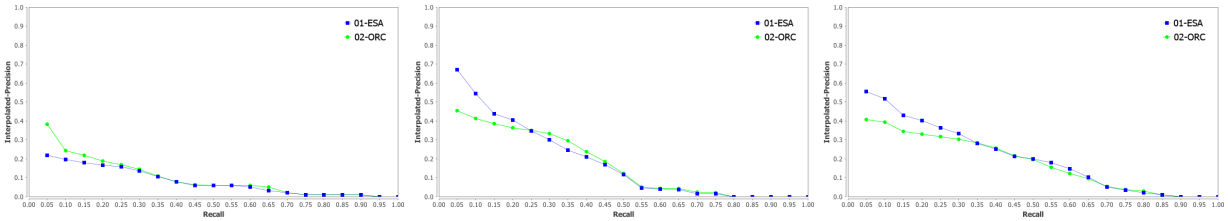
(a) Chinese (b) Japanese (c) Korean

Figure 2: E2CJK F2F evaluation results with Wikipedia ground truth



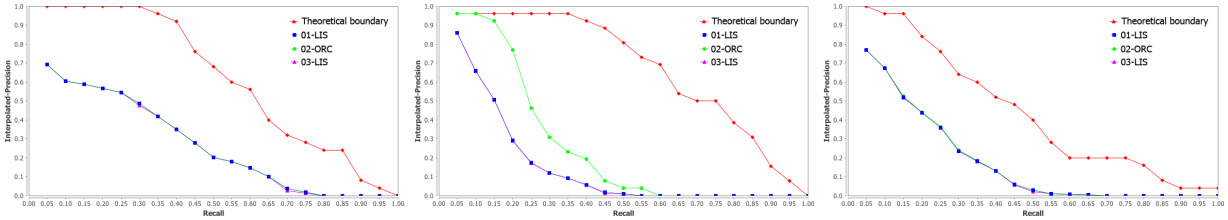
(a) Chinese (b) Japanese (c) Korean

Figure 3: E2CJK F2F evaluation results with manual assessment



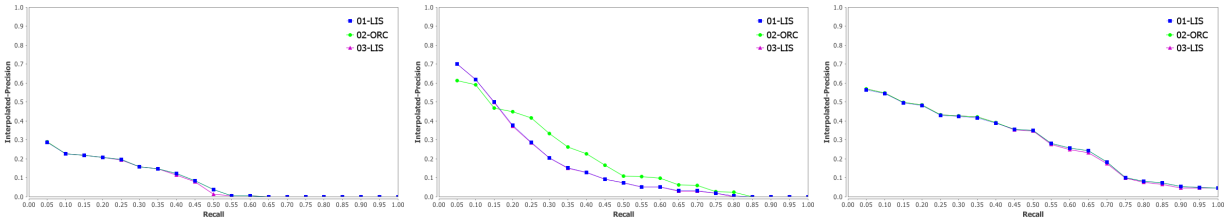
(a) Chinese (b) Japanese (c) Korean

Figure 4: E2CJK A2F evaluation results with manual assessment



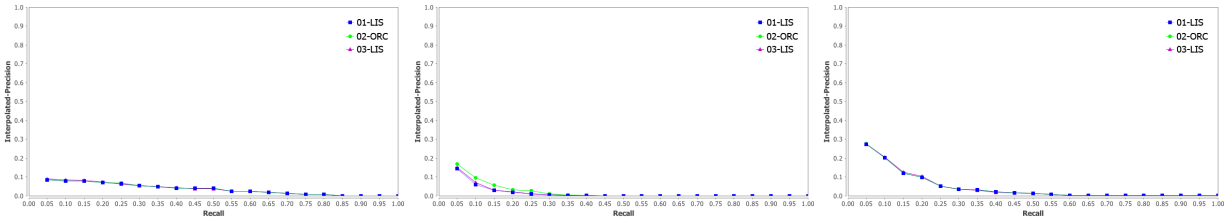
(a) Chinese (b) Japanese (c) Korean

Figure 5: CJK2E F2F evaluation results with Wikipedia ground truth



(a) Chinese (b) Japanese (c) Korean

Figure 6: CJK2E F2F evaluation results with manual assessment



(a) Chinese (b) Japanese (c) Korean

Figure 7: CJK2E A2F evaluation results with manual assessment

4.2 Evaluation methodology

The existence of a good evaluation framework which makes it possible to recognise and justify (both major and minor) improvements to the methods or reject method updates that do not improve performance is critical to the continuous technology progress of link discovery systems. A good evaluation framework will have the characteristic of assigning a system that produces better results a higher score than to a systems that produces worse results. This behaviour will be primarily stable (consistent from one set of topics to another) and reliable (an improvement in score will truly correspond to an improvement in user experience). The key to designing such an evaluation framework is to understand what is expected from an ideal system. The resemblance of the system's characteristics to the characteristics of the ideal system should then be captured by the framework as accurately as possible. As we will show now, designing an evaluation framework with these properties is certainly one of the main challenges of link discovery.

Since the system output is in CrossLink defined as a ranked list of anchor-target pairs, the performance of two systems can be compared by assessing their ranked lists. To do this, an evaluation framework will typically define (a) the set of (possibly graded) correct answers (ground truth - GT) and (b) the methods for calculating the score based on the system's answers (evaluation metrics). The CrossLink evaluation task [Tang et al.,2013] defines two GTs (the Wiki and the Manual assessment) and a set of evaluation measures, which are based on standard information retrieval metrics (MAP, R-Prec, Precision-at-n) and are applied on the participants' result sets at the A2F or F2F granularity.

Some of the limitations of the current evaluation approach, such as the inaccuracy/subjectivity of the Wiki GT, have already been widely known by both the participants and the task organisers. However, as we were designing and evaluating our methods for CrossLink-2, we identified a few more evaluation pitfalls about which we informed the task organisers. From our email conversation, it became clear that even they did not have a unanimous view on how these issues should be approached. As the knowledge of these issues contributes to the better understanding of the link discovery task, we discuss them here and propose how the evaluation framework can be improved in the future.

GT definition – the Wiki GT set for a given Wiki page (*topic*) is defined in CrossLink as the union of the concepts linked from either the source or the destination Wiki version (the source language concepts are mapped to their equivalent concepts in the destination Wiki version). Since equivalent pages in different Wiki versions often provide substantially different information on the same topic, there is consequently a low correlation (typically less than 0.2) of their respective link structures [Knoth et al.,2011b].

Therefore, the current approach has certain disadvantages one should be aware of. 1) An ideal system that will correctly identify all relevant anchors in the orphan document and will correctly link them to their relevant concepts in the destination Wikipedia version will not achieve 100% recall, because there is typically a large set of links in GT for which no relevant anchor in the orphan document exists. 2) Since Wiki GT evaluation is carried out only at the F2F level, a possible way how to achieve close to 100% recall would be to guess concepts, which are linked in the target language version of the orphan document and for which there does not exist any relevant anchor in the source document, and assign them any (even irrelevant) anchor in the orphan document. Although this strategy could potentially lead to better performance, we think it should be discouraged as it exploits a particular weakness in the evaluation methodology and changes the meaning of the CrossLink task.

The theoretical performance boundary – The findings reported in the previous paragraph lead us to measure the theoretical boundary in CrossLink-2 (F2F evaluation with Wiki GT). This boundary gives us the performance of an ideal system, which is constructed as follows: we take the original GT and remove from it all target language concepts for which there does not exist any relevant term (or even substring of a term) in the orphan document that could be used as an anchor pointing to this concept. The run submission is then constructed only from the remaining (correct) concepts in GT. The idea of the theoretical boundary is to find the maximum performance a CLLD system can achieve in this task. The calculation of the theoretical boundary is based on the November 2012 dump of Wikipedia with the CrossLink-2 GT. Although the calculated theoretical boundary can slightly change according to the Wikipedia version used, we consider the produced boundary depicted in Figures 2, 5 sufficiently accurate for the purposes of the CrossLink-2 evaluation. We believe that comparing the submitted runs with the theoretical boundary is more informative of systems' performance than the absolute evaluation scores. While the achieved absolute scores might seem in many cases quite low, it is possible to see from the comparison that, in particular in the E2CJK task, the performance of the CLLD systems is actually fairly good.

Ranking largely determines performance - We experimented with different ranking strategies for Wiki GT including the extreme cases where a system gets all the correct answers on the top or the bottom positions in the result list. We observed that ranking largely influences how successful a system is in the evaluation. Typically, by changing the order of anchors in the output file, we were able to get LMAP corresponding to both a top performing system as well as a system at the bottom of the evaluation chart. It directly follows from the way how LMAP is calculated that providing correct answers on the top positions is critical. Consequently, one of the problems with the application of LMAP in CrossLink is that the GT is unstable/subjective and the retrieved links are not equal, because some of the links are much more relevant than others. For example, in an article about *Japan* the link to *Tokyo* is certainly more important than the link to the *Michelin Guide*, yet systems are rewarded in the same way for retrieving any of them. This can lead to situations where systems with very different qualitative properties are assigned the same LMAP score. We think that a way to mitigate this issue (apart from the already used Manual Assessment) would be to apply one of the existing graded relevance evaluation metrics [Sakai,2009]. The graded GT could be constructed as a multiset union of links in all Wikipedia languages (instead of a set union of the two considered languages). We think this approach would not only lead to more informative results, but might also help stabilise the fluctuations in results of participants in different language combinations and evaluation settings.

The evaluation metric rewards certainty, not relevance – CrossLink aims to encourage the development of systems that can link an anchor to multiple concepts. The reas-

ons why this is useful are explained in Section 3.4. Consequently, the run submission format allows participants to report more than one target concept per anchor. However, the only allowed way of expressing this is by assigning all the concepts associated with that anchor a single position in the result list. This means, for example, that a system can in an article about *India* generate anchor *Gandhi* with links to *Mahatma Gandhi*, *Gandhi (film)* and *Gandhi (American band)* and must assign them a single position in the result list. The first link is certainly correct, the second link seems useful and the third link is certainly incorrect. The problem of this approach is that: A system (a) cannot provide any ranking for the generated concepts, i.e. all concepts are treated equal and the correctness of the anchor is evaluated according to Equation 5 in [Tang et al.,2013] as the proportion of those concepts that were correct and (b) cannot decide to link a concept with high relevance for a given anchor, then generate other anchors and eventually additional concepts with lower relevance for the given anchor.

Since the performance of a system is critically influenced by the links generated in the first positions, this leads to a situation in which systems are encouraged to first generate "low risk" anchors. Unambiguous anchors, which are by its nature difficult to get wrong, constitute this low risk. Therefore, an effective strategy is to choose less relevant, but certain anchors, before highly relevant but ambiguous anchors. As acknowledged by one of the organisers, the problem is that this approach rewards certainty, not precision. Also, according to Equation 5 in [Tang et al.,2013], a system cannot be rewarded for generating more than one target per anchor as from a strategic point of view, it is better to select one concept (about which a system is the most certain) rather than more concepts. The solution would be to allow the ranking in the output file at the granularity of targets (rather than at the granularity of anchors).

5. RELATED WORK

KMI @ CrossLink-1 vs KMI @ CrossLink-2 – The methods we applied in CrossLink-2 follow quite different strategies than the methods we used in CrossLink-1 [Knoth et al.,2011a]. While in CrossLink-1 we approached the problem as a similarity search task, in CrossLink-2 we see it rather as a disambiguation and ranking exercise. Both approaches have advantages and disadvantages. In CrossLink-1, we designed methods that are quite general and flexible in their ability to be applied to interlinking of non-Wikipedia contexts (e.g. newspapers, blogs or books) instead of just *wikifying*. On the other hand, the CrossLink-2 methods are much more tailored to the Wiki (or even Wikipedia) environment (and thus also closer to the methods of most other CrossLink participants). These methods consequently achieve better results on the CrossLink dataset. We think that in the future, such advantages and disadvantages of methods should be highlighted in the results overview and task participants would be also more encouraged in the development of methods that are applicable in a wider context. Currently, as such methods are unlikely to perform as well as methods specifically tailored to the Wikipedia collection, there is little incentive to develop them and submit them for evaluation.

Emphasis on ranking rather than classification – One of the fundamental differences of our approach (apart from its multilinguality) from the wikification approach of [Milne and Witten,2008] is the emphasis on ranking rather than classification. The system of Milne & Witten is configured to classify the set of generated candidate links into positive or negative categories and produce the positive links as a result set. Our approach (and in fact the CrossLink task specification) emphasises the importance of the ranking phase. The fact that the generated links are ranked makes it possible, for example, to develop a user interface where one controls the number of the generated links using a slider. Based on the number of links that should be generated, the system should, from the perspective of the user, display those that are the most relevant.

6. CONCLUSION

In this paper, we presented methods for Cross-Language Link Discovery (CLLD). The methods of our team achieved the best results in the E2CJK task and were the top performer in the CJK2E task, where we did not make use of such a solid disambiguation system as we deployed in the E2CJK task.

However, we believe the most important is the knowledge we acquired while carrying out experiments. We understood the importance of the ranking phase, experimentally confirmed the impact of high variance in the ground-truth on the CLLD results, measured the maximum (theoretical boundary) performance of an ideal CLLD system and analysed some of the evaluation pitfalls. We believe this knowledge will help us to better understand how to more representatively measure the performance in the future, which will, in turn, enable further evidence-based improvements of link discovery systems.

References

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *In Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1606–1611.

Thorsten Joachims. 2006. Training linear svms in linear time. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pages 217–226, New York, NY, USA. ACM.

Petr Knoth, Lukas Zilka, and Zdenek Zdrahal. 2011a. Kmi, the open university at ntcir-9 crosslink: Cross-lingual link discovery in wikipedia using explicit semantic analysis. In *NTCIR-9*.

Petr Knoth, Lukas Zilka, and Zdenek Zdrahal. 2011b. Using explicit semantic analysis for cross-lingual link discovery. In *CLIA - IJC-NLP 2011*, Chiang Mai, Thailand, November.

Tie-Yan Liu. 2009. Learning to rank for information retrieval. *Found. Trends Inf. Retr.*, 3(3):225–331, March.

David Milne and Ian H. Witten. 2008. Learning to link with wikipedia. In *Proceeding of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 509–518, New York, NY, USA. ACM.

Tetsuya Sakai. 2009. On the robustness of information retrieval metrics to biased relevance assessments. *JIP*, 17:156–166.

Ling-Xiang Tang, In-Su Kang, Fuminori Kimura, Yi-Hsun Lee, Andrew Trotman, Shlomo Geva, and Yue Xu. 2013. Overview of the ntcir-10 cross-lingual link discovery task. In *NTCIR-10*.