# KMI at NTCIR-10 CrossLink-2

# Simple Yet Effective Methods for Cross-Lingual Link Discovery (CLLD)

Petr Knoth petr.knoth@open.ac.uk, Drahomira Herrmannova d.herrmannova@open.ac.uk

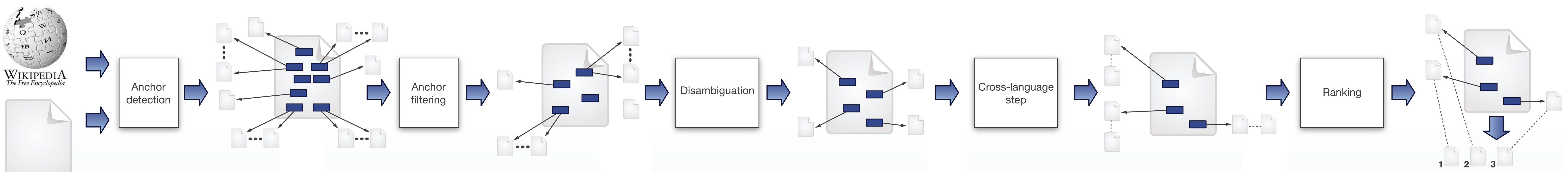
## Introduction

KMI submitted 15 runs in the NTCIR-10 CrossLink-2 evaluation achieving the best overall results in the English to Chinese, Japanese and Korean (E2CJK) task and being the top performer in the Chinese, Japanese, Korean to English task (CJK2E). All KMI methods are language agnostic and can be easily applied to any other language combination with sufficient corpora and available pre-processing tools.

## Submitted runs

| Run Suffix | Similarity method | Adding | Ranking |
|---|---|---|---|
| **E2CJK runs** | | | |
| 01-ESA | Explicit Semantic Analysis | Yes | Anchor probability ranking |
| 02-ORC | Explicit Semantic Analysis | Yes | Oracle ranking |
| **CJK2E runs** | | | |
| 01-LIS | Link similarity | Yes | Anchor probability ranking |
| 02-ORC | Link similarity | Yes | Oracle ranking |
| 03-LIS | Link similarity | No | Anchor probability ranking |

## Link discovery methods




### 1. Anchor detection

Look up all occurrences of dictionary terms in the orphan document.

Dictionaries of candidate anchors are pre-compiled for each source 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.

### 2. Anchor filtering

Discard anchors with low probability

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

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.

### 3. Disambiguation

Out of *n* possible concepts, select the one with the highest score

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

where *p*(c|a) is the conditional probability of concept *c* given anchor *a* and $sim(ctx_a, ctx_c)$ is the similarity of anchor's context $ctx_a$ with the text describing concept $ctx_c$, calculated using either Explicit Semantic Analysis (ESA) or link similarity (LIS) method.

### 4. Cross-language step

Find an equivalent concept in the target Wikipedia version to the concept selected in the disambiguation step.

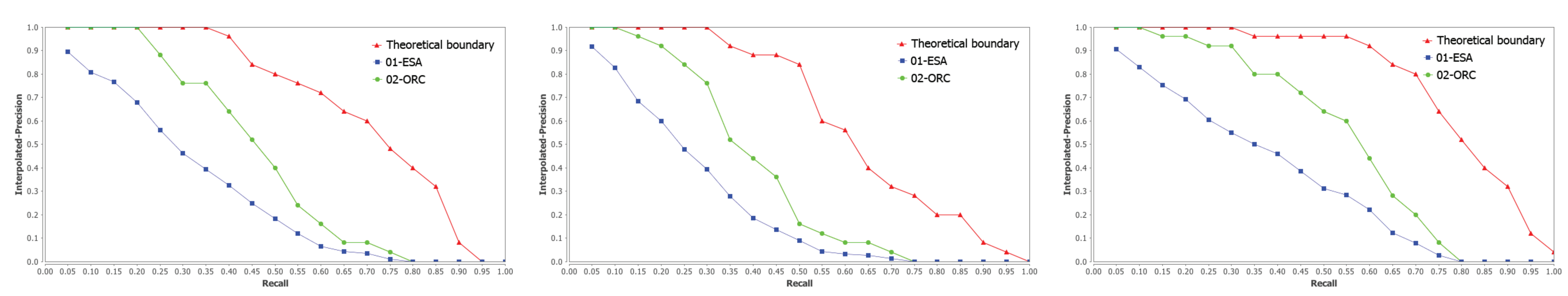
If a cross-language link is missing for the desired language combination, we make use of the fact that the cross-language relation is transitive. Therefore, the cross-language link can be sometimes acquired using other Wikipedia language versions.

### 5. Ranking

All anchor-concept pairs are ranked, sorted and returned in the specified output format.

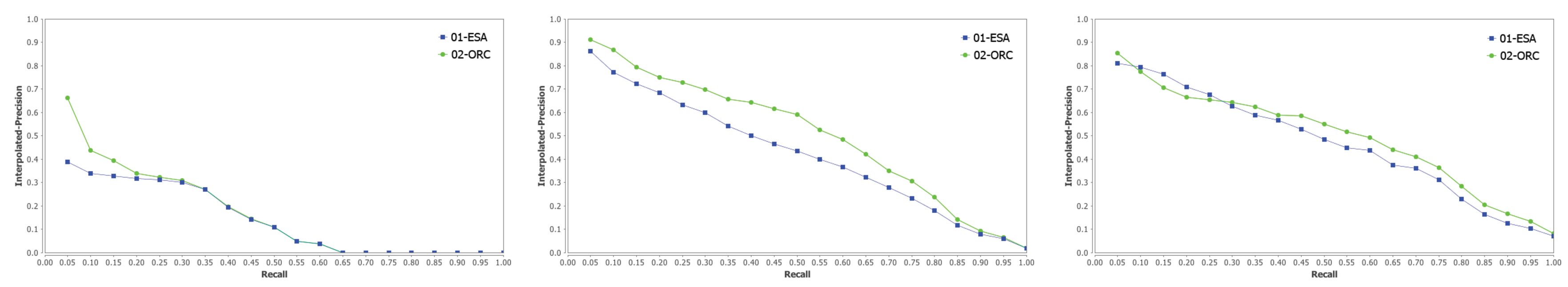
We have experimented with 3 ranking methods:
**1.** Anchor probability ranking
**2.** Machine learned ranking
**3.** Oracle ranking

## E2CJK Results



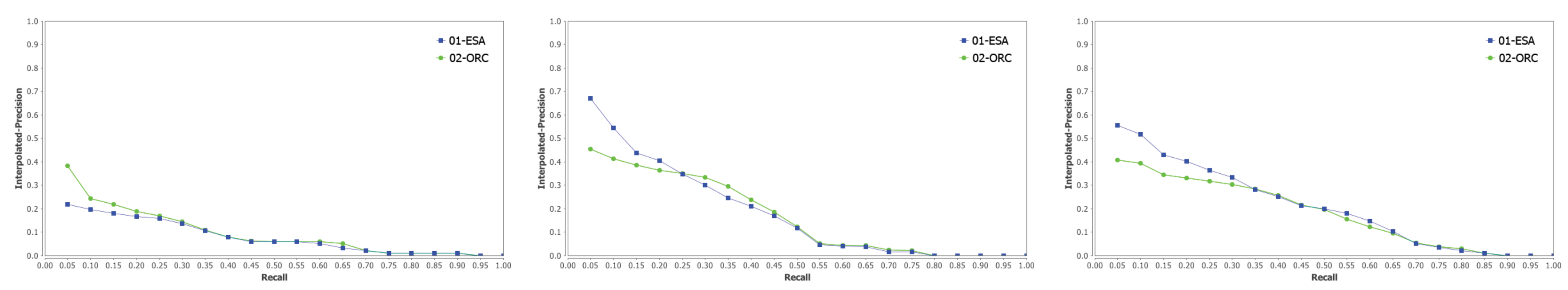
Chinese — Japanese — Korean

**Ground truth assessment**



Chinese — Japanese — Korean

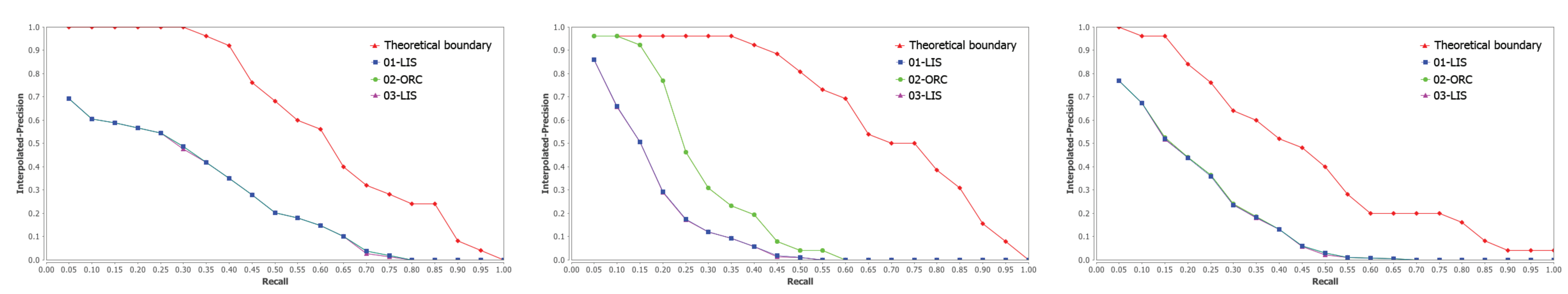
**Manual assessment, F2F**



Chinese — Japanese — Korean

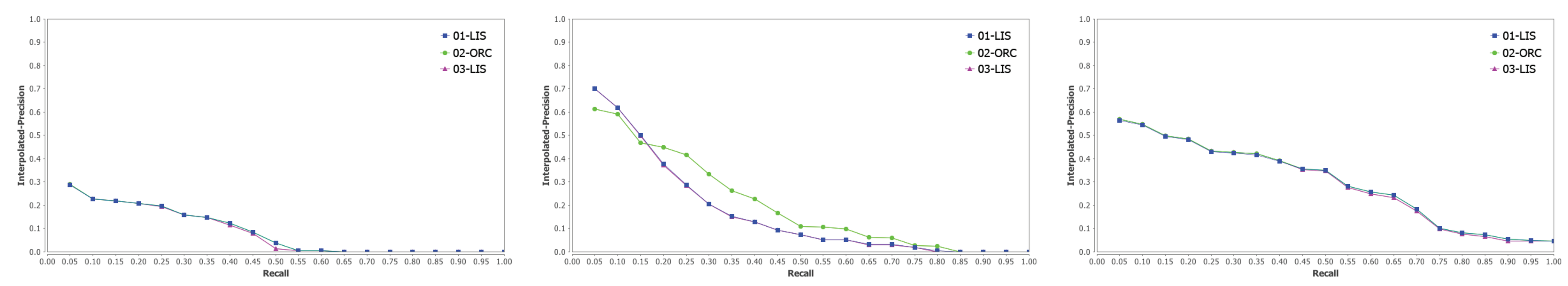
**Manual assessment, A2F**

## CJK2E Results



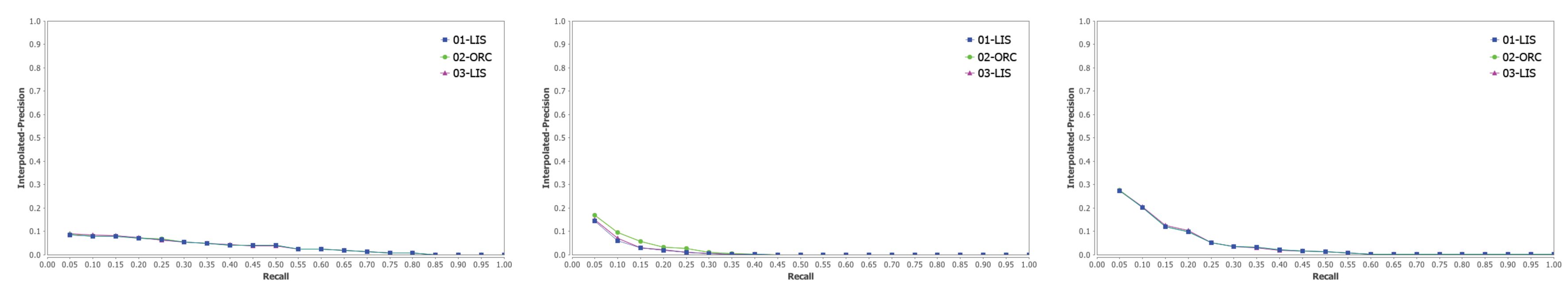
Chinese — Japanese — Korean

**Ground truth assessment**



Chinese — Japanese — Korean

**Manual assessment, F2F**



Chinese — Japanese — Korean

**Manual assessment, A2F**

## How to improve performance?

**The use of ESA for disambiguation in CJK2E:** ESA was applied in E2CJK tasks where it performed consistently better than link similarity.
**Anchor detection:** Our system did not detect anchors that were only part of a term, contrary to other systems.
**Tuning parameters in the disambiguation step:** It might be possible to determine more optimal disambiguation parameters by further tuning or machine learning.
**Considering more than one disambiguation per anchor in the first step:** our methods currently select the best disambiguation for each anchor in the first round and the second best, third best, etc. disambiguation only in the following rounds. It might be possible to achieve better performance in manual assessment if more than one disambiguation is assigned in the first round.

## What have we learned?

**ESA vs link similarity disambiguation:** Our experiments show that ESA outperforms link similarity.
**Ranking strategy:** While the optimal ranking technique (ORC runs) with the Wiki ground truth (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.

## 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. We think the evaluation framework can be improved in the following aspects.
**The theoretical performance boundary:** The theoretical boundary (see Graphs) gives us the maximum 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 evaluation metric rewards certainty, not relevance.** At the moment 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 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. The solution would be to allow the ranking in the output file at the granularity of targets (rather than at the granularity of anchors).
**GT definition and ranking strategy:** The currently established Wiki GT set is defined in a way which does not allow even an ideal system to achieve 100% recall. In addition, ranking largely determines how successful a system is in the evaluation. We think that a way to mitigate these issues 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).

## Conclusion

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.