KMI at NTCIR-10 CrossLink-2

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

Petr Knoth and Drahomira Herrmannova

KMi

Introduction

- Method overview
- What have we learned
- Evaluation methodology

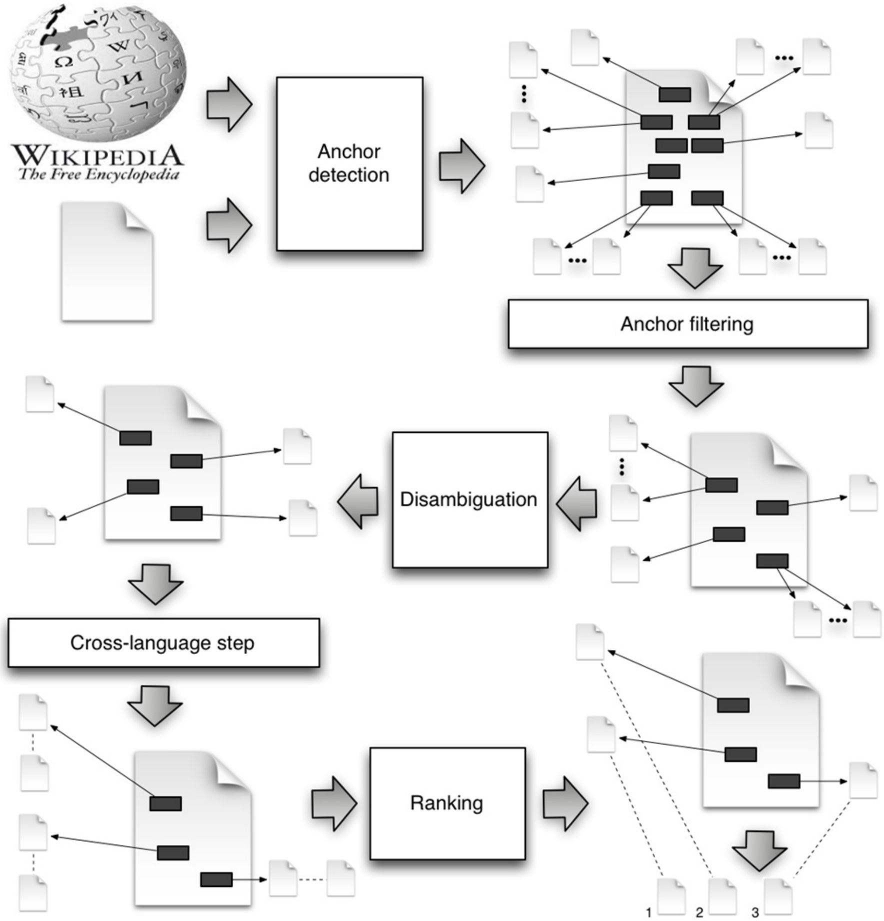
Method introduction

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

Definitions

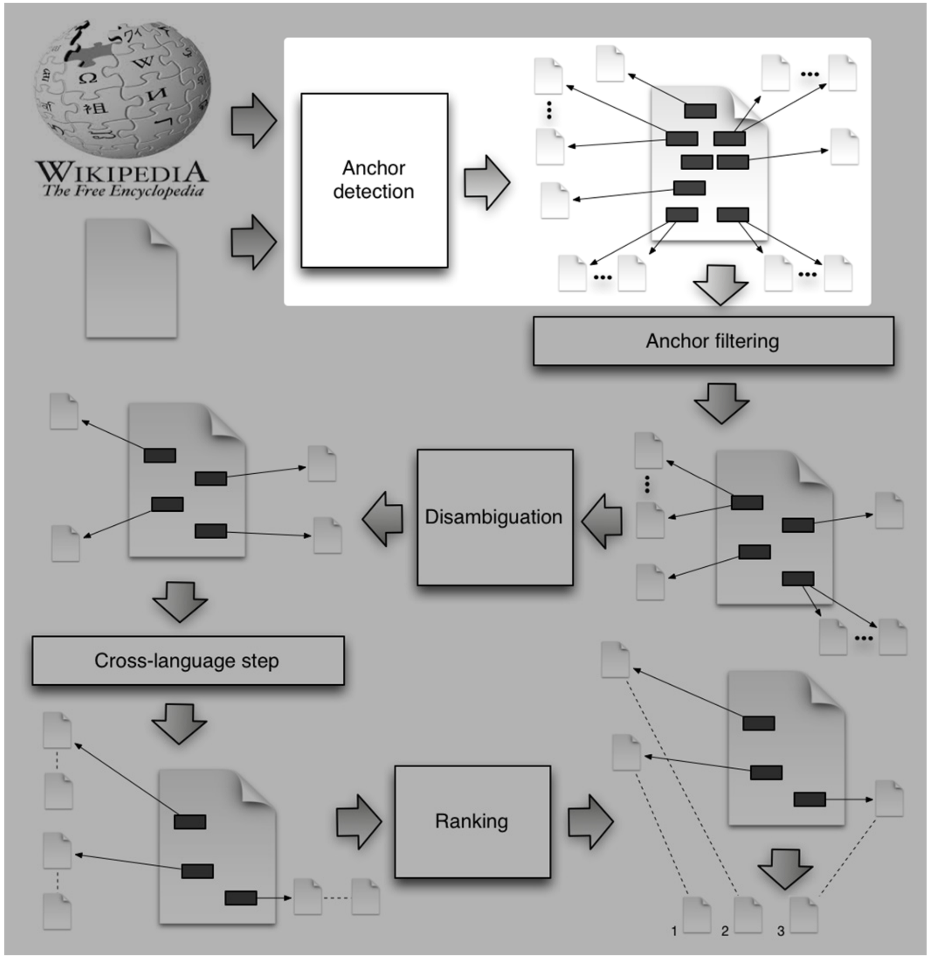
- *Term* – any textual fragment (typically a noun phrase) that can be potentially used as the (clickable) body of a hypertext.
- *Anchor* – an actual instance of a term used as the body of a hypertext link.
- *Wikipedia (language) version* – an instance of the Wikipedia collection written in a specific language
- *Concept* – every Wikipedia page describes a concept (its name provided as page title).
- *Link* – an anchor-concept pair
- *Target* – refers to the concept linked by an anchor

Method overview



KMi

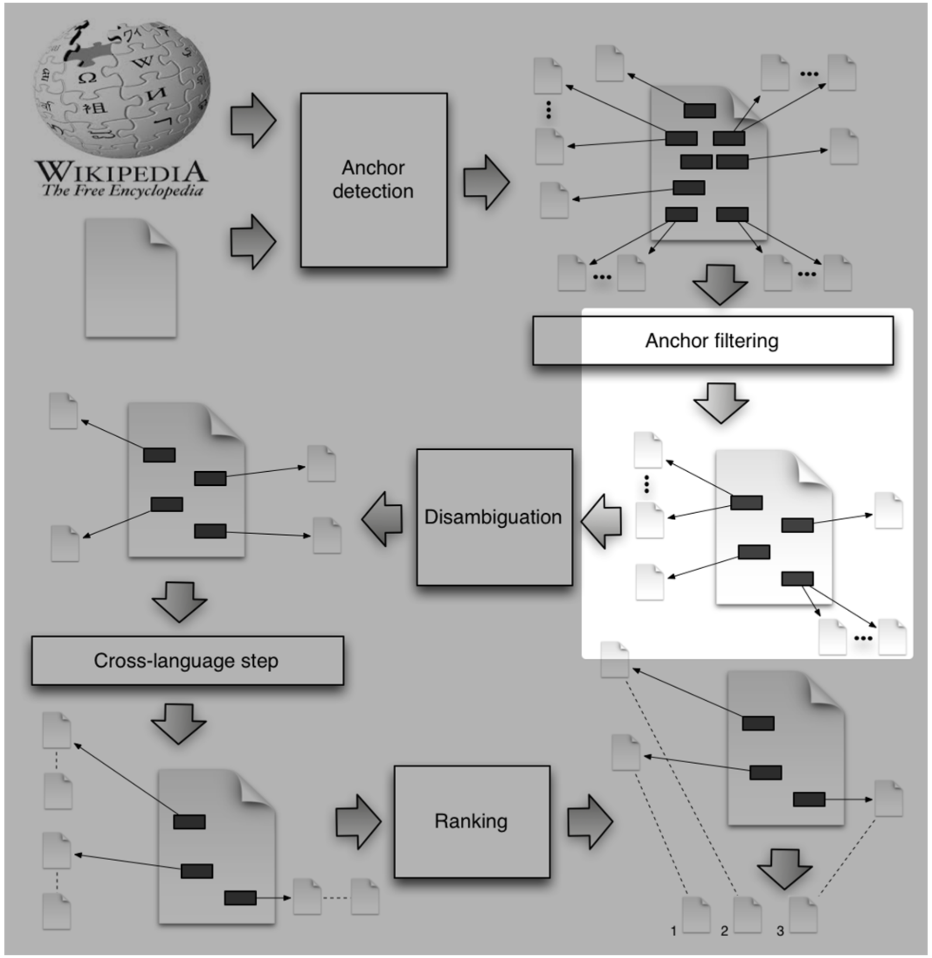
1. Anchor detection



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

2. Anchor filtering



KMi

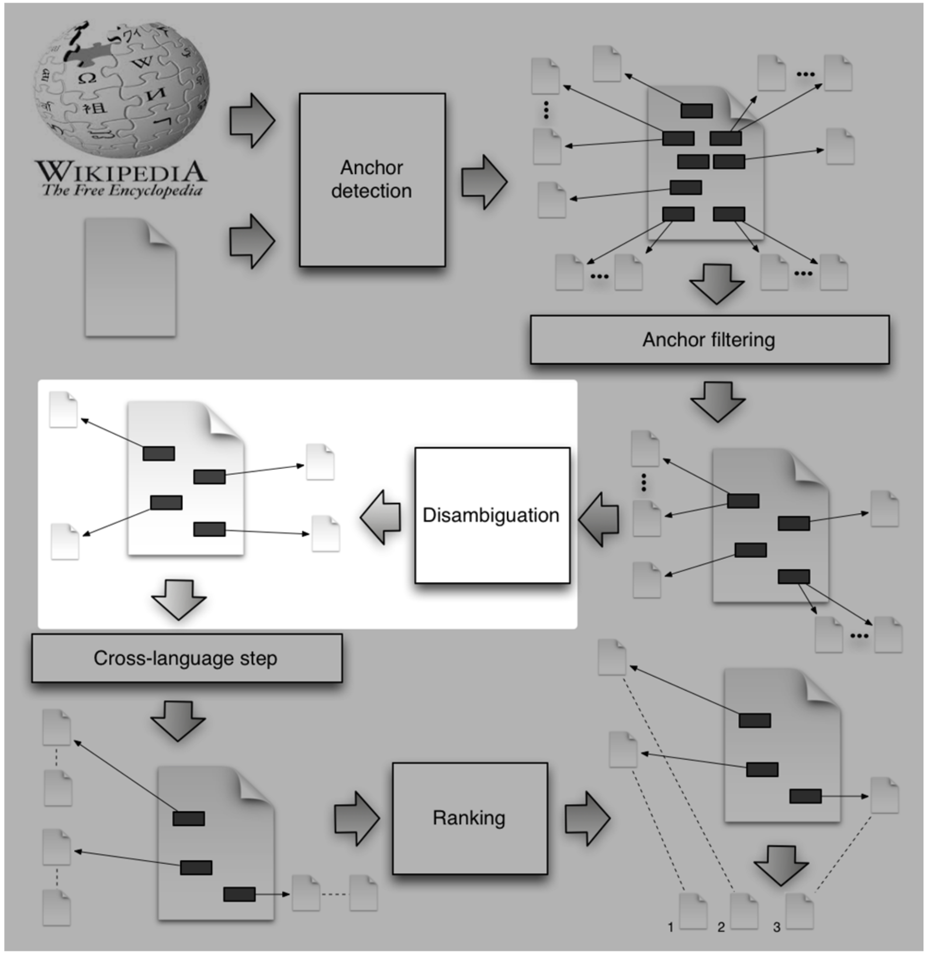
2. Anchor filtering

- 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*
- N_t is the number of terms *t* in the collection

3. Disambiguation



KMi

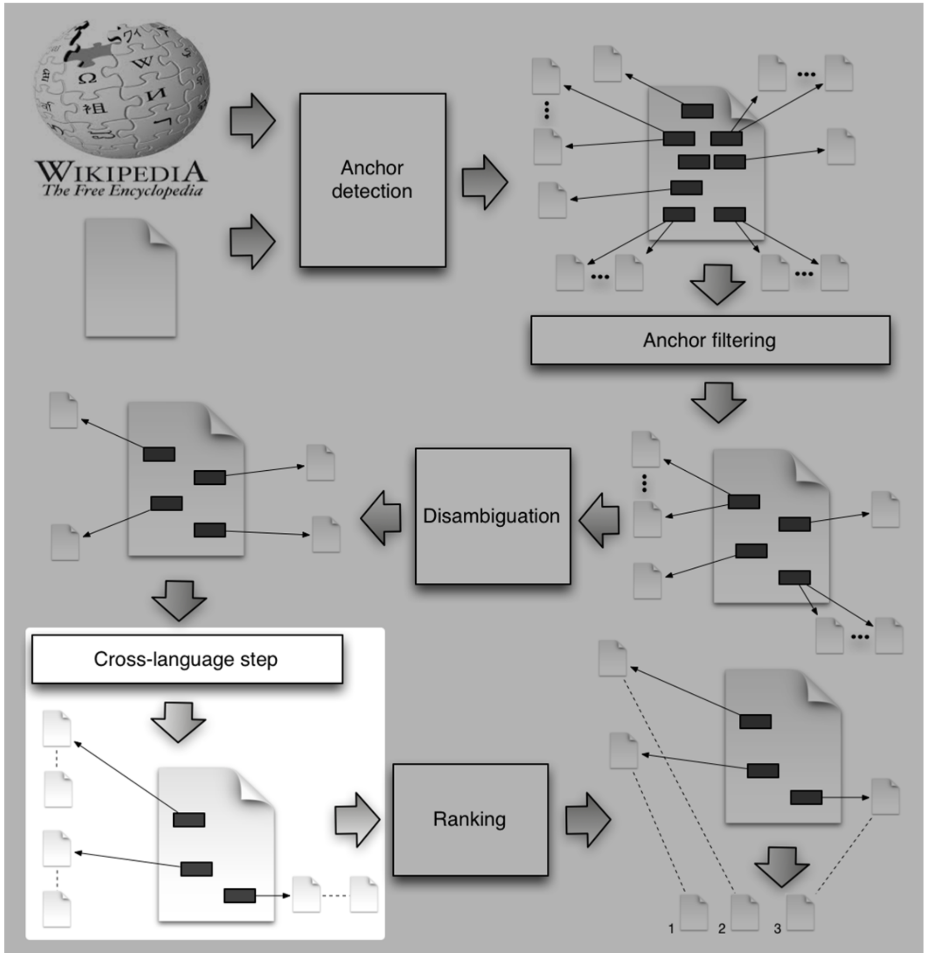
3. Disambiguation

- 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*
- $sim(ctx_a, ctx_c)$ is the similarity of anchor's context ctx_a with the text describing concept ctx_c, calculated using
  - Explicit Semantic Analysis (ESA)
  - Link similarity (LIS)

4. Cross-language step



4. Cross-language step

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

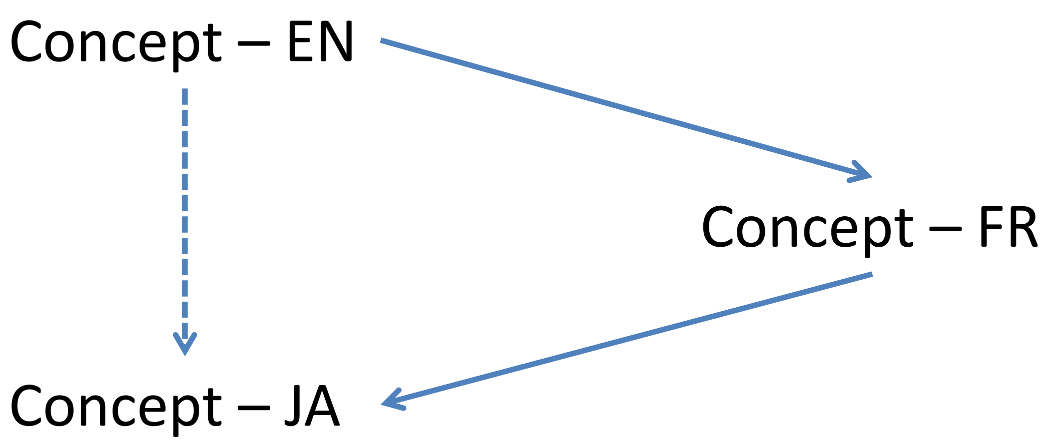
Concept – EN

↓

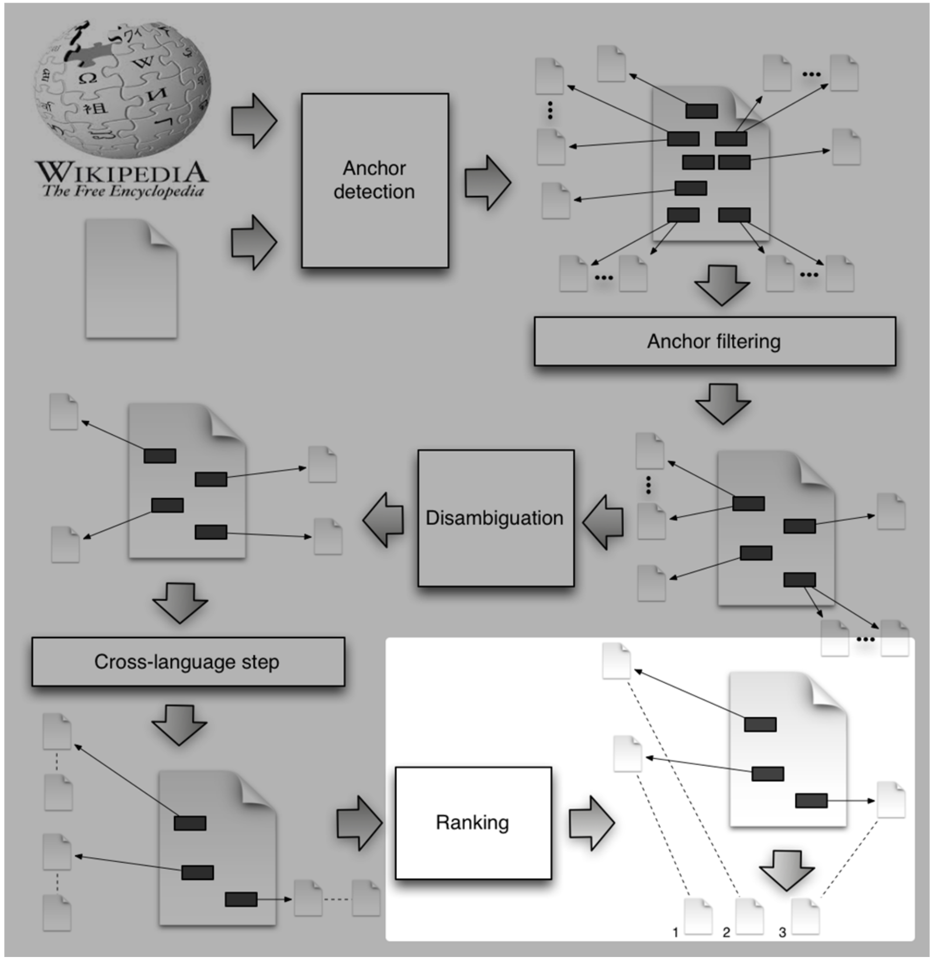
Concept – JA

4. Cross-language step – transitivity

- 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



KMi

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

Learning to rank features 1/2

- **Generality** - the depth of the concept page in the Wiki-pedia 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.

KMi

Learning to rank features 2/2

- **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.

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

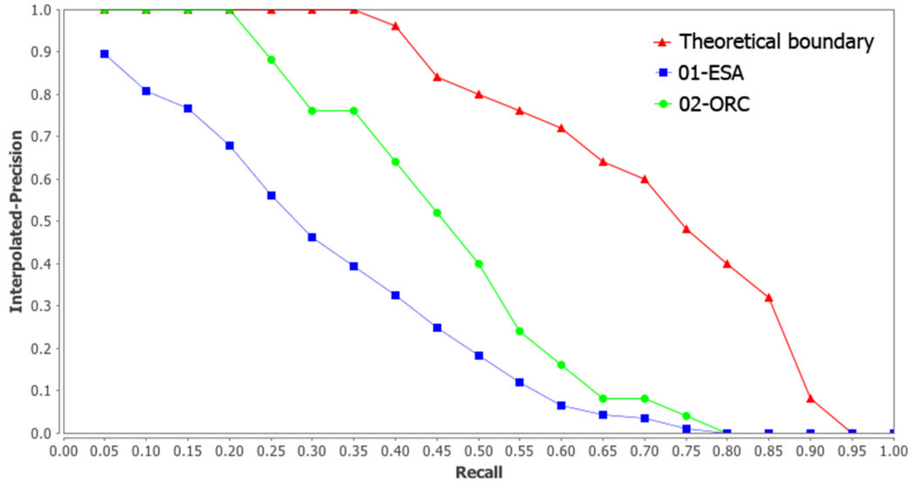
How to improve performance?

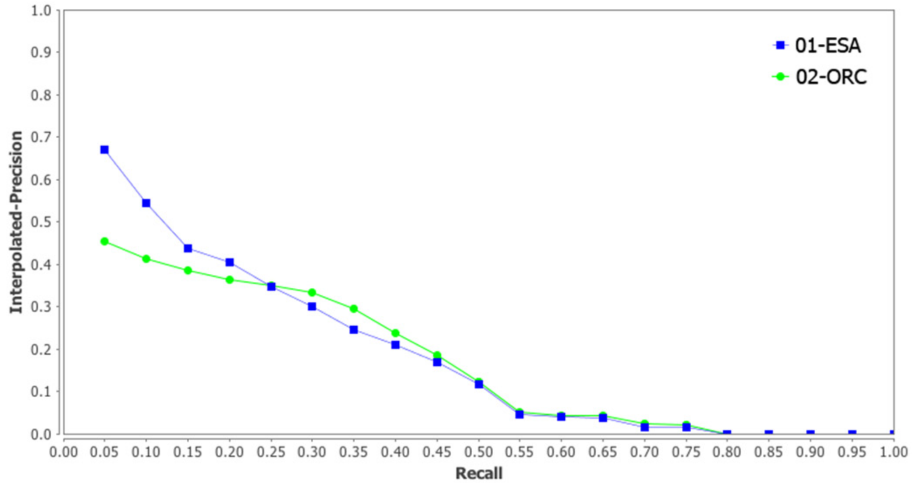
- The use of ESA for disambiguation in CJK2E
- Anchor detection
- Tuning parameters in the disambiguation step
- Considering more than one disambiguation per anchor in the first step

What have we learned?

- ESA vs link similarity disambiguation
- Ranking strategy – anchor ranking works as well as oracle ranking

Anchor ranking vs oracle ranking





KMi

Evaluation methodology

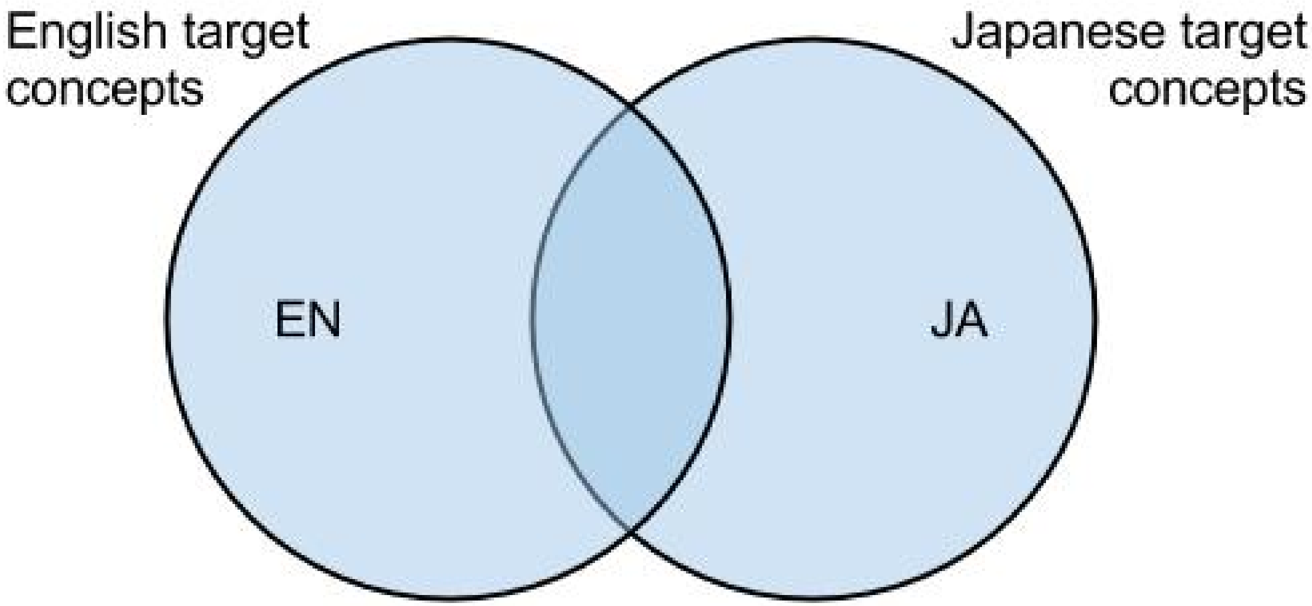
The existence of a good evaluation framework, which makes it possible to recognise and justify (both major and minor) improvements to the methods or reject method updates that do not improve performance, is critical to the continuous technology progress of link discovery systems. We think the evaluation framework can be improved in the following aspects.

KMi

Evaluation methodology

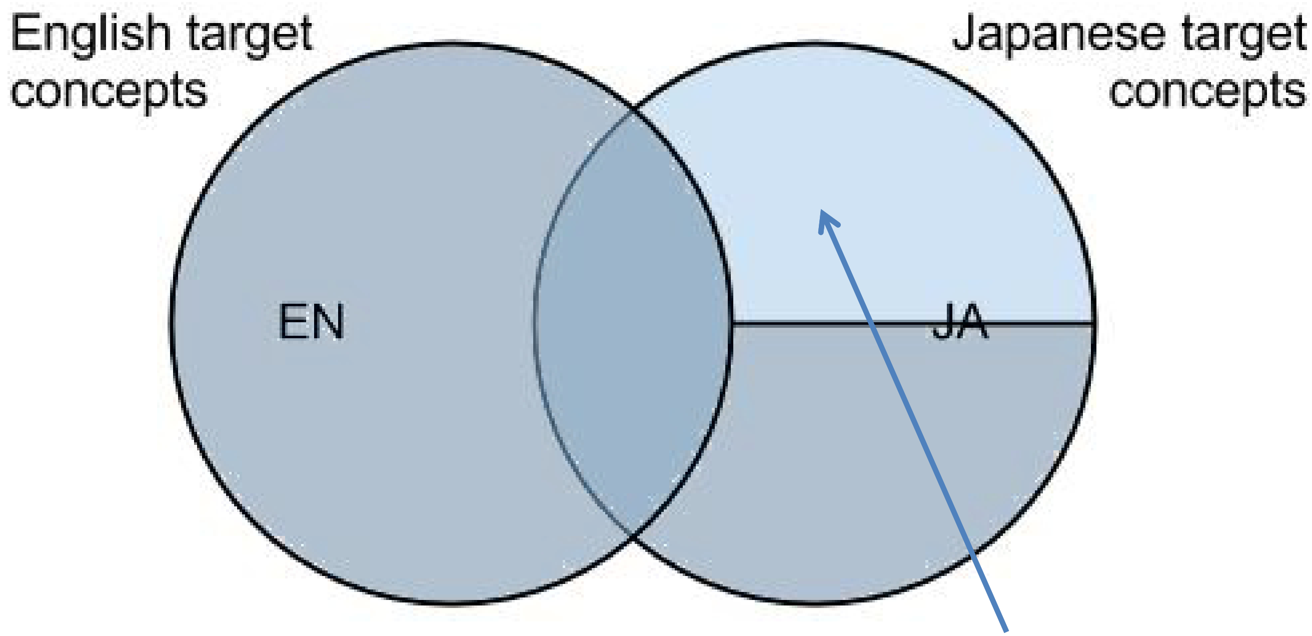
- GT definition
- The theoretical performance boundary
- The evaluation metric rewards certainty, not relevance

Ground truth definition



Definition of GT for each language combination

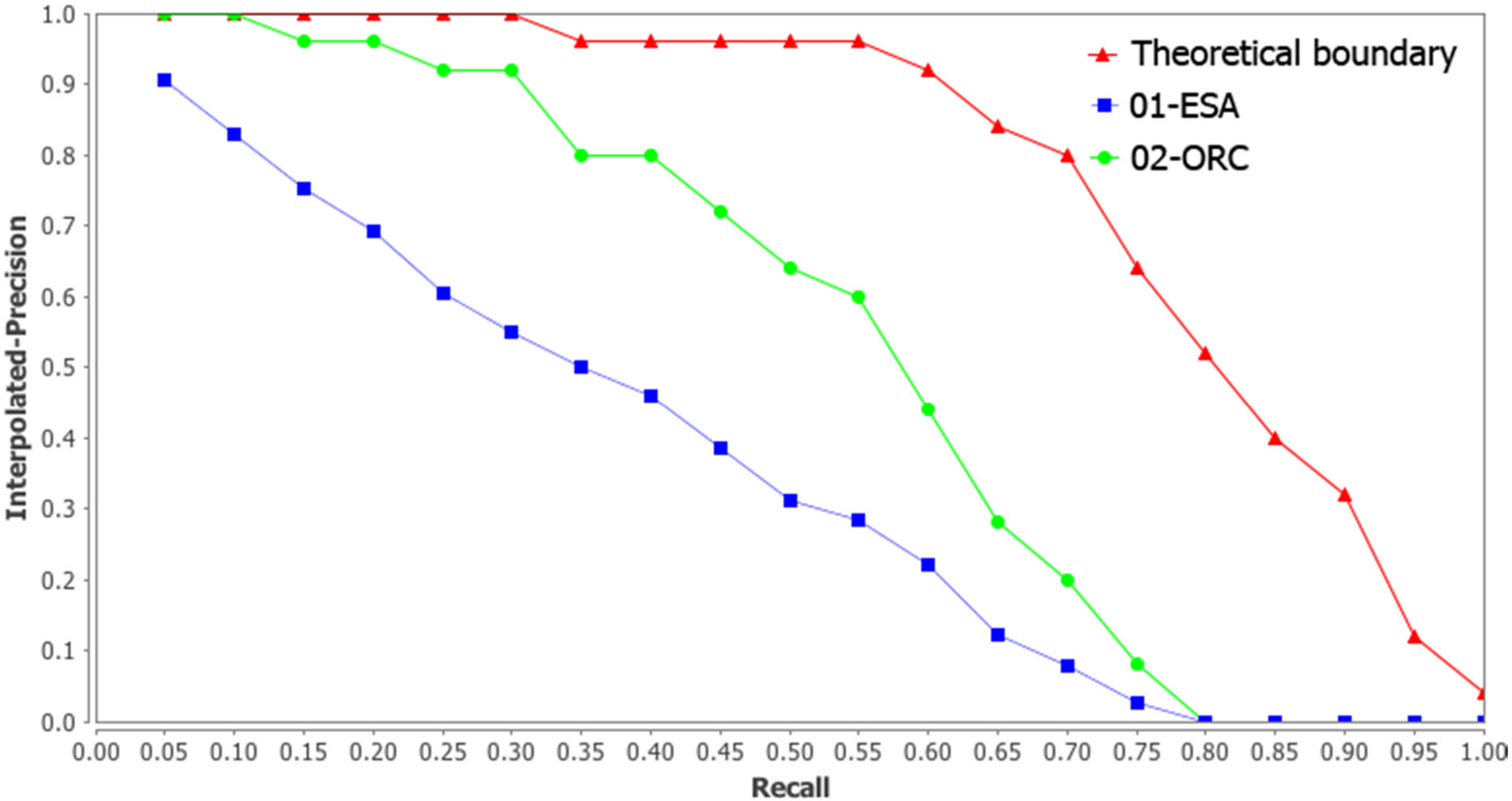
Ground truth definition



Concepts for which there is no suitable anchor in the orphan EN document

KMi

Theoretical performance boundary



Gives us the preformance of an ideal system

KMi

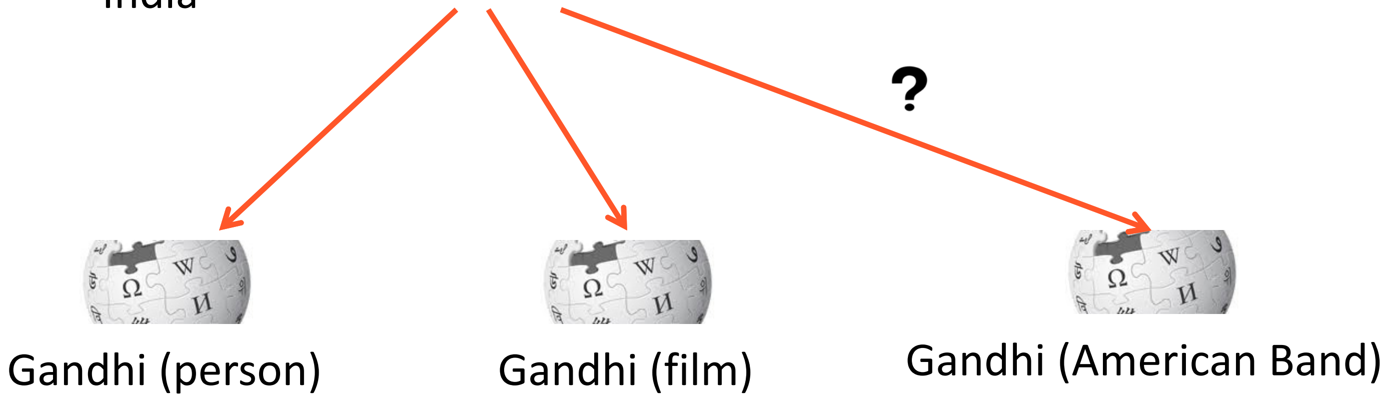
Alternative GT definition

- Very low agreement between GTs (~0.2) [Knoth, 2011]
- GT created as a multiset union of many Wikipedia versions' GT
- System answers are not binary but graded

The evaluation metric rewards certainty, not relevance

India

India became an independent nation in 1947 after a struggle for independence that was marked by non-violent resistance led by Mahatma Gandhi.



The evaluation metric rewards certainty, not relevance

Result set 1
Position 1
….
Position 2:
Gandhi (person)
Gandhi (film)
Gandhi (American Band)
Position 3:
….

Result set 2
Position 1
Gandhi (person)
Position 2:
Gandhi (film)
Position 3:
…
Position N:
Gandhi (American Band)

Result set 1 will get a lower MAP than result set 2.

An effective strategy is to prefer obvious unambigous links (such as India) over ambiguous relevant links (Gandhi).

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.