The Effectiveness Of Cross-lingual Link Discovery

Ling-Xiang (Eric) Tang†
Kelly Itakura†
Shlomo Geva†
Andrew Trotman‡
Yue Xu†

† QUT (Brisbane, Australia)
‡ University of Otago
Wikipedia

- Online wiki-based hypertext encyclopedia
- Contains articles on over 20 million topics
- Contains articles in 281 languages

- Has *extensive* hypertext links between documents in the same language
- Has *few* hypertext links between documents in different languages
## Our View Of Wikipedia

<table>
<thead>
<tr>
<th>No</th>
<th>Language</th>
<th>Wiki</th>
<th>Articles</th>
<th>Images</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>English</td>
<td>en</td>
<td>3,807,882</td>
<td>825,432</td>
</tr>
<tr>
<td>9</td>
<td>Japanese</td>
<td>ja</td>
<td>779,656</td>
<td>77,107</td>
</tr>
<tr>
<td>12</td>
<td>Chinese</td>
<td>zh</td>
<td>386,596</td>
<td>27,175</td>
</tr>
<tr>
<td>20</td>
<td>Korean</td>
<td>ko</td>
<td>182,327</td>
<td>10,241</td>
</tr>
</tbody>
</table>

List of Wikipedia languages ranked on number of articles in that language.
The Reality Of Wikipedia For Many

<table>
<thead>
<tr>
<th>No</th>
<th>Language</th>
<th>Wiki</th>
<th>Articles</th>
<th>Images</th>
</tr>
</thead>
<tbody>
<tr>
<td>276</td>
<td>Marshallese</td>
<td>mh</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>277</td>
<td>Afar</td>
<td>aa</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>278</td>
<td>Kuanyama</td>
<td>kj</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>279</td>
<td>Hiri Motu</td>
<td>ho</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>280</td>
<td>Muscogee</td>
<td>mus</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>281</td>
<td>Kanuri</td>
<td>kr</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

List of Wikipedia languages ranked on number of articles in that language

“Kanuri is a dialect continuum spoken by some four million people, as of 1987, in Nigeria, Niger, Chad and Cameroon, as well as small minorities in southern Libya and by a diaspora in Sudan.”
http://en.wikipedia.org/wiki/Kanuri_language
Wylam does not appear to exist if you speak French (or Chinese, Japanese, Korean, or ...)!
Problem 1

• There are many languages that have insufficient topical coverage in Wikipedia.

• We believe that it is too restrictive to only have same-language links in Wikipedia, especially if the reader is multi-lingual.
  • “Most first-language speakers speak Hausa or Arabic as a second language”

http://en.wikipedia.org/wiki/Kanuri_language
Our View Of Wikipedia

- Wikipedia articles exit in multiple languages
The Reality Of Wikipedia For Many

- Different articles are written by different sets of authors and are not necessarily the same.
E.G. Wylam

English

Polish
Problem 2

- There are many articles that have different coverage in the different language versions of Wikipedia.

- We believe that it is too restrictive to only have same-language links in Wikipedia, especially if the reader is multi-lingual.
Our View Of Wikipedia

• Cross-language links address these problems
• Such links already exist in Wikipedia:

Chinese article “諾森伯蘭郡” links to the English article “List of MPs elected in the United Kingdom general election, 2005”. The page also exists in many languages including English as “Northumberland”.
The Reality Of Wikipedia For Many

- Links are largely same-language
- Not all cross-language equivalent links exist
  - The English “Custard” is not linked to Italian “Crema pasticcera” (and vice versa)
- Cross-language links are not always correct
  - Chinese “奶黄” is incorrectly linked to Italian “Budino” (and vice versa)
  - It should go to “custard”
E.G. Custard
Research Question

• *Can we build systems that automatically recommend correct cross language links (anchors and targets)?*

• We proposed this as a task and ran a pilot at NTCIR-9 (this will run again at NTCIR-10)

• This is an extension of the Link-the-Wiki track that ran in English at INEX (which is now finished)
CrossLink Task at NTCIR

• Task
  • Given English and a CJK Wikipedia, propose links from English into one of the other collections

• That is:
  • Choose anchors in English documents
  • Choose target documents in one of the other languages
    • Three tasks in total (Chinese, Japanese, Korean)
Document Collection

- Four language versions of Wikipedia

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Articles</th>
<th>Pre-existing Cross-lingual links</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>3,484,250</td>
<td>169,974 (en→zh, 4.9%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>292,548 (en→ja, 8.4%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>87,367 (en→ko, 2.5%)</td>
</tr>
<tr>
<td>Chinese</td>
<td>316,251</td>
<td>170,637 (zh→en, 54.0%)</td>
</tr>
<tr>
<td>Japanese</td>
<td>715,911</td>
<td>289,579 (ja→en, 40.4%)</td>
</tr>
<tr>
<td>Korean</td>
<td>201,512</td>
<td>89,230 (ko→en, 44.3%)</td>
</tr>
</tbody>
</table>
Topics

• Topics were 25 documents chosen at random from the English Wikipedia collection

• 4 sub-tasks
  • en→zh (English to Chinese)
  • en→ja (English to Japanese)
  • en→ko (English to Korean)

• Runs:
  • 250 links per document, 5 targets per link
    • Multi-target linking
Algorithms

• See NTCIR session 5
  • December 8th at 2pm
Runs

- 11 groups participated
- 57 runs were submitted
- Runs were submitted for all tasks
- English to Chinese was the most popular task

<table>
<thead>
<tr>
<th>Task</th>
<th>Runs</th>
<th>Mean links/topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>en→zh</td>
<td>25</td>
<td>2969</td>
</tr>
<tr>
<td>en→ja</td>
<td>11</td>
<td>666</td>
</tr>
<tr>
<td>en→ko</td>
<td>21</td>
<td>924</td>
</tr>
</tbody>
</table>
Assessment Methods

• Automatic Assessment
  • File to File (F2F) assessment ("see also" links)
  • Derived from the Wikipedia itself

• Manual Assessment
  • Anchor to file (A2F) assessment ("inline" links)
  • Human decisions on the links in the runs
F2F: Assessment

- Ground-truth (qrels) derived from links already in Wikipedia articles through triangulation
  - The mono-lingual links from the translation of the source article
  - The cross-lingual page of the mono-lingual links from the source article

- E.g. English article “Martial Arts”
  - Relevant Chinese links are those links out of the Chinese “Martial Arts” (武術) article, and the Chinese counterpart for all links out of the English “Martial Arts” article
F2F: Assessment

Martial Arts

Self-Defence

Combat

武術

肢 (Limb (anatomy))

冷兵器 (Cold weapon)

Chinese of links from “Martial Arts”

Links from “武術”
A2F: Assessment

• Pooled the runs
  • Some anchors (from different runs) overlapped
  • They could be judged as separate anchors or one long anchor (the assessor decided)
    • Which is better “George Stephenson” or “Stephenson”

• Manually assess each anchor in each document using a custom-built assessment tool
A2F: Assessors

- QUT students and staff
  - en→zh: Difficult to recruit
    - 3 topics were not assessed
  - en→ko: Easy to recruit
  - en→ja: Done by Kelly
- All were compensated with cinema tickets

<table>
<thead>
<tr>
<th>Task</th>
<th>Assessors</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>en→zh</td>
<td>15</td>
<td>PhD students, and undergrads</td>
</tr>
<tr>
<td>en→ja</td>
<td>1</td>
<td>Postdoc</td>
</tr>
<tr>
<td>en→ko</td>
<td>5</td>
<td>Undergrads</td>
</tr>
</tbody>
</table>
# A2F: Assessment Tool

**Source document**

**Unassessed anchor**

**Current anchor**

**Target document**

---

### Sushi

28271 3750990162 2010-07-23T20:00:43Z

**Japanese**

*The image of a sushi roll with a salmon slice is shown.*

**Text:**

>Sushi is a Japanese dish consisting of sliced raw fish which is customarily served with other ingredients, such as vinegared rice, seaweed, pickled vegetables, and sometimes fish eggs. It is generally considered to be an acquired taste, as distinct from sushi. Sushi that is served rolled inside a tube of nori (seaweed) is makizushi (巻き寿司). Slices of sushi rice with seaweed scattered over it is called chazuke (茶漬け), which is a type of sushi. Sushi is served with soy sauce, wasabi, and pickled ginger. The sushi is traditionally served cold.

---

### Seaweed

698091 66097982 2008-01-25T07:18:56Z

**Current anchor**

**Text:**

>Seaweed is a type of marine algae that is used as a food source in many cultures around the world. It is often used in sushi rolls and as a garnish. Seaweed is high in nutrients, including vitamins and minerals.

---

Right click irrelevant

Left click relevant
Assessments

• Many thousands of relevant (and non-relevant) links were assessed

<table>
<thead>
<tr>
<th>Assessment set</th>
<th>Relevant links</th>
<th>Overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>en→zh automatic</td>
<td>2,116</td>
<td>1134</td>
</tr>
<tr>
<td>en→zh manual</td>
<td>4,309</td>
<td></td>
</tr>
<tr>
<td>en→ja automatic</td>
<td>2,939</td>
<td>781</td>
</tr>
<tr>
<td>en→ja manual</td>
<td>1,118</td>
<td></td>
</tr>
<tr>
<td>en→ko automatic</td>
<td>1,681</td>
<td>821</td>
</tr>
<tr>
<td>en→ko manual</td>
<td>2,786</td>
<td></td>
</tr>
</tbody>
</table>

• Note the overlap, new links were found
• Next year we’ll assess the automatic pool
  • At INEX this found many non-relevant links!
Evaluation

• Evaluation was with standard IR metrics adapted to link-discovery

• MAP, R-PREC and P@n

• Will only present some en→zh result here
F2F: Precision & Recall

\[
\text{Precision}_{f2f} = \frac{\text{Found & Relevant}}{\text{Found}}
\]

\[
\text{Recall}_{f2f} = \frac{\text{Found & Relevant}}{\text{Relevant}}
\]

• Nothing unexpected here!
A2F: Precision & Recall

\[ f_{\text{anchor}}(i) = \begin{cases} 1, & \text{if relevant with } \geq 1 \text{ relevant targets} \\ 0, & \text{otherwise} \end{cases} \]

An anchor is relevant if one or more of its targets is relevant

\[ f_{\text{link}}(j) = \begin{cases} 1, & \text{if relevant} \\ 0, & \text{otherwise} \end{cases} \]

A target is relevant if the assessor assessed it as relevant

\[
\text{Precision}_{a2f} = \left( \sum_{i=1}^{n} (f_{\text{anchor}}(i)) \times \frac{\sum_{j=1}^{k_i} f_{\text{link}}(j)}{k_i} \right) / n
\]

Precision of an article is mean of the anchor-target precisions

\[
\text{Recall}_{a2f} = \left( \sum_{i=1}^{n} (f_{\text{anchor}}(i)) \times \frac{\sum_{j=1}^{k_i} f_{\text{link}}(j)}{k_i} \right) / N
\]

And likewise for recall.
Evaluation Metrics

\[ MAP = \left( \frac{\sum_{t=1}^{n} \sum_{k=1}^{m} p_{kt}}{m} \right) / n \]  
That is, MAP as usual

\[ RPrec = \sum_{t=1}^{n} \frac{P_t @ R}{n} \]  
That is, RPREC as usual

**Precision-at-N**

\[ N = 5, 10, 20, 30, 50, 250 \]
Evaluation Results MAP

- Full details in NTCIR track overview paper
- Note, however, different rank order and MAP scores

<table>
<thead>
<tr>
<th></th>
<th>F2F (Automatic)</th>
<th></th>
<th>A2F (Manual)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Participant</td>
<td>MAP</td>
<td>Participant</td>
</tr>
<tr>
<td>HITS</td>
<td>0.373</td>
<td></td>
<td>UKP</td>
</tr>
<tr>
<td>UKP</td>
<td>0.314</td>
<td></td>
<td>QUT</td>
</tr>
<tr>
<td>KMI</td>
<td>0.260</td>
<td></td>
<td>HITS</td>
</tr>
<tr>
<td>IASL</td>
<td>0.225</td>
<td></td>
<td>KMI</td>
</tr>
<tr>
<td>QUT</td>
<td>0.179</td>
<td></td>
<td>IASL</td>
</tr>
<tr>
<td>WUST</td>
<td>0.108</td>
<td></td>
<td>WUST</td>
</tr>
<tr>
<td>ISTIC</td>
<td>0.032</td>
<td></td>
<td>ISTIC</td>
</tr>
</tbody>
</table>

Automatic and Manual MAP for en→zh
F2F: Results Precision / Recall
A2F: Results Precision / Recall
Unique Relevant Links

- Some systems were good at finding relevant links but not ranking them

<table>
<thead>
<tr>
<th>Assessment</th>
<th>Total (%)</th>
<th>Team</th>
<th>Rel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automatic</td>
<td>245 (11.6%)</td>
<td>UKP</td>
<td>97</td>
</tr>
<tr>
<td>Manual</td>
<td>1397 (32.4%)</td>
<td>QUT</td>
<td>1103</td>
</tr>
</tbody>
</table>

Unique Relevant en→zh Links
Cross Language Agreement

- Two groups (HITS & UKP) submitted runs to all three (CJK) tasks
- These groups consistently performed well regardless of language
- Their algorithms are language independent!
- So, which task was “easiest”? 
Cross-language Agreement (Manual)

Performance of HITS (1-3) and UKP (4-8), manual F2F

Japanese is easier than Chinese
Cross-language Agreement (Automatic)

Performance of HITS (1-3) and UKP (4-8), Automatic F2F

Korean is easier than Chinese than Japanese
The Effectiveness of CLLD

• Effectiveness of CLLD is at the same level as the first year INEX ran a Link Discovery track.

• We’re more effective at copying what’s there than suggesting new links.

• Systems are either effective at recommending new links or ranking old ones, not both.

• More effective in “easier” languages.
Questions?