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

No	Language	Wiki	Articles	Images
1	English	en	3,807,882	825,432
9	Japanese	ja	779,656	77,107
12	Chinese	zh	386,596	27,175
20	Korean	ko	182,327	10,241

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

The Reality Of Wikipedia For Many

No	Language	Wiki	Articles	Images
276	Marshallese	mh	10	2
277	Afar	aa	6	0
278	Kuanyama	kj	5	0
279	Hiri Motu	ho	3	0
280	Muscogee	mus	2	0
281	Kanuri	kr	1	0

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

E.G. Wylam

Languages
Deutsch
Italiano
Nederlands
Polski

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

English

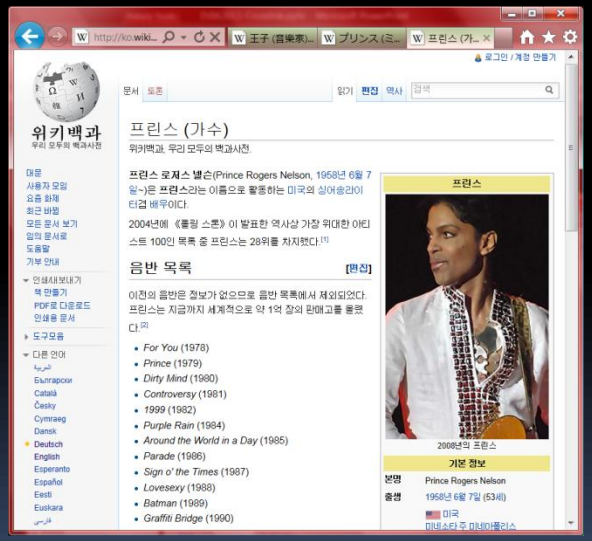
German

The Reality Of Wikipedia For Many

- Different articles are written by different sets of authors and are not necessarily the same

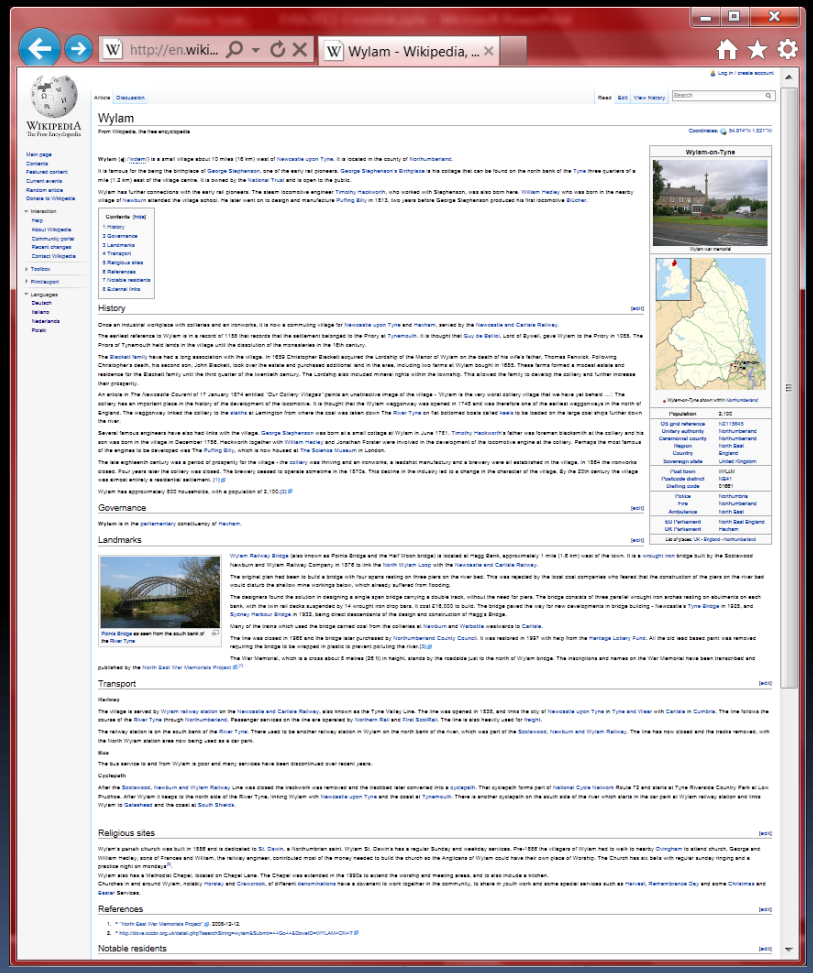
Chinese

Japanese



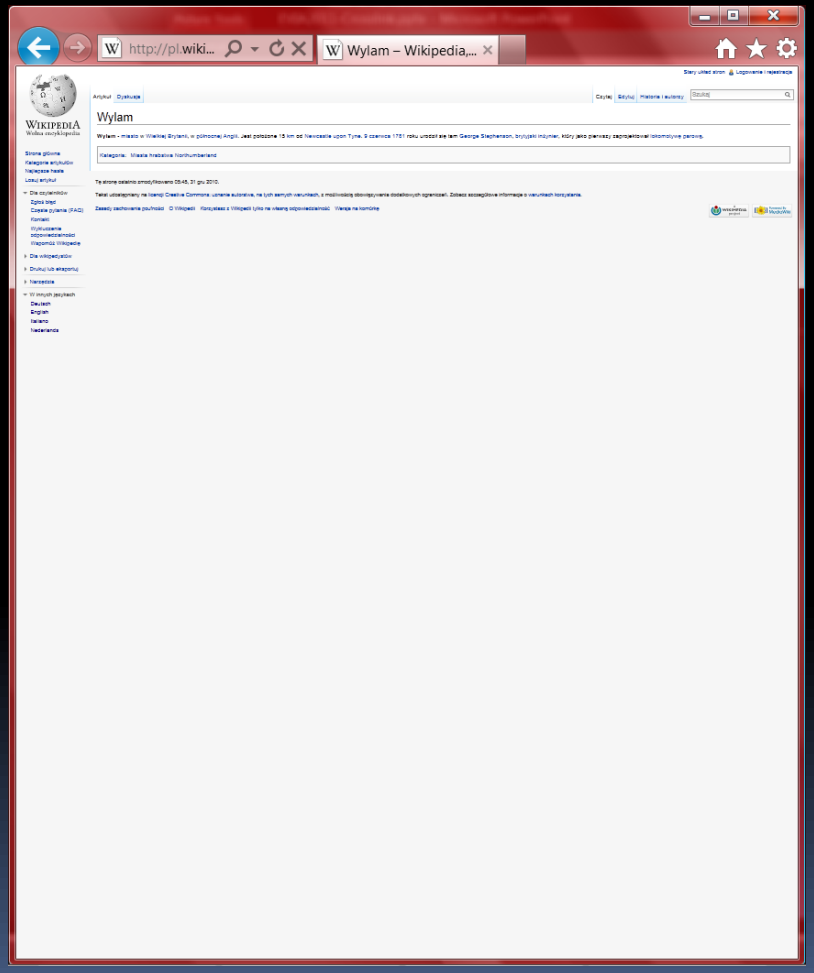
Korea

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

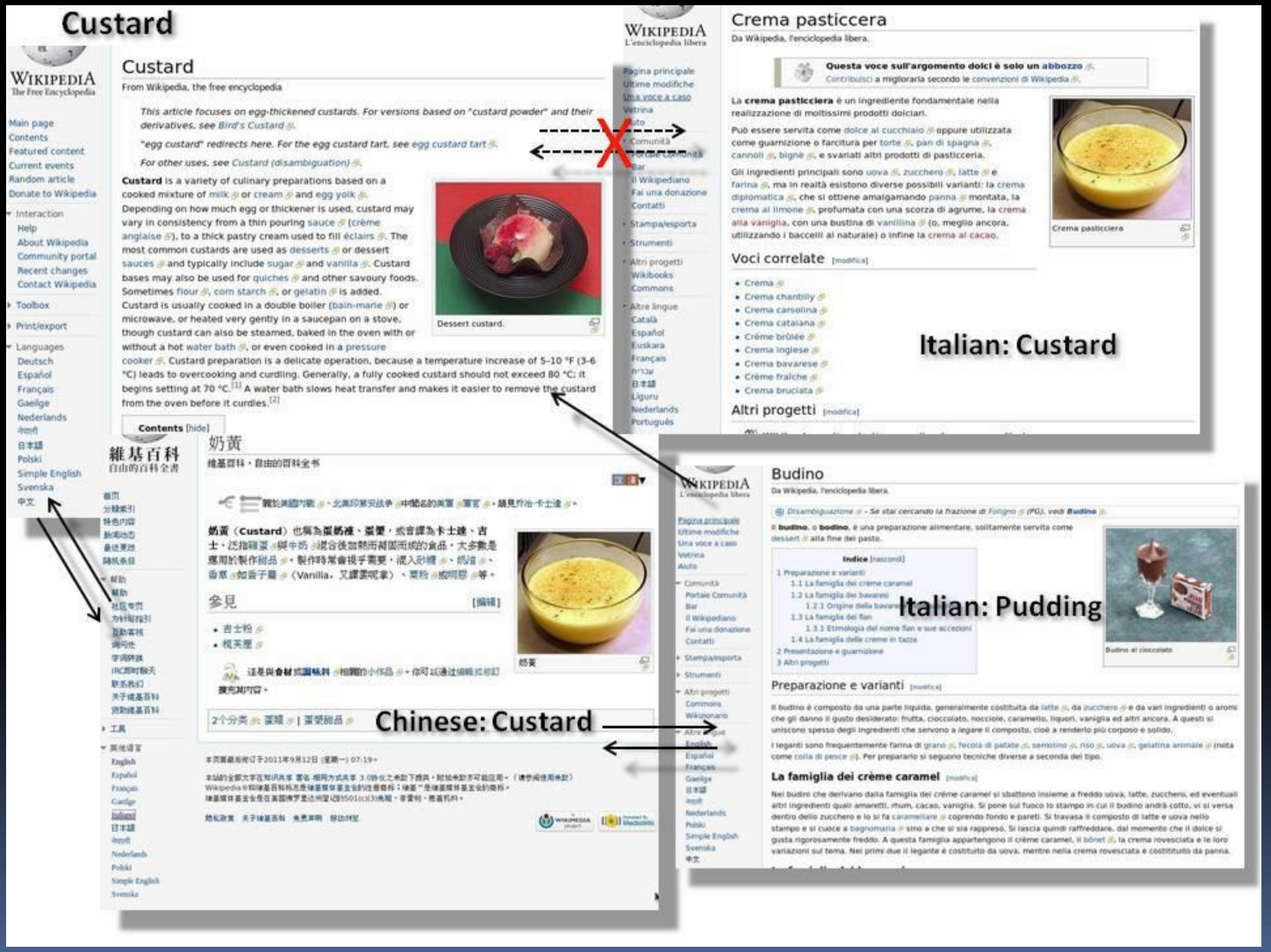
- Cross-language links address these problems
  - Such links already exist in Wikipedia:

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


Custard
Custard
From Wikipedia, the free encyclopedia
Crema pasticcera
Da Wikipedia, l'enciclopedia libera.
Italian: Custard
奶黃
Chinese: Custard
Budino
Italian: Pudding

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

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

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

Task	Runs	Mean links/topic
en→zh	25	2969
en→ja	11	666
en→ko	21	924

Assessment Methods

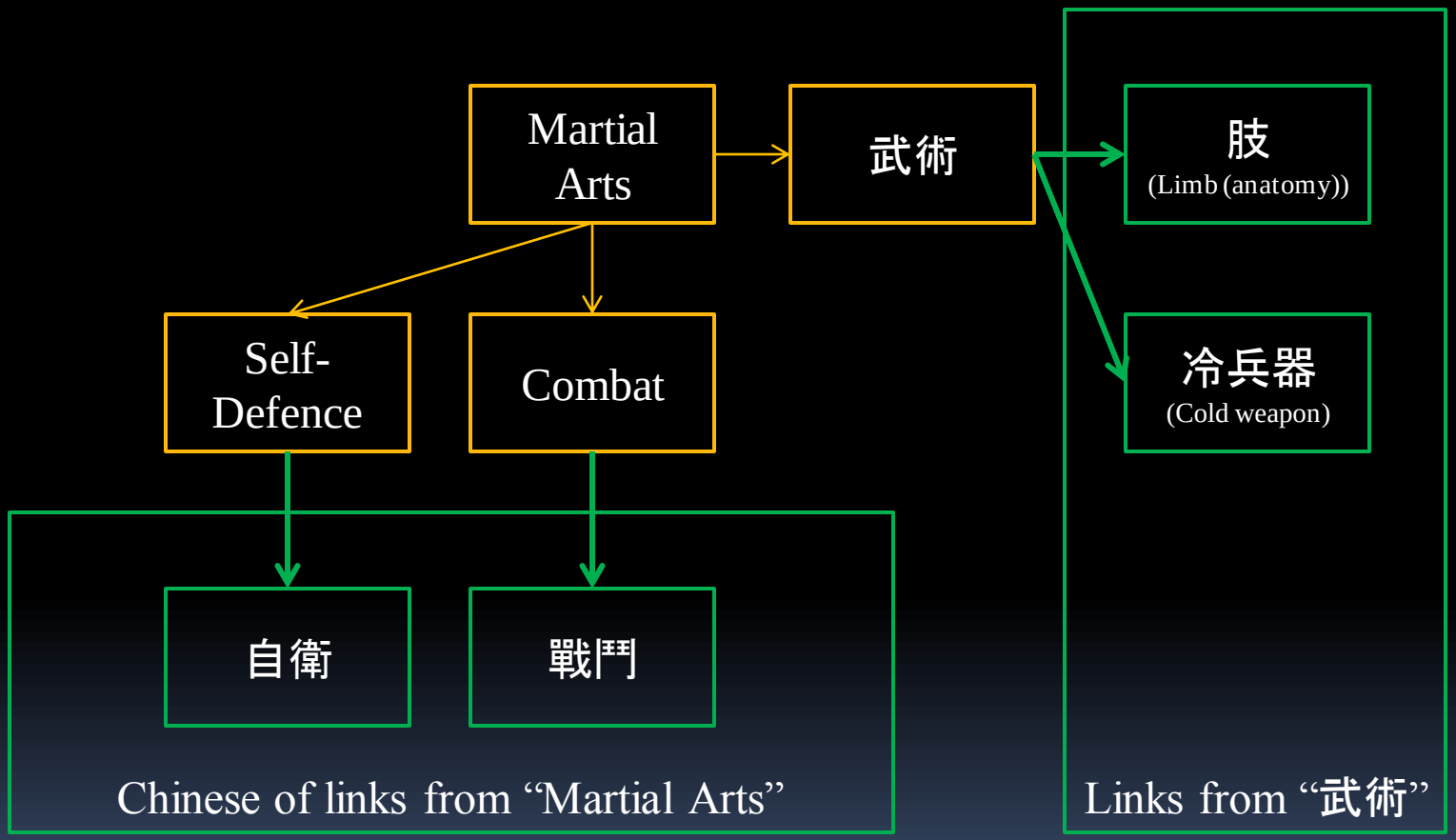
- Automatic Assessment
  - File to File (F2F) assessment (“see also” links)
  - Derived from the Wikipedia itself
- Manual Assessment
  - Anchor to file (A2F) assessment (“inline” links)
  - Human decisions on the links in the runs

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



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

Task	Assessors	Description
en→zh	15	PhD students, and undergrads
en→ja	1	Postdoc
en→ko	5	Undergrads

A2F: Assessment Tool

Source document

Unassessed anchor

Current anchor

Target document

Right click irrelevant
Left click relevant

Assessments

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

Assessment set	Relevant links	Overlap
en→zh automatic	2,116	1134
en→zh manual	4,309	
en→ja automatic	2,939	781
en→ja manual	1,118	
en→ko automatic	1,681	821
en→ko manual	2,786	

- 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

$$Precision_{f2f} = \frac{\text{Found \& Relevant}}{\text{Found}}$$

$$Recall_{f2f} = \frac{\text{Found \& Relevant}}{\text{Relevant}}$$

- Nothing unexpected here!

A2F: Precision & Recall

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

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

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

A *target* is relevant if the assessor assessed it as relevant

$$Precision_{a2f} = \left(\sum_{i=1}^{n} \left(f_{anchor}(i) \right) \times \frac{\sum_{j=1}^{k_i} f_{link}(j)}{k_i} \right) / n$$

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

$$Recall_{a2f} = \left(\sum_{i=1}^{n} \left(f_{anchor}(i) \right) \times \frac{\sum_{j=1}^{k_i} f_{link}(j)}{k_i} \right) / N$$

And likewise for recall

Evaluation Metrics

$$MAP = (\sum_{t=1}^{n} \frac{\sum_{k=1}^{m} p_{kt}}{m})/n$$

That is, MAP as usual

$$RPrec = \sum_{t=1}^{n} P_t\,@\,R\,/n$$

That is, RPREC as usual

Precision-at-N

N = 5, 10, 20, 30, 50, 250

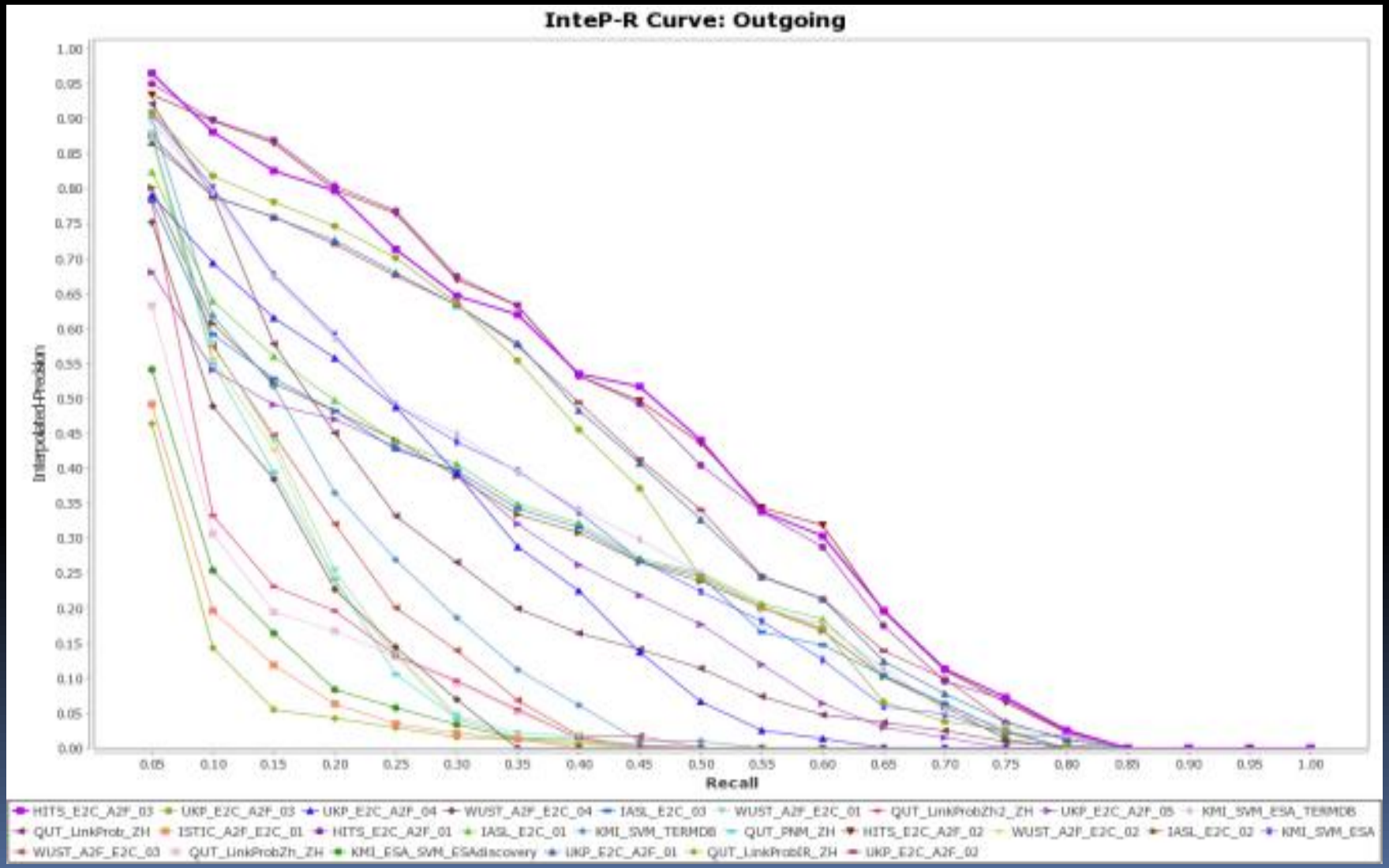
Evaluation Results MAP

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

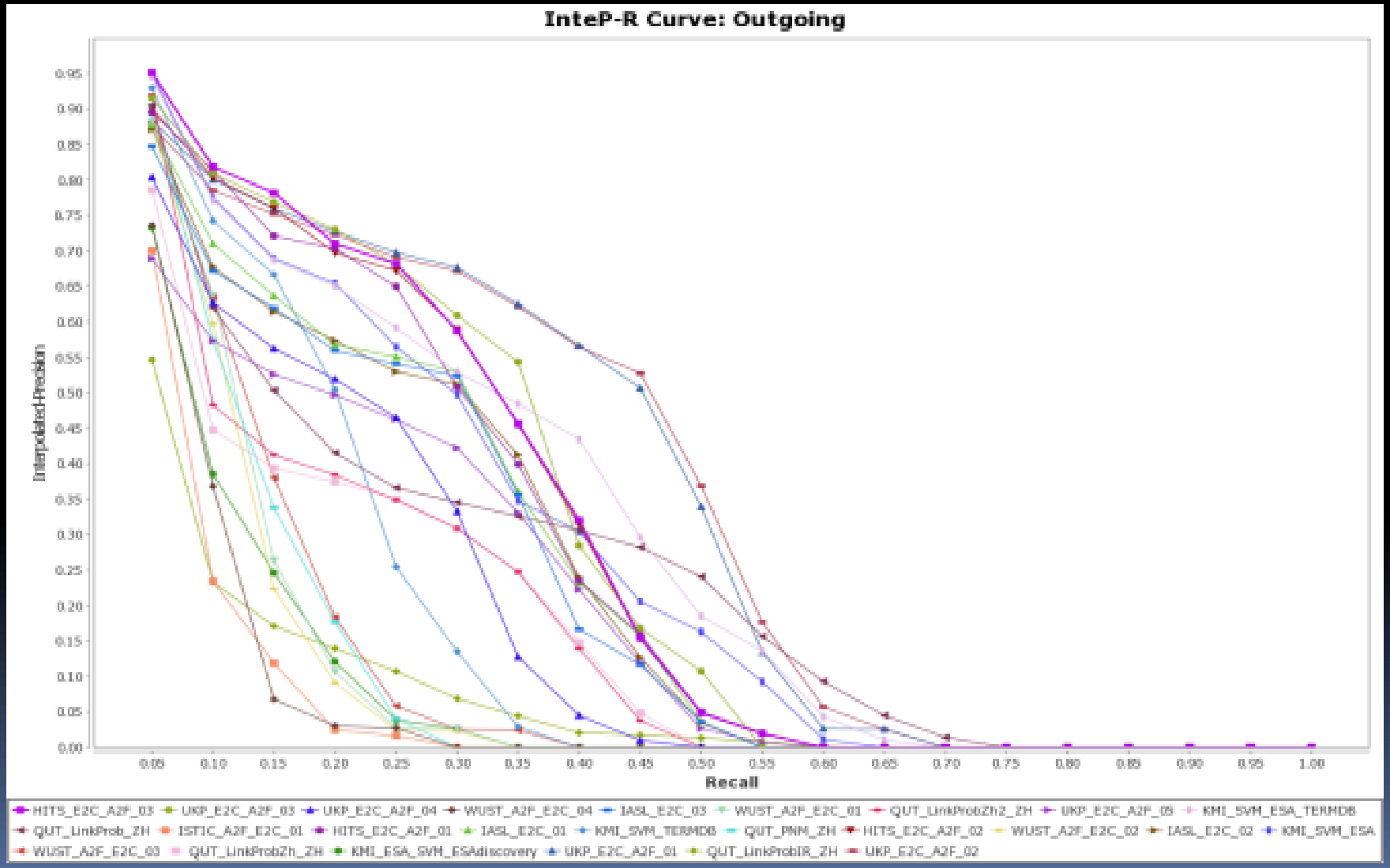
F2F (Automatic)		A2F (Manual)	
Participant	MAP	Participant	MAP
HITS	0.373	UKP	0.157
UKP	0.314	QUT	0.115
KMI	0.260	HITS	0.102
IASL	0.225	KMI	0.097
QUT	0.179	IASL	0.037
WUST	0.108	WUST	0.012
ISTIC	0.032	ISTIC	0.000

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

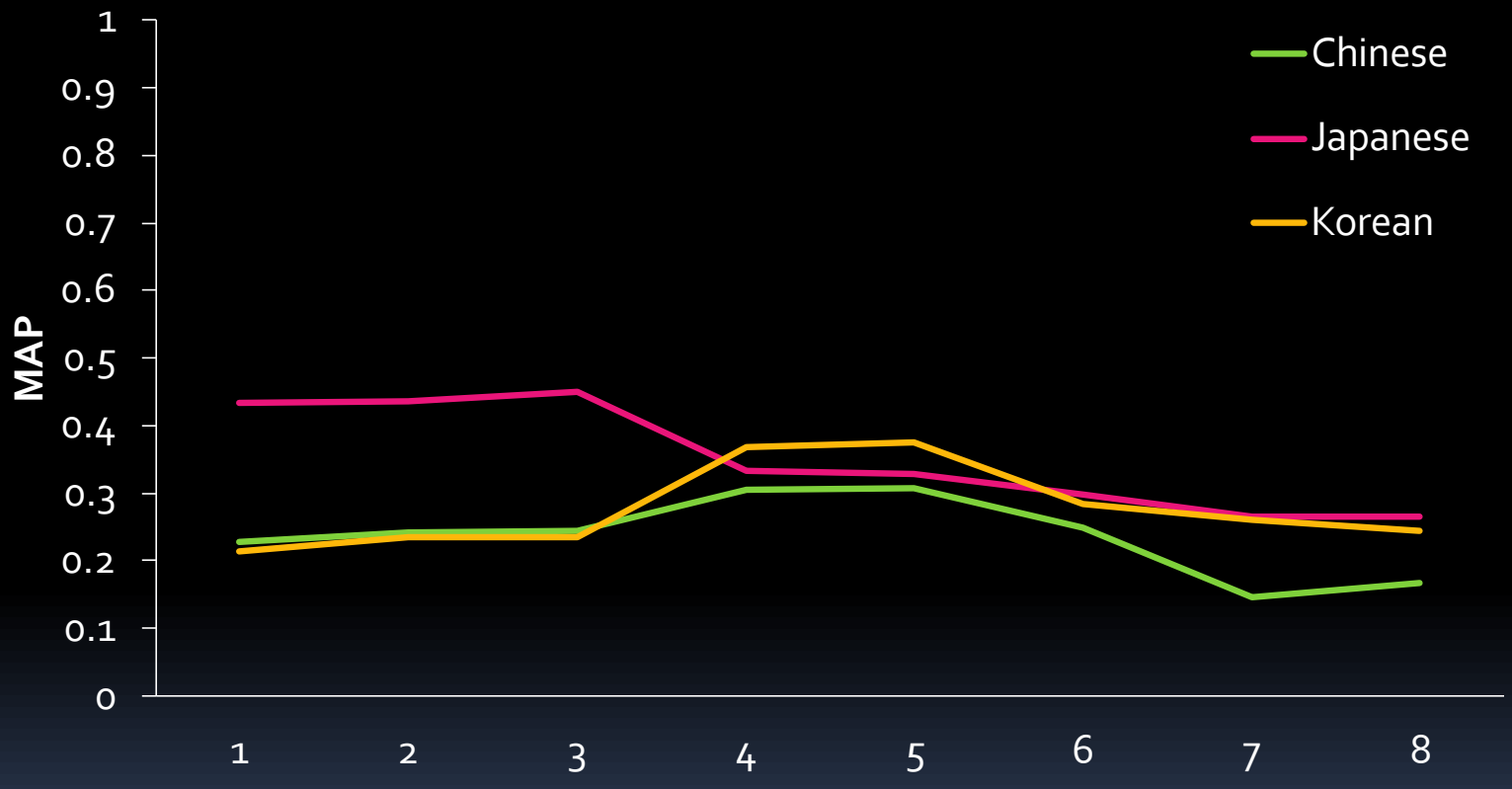
Assessment	Total (%)	Team	Rel
Automatic	245 (11.6%)	UKP	97
Manual	1397 (32.4%)	QUT	1103

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”?

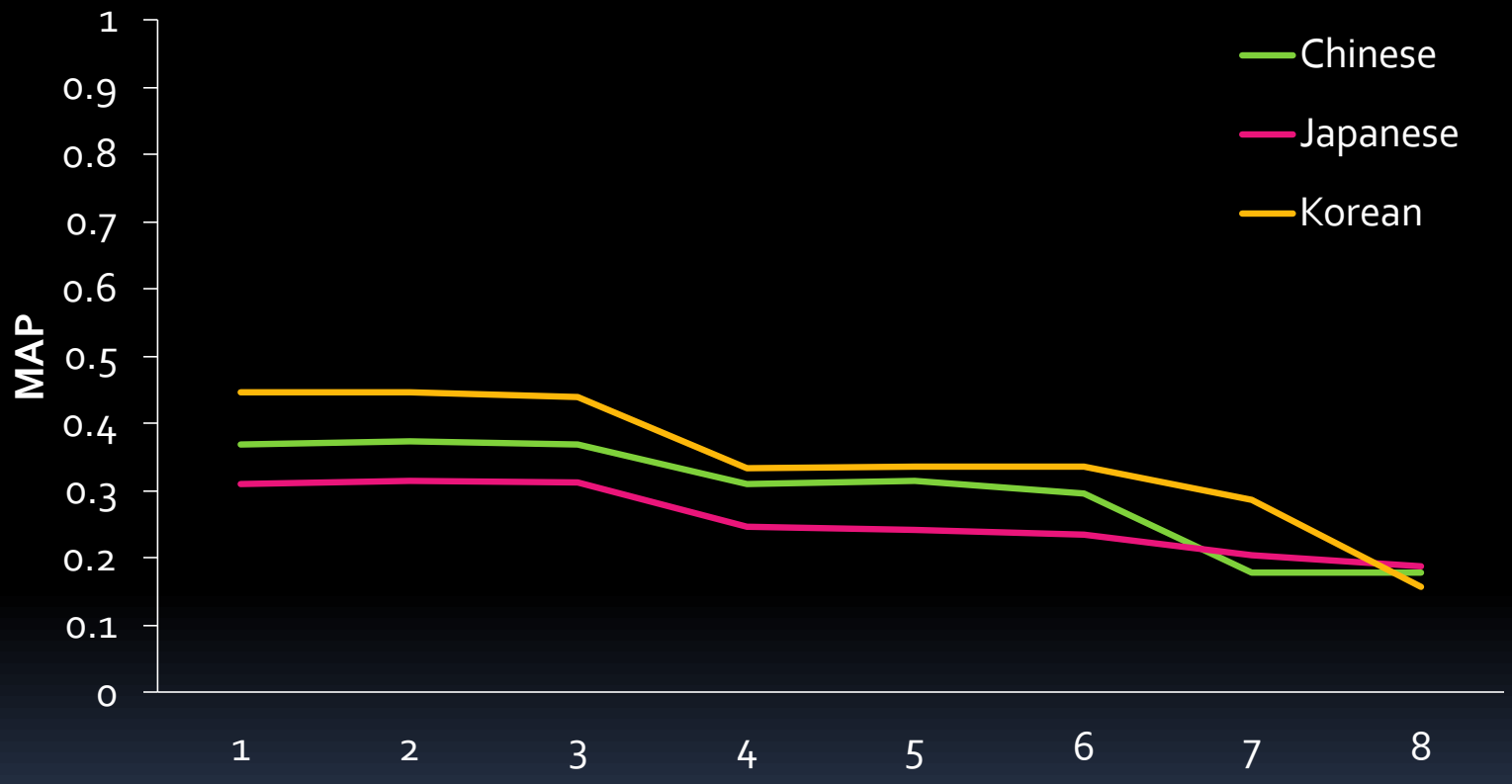
Cross-language Agreement (Manual)



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), AutomaticF2F

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?