

# The Effectiveness Of Cross-lingual Link Discovery

Ling-Xiang (Eric) Tang<sup>†</sup>

Kelly Itakura<sup>†</sup>

Shlomo Geva<sup>†</sup>

Andrew Trotman<sup>‡</sup>

Yue Xu<sup>†</sup>

<sup>†</sup> QUT (Brisbane, Australia)

<sup>‡</sup> 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](http://en.wikipedia.org/wiki/Kanuri_language)

# E.G. Wylam

The screenshot shows the Wikipedia page for 'Wylam'. The page is in English. The left sidebar contains a 'Languages' section with links for Deutsch, Italiano, Nederlands, and Polski. The main content area includes the article title 'Wylam', a brief description, and a detailed paragraph about its history and connection to George Stephenson. A 'Contents' table of contents is visible, listing sections from History to External links. On the right, there is a map of Wylam-on-Tyne and a photo of the Wylam war memorial. The page also features a search bar and a 'Log in / create account' link in the top right corner.

Wylam

From Wikipedia, the free encyclopedia

Coordinates: 54°97′N 1°82′W

**Wylam** (<sup>i</sup>/ˈwɪləm/) is a small village about 10 miles (16 km) west of [Newcastle upon Tyne](#). It is located in the county of [Northumberland](#).

It is famous for the being the birthplace of [George Stephenson](#), one of the early rail pioneers. [George Stephenson's Birthplace](#) is his cottage that can be found on the north bank of the [Tyne](#) three quarters of a mile (1.2 km) east of the village centre. It is owned by the [National Trust](#) and is open to the public.

Wylam has further connections with the early rail pioneers. The steam locomotive engineer [Timothy Hackworth](#), who worked with Stephenson, was also born here. [William Hedley](#) who was born in the nearby village of [Newburn](#) attended the village school. He later went on to design and manufacture [Puffing Billy](#) in 1813, two years before George Stephenson produced his first locomotive [Blücher](#).

**Contents** [hide]

- History
- Governance
- Landmarks
- Transport
- Religious sites
- References
- Notable residents
- External links

**History** [edit]

Once an industrial workplace with collieries and an ironworks, it is now a commuting village for [Newcastle upon Tyne](#) and [Harburn](#), served by the [Newcastle and Carlisle Railway](#).

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](http://en.wikipedia.org/wiki/Kanuri_language)

# Our View Of Wikipedia

- Wikipedia articles exist in multiple languages



The screenshot shows the English Wikipedia page for "Prince (musician)". The browser's address bar displays "http://en.wiki...". The page features the Wikipedia logo, a sidebar with navigation links, and the main article content. The article text describes Prince Rogers Nelson, born June 7, 1958, as an American singer, songwriter, musician, and actor. It mentions his production of ten platinum albums and thirty Top 40 singles, his founding of his own recording studio and label, and his role as a "talent promoter" for artists like Sheila E., Carmen Electra, and Sinéad O'Connor. A photo of Prince in 2009 in Paris is included, with a caption "Prince in 2009 in Paris, France". Below the photo is a "Background information" table listing his birth name, aliases, birth date, and genres.

| Background information |  |
|------------------------|--|
| Birth name             | Prince Rogers Nelson   |
| Also known as          | Jamie Starr<br>Christopher<br>Alexander Nevermind<br>Joey Coco |
| Born                   | June 7, 1958 (age 53)<br>Minneapolis, Minnesota, US            |
| Genres                 | Funk, R&B, rock, pop, new wave, Minneapolis sound,             |

English



The screenshot shows the German Wikipedia page for "Prince". The browser's address bar displays "http://de.wiki...". The page features the Wikipedia logo, a sidebar with navigation links, and the main article content. The article text describes Prince Rogers Nelson (\* 7. Juni 1958 in Minneapolis, Minnesota) as an US-amerikanischer Sänger, Komponist, Songwriter, Musikproduzent und Multiinstrumentalist. It mentions his production of ten platinum albums and thirty Top 40 singles, his founding of his own recording studio and label, and his role as a "talent promoter" for artists like Sheila E., Carmen Electra, and Sinéad O'Connor. A photo of Prince in 2009 in Paris is included, with a caption "Prince im Jahr 2009 in Paris". Below the photo is a "Background information" table listing his birth name, aliases, birth date, and genres.

| Background information |  |
|------------------------|--|
| Birth name             | Prince Rogers Nelson   |
| Also known as          | Jamie Starr<br>Christopher<br>Alexander Nevermind<br>Joey Coco |
| Born                   | June 7, 1958 (age 53)<br>Minneapolis, Minnesota, US            |
| Genres                 | Funk, R&B, rock, pop, new wave, Minneapolis sound,             |

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



# English

[illegible]

# Polish

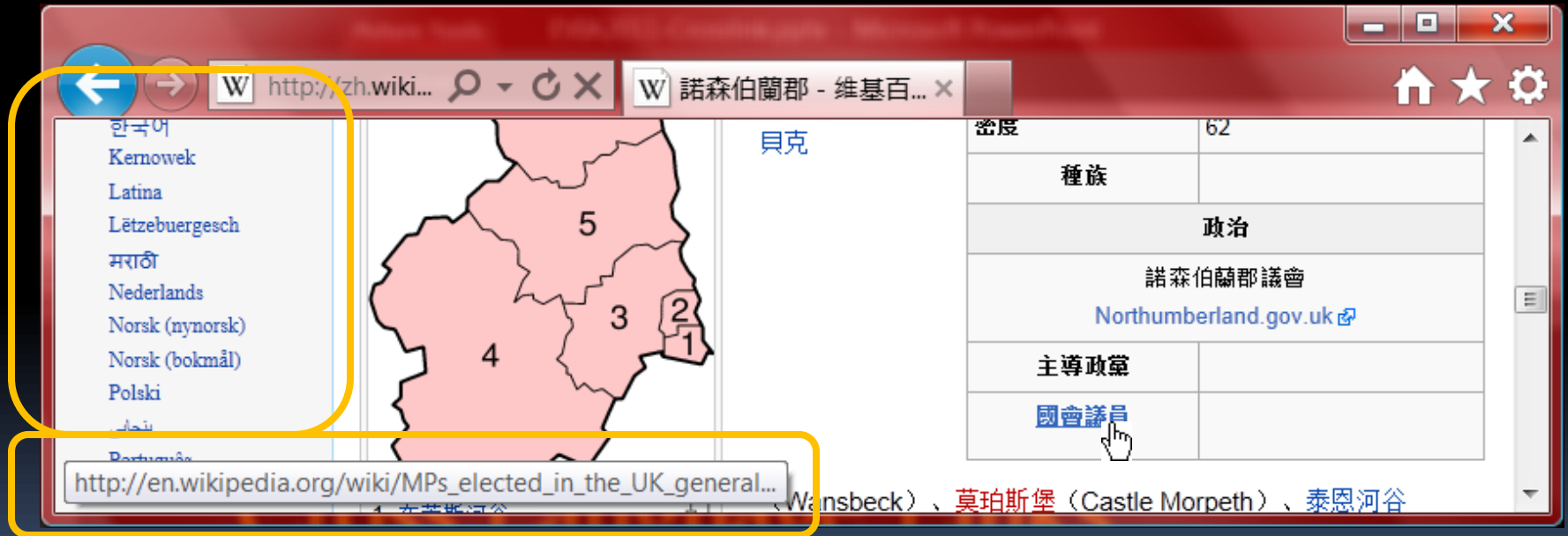
[illegible]

## 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 “奶黃” 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

| Corpus   | Articles  | Pre-existing<br>Cross-lingual links                                    |
|----------|-----------|--|
| English  | 3,484,250 | 169,974 (en→zh, 4.9%)<br>292,548 (en→ja, 8.4%)<br>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 8<sup>th</sup> 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              |

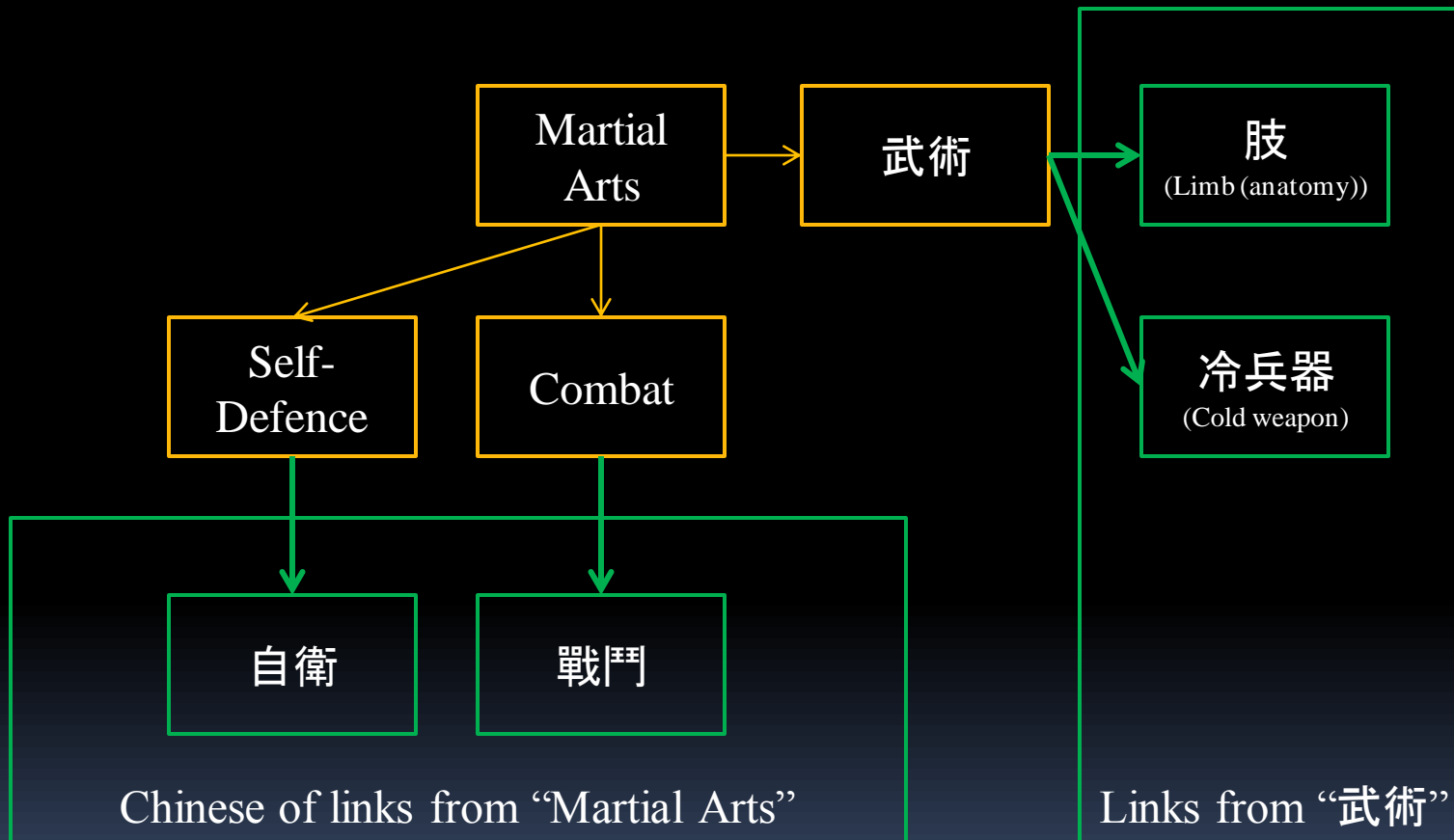
# Assessment Methods

- 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

NTCIR 9 Crosslink: Manual Assessment

File Utility Linking Language Help

Source ID: 28271 Target ID: 698091 Completion: 0 / 2755 Current subanchor: seaweed Belongs to: seaweed

Topic Title: Sushi Target Title: 鋸吻刺刀魚

Source document

Unassessed anchor

Current anchor


Target document

Sushi


28271 375090162 2010-07-23T20:06:43Z Nick Number 1526960

:Sushi

:Japanese cuisine

Image:  Currently Unavailable

Different types of nigiri-zushi ready to be eaten

Image:  Currently Unavailable

Type of Sushi

is a Japanese dish consisting of cooked vinegared rice which is commonly topped with other ingredients, such as fish or other seafood, or put into rolls. Sliced raw fish by itself is called sashimi, as distinct from sushi. Sushi that is served rolled inside or around bread and processed with seaweed (or nori) is makizushi (巻き). Toppings stuffed into a small pouch of fried tofu is inarizushi. A bowl of sushi rice with toppings scattered over it is called chirashi-zushi (ちらし). History

Image:Hiroshige Bowl of Sushi.jpg[thumb|180px|right|Sushi by Hiroshige in Edo period]] The traditional form of sushi is fermented fish and rice, preserved with salt in a process that has been traced to Southeast Asia, where it remains popular today.<sup>2</sup> The term sushi comes from an archaic grammatical form no longer used in other contexts; literally, "sushi" means "it's sour",<sup>3</sup> a reflection of its historic fermented roots.

The science behind the fermentation of fish packed in rice is that the vinegar produced from fermenting rice breaks the fish down into amino acids. This results in one of the five basic tastes, called umami in Japanese.<sup>4</sup> The oldest form of sushi in Japan, Narezushi, still very closely resembles this process. In Japan, Narezushi evolved into Oshizushi and ultimately Edomae nigirizushi, which is what the world today knows as "sushi."

Contemporary Japanese sushi has little resemblance to the traditional lacto-fermented rice dish. Originally, when the fermented fish was taken out of the rice, only the fish was consumed and the fermented rice was discarded.

鋸吻刺刀魚

698091 6097982 2008-01-25T07:18:56Z Alan li 232717

:絨背魚目

鋸吻刺刀魚，又稱藍鰐刺刀魚、漂鰐魚為 輻鰭魚綱 絨背魚目 海龍亞目 溝口魚科的其中一種。分布

本魚分布於 印度、西太平洋、從 東非到 馬紹爾群島、南 日本即 大堡礁，均可見。

深度

水深2公尺~20公尺。

特徵

體側扁，無側線，體色多變化，由褐色至粉紅色或黃色均有，體側布滿許多小黑點及白點；吻延長為扁管狀，無齒，具有一對鬚；第一背鰭無絨，胸鰭小，第二背鰭及尾鰭特大，尾鰭幾乎與軀幹同長，臀鰭圓形。第一背鰭有2枚大型暗色斑，皮膚具有數列星狀突起；吻部背面無硬齒，平直而不彎曲。母魚腹鰭的一部分下緣變形為相接而成的鰾卵袋，體型也比公魚大。體長可達17公分。

本魚體色適其棲地環境而定，多半生活或靠近海藻或藻床，以擬態方式模仿海藻，不易被發現。常成對出現，以吸食方式攝取浮游生物。

經濟利用

多作為海水觀賞魚，不具食用價值。

參考資料

台灣魚類資料

Anchor color legend: Current anchor Not assessed Incomplete Relevant Irrelevant

Previous Next

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 if 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

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

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

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

And likewise for recall

# Evaluation Metrics

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

That is, MAP as usual

$$R\text{Prec} = \sum_{t=1}^n P_t @ R / n$$

That is, RPREC as usual

*Precision-at-N*

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

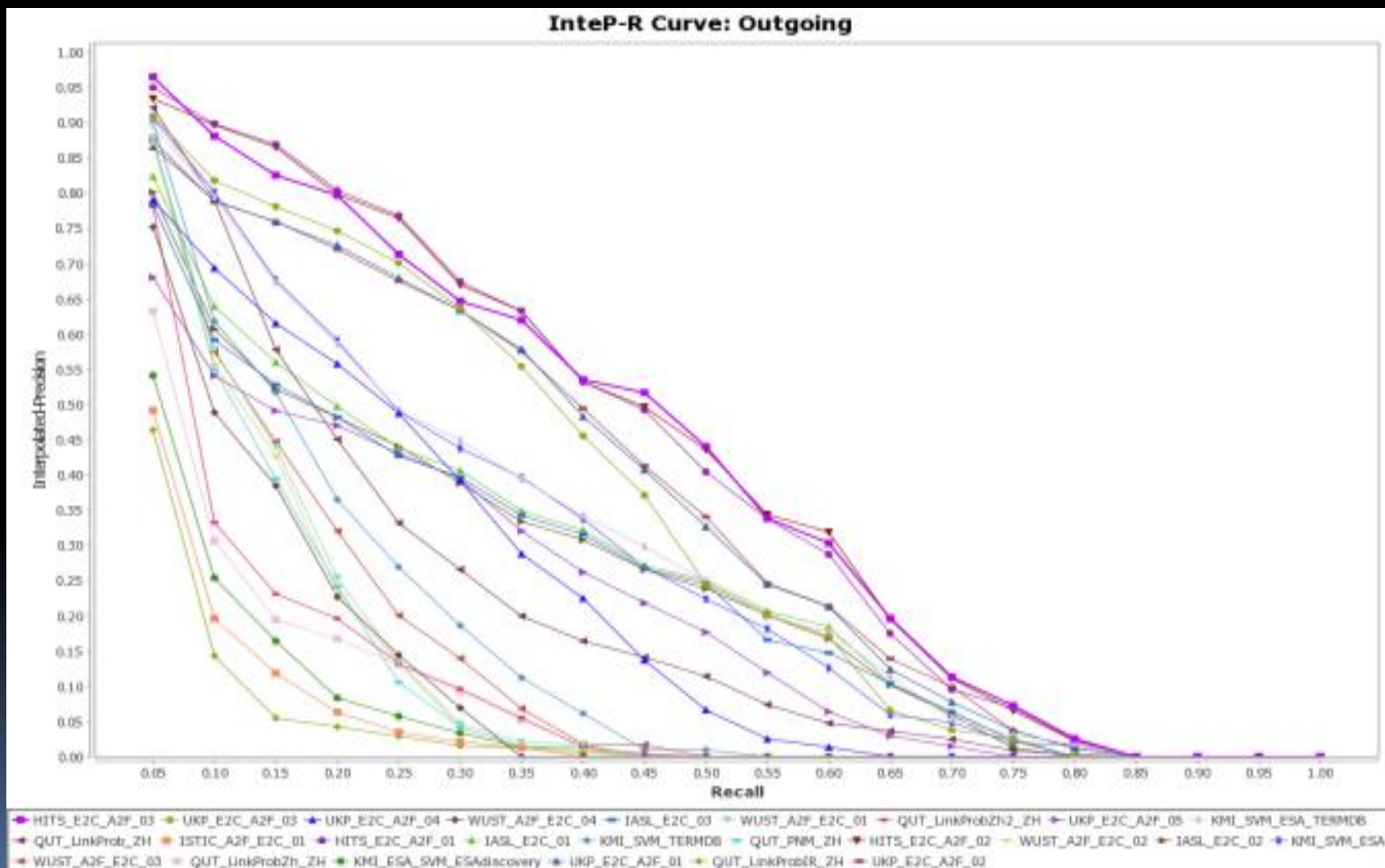
# Evaluation Results MAP

- 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

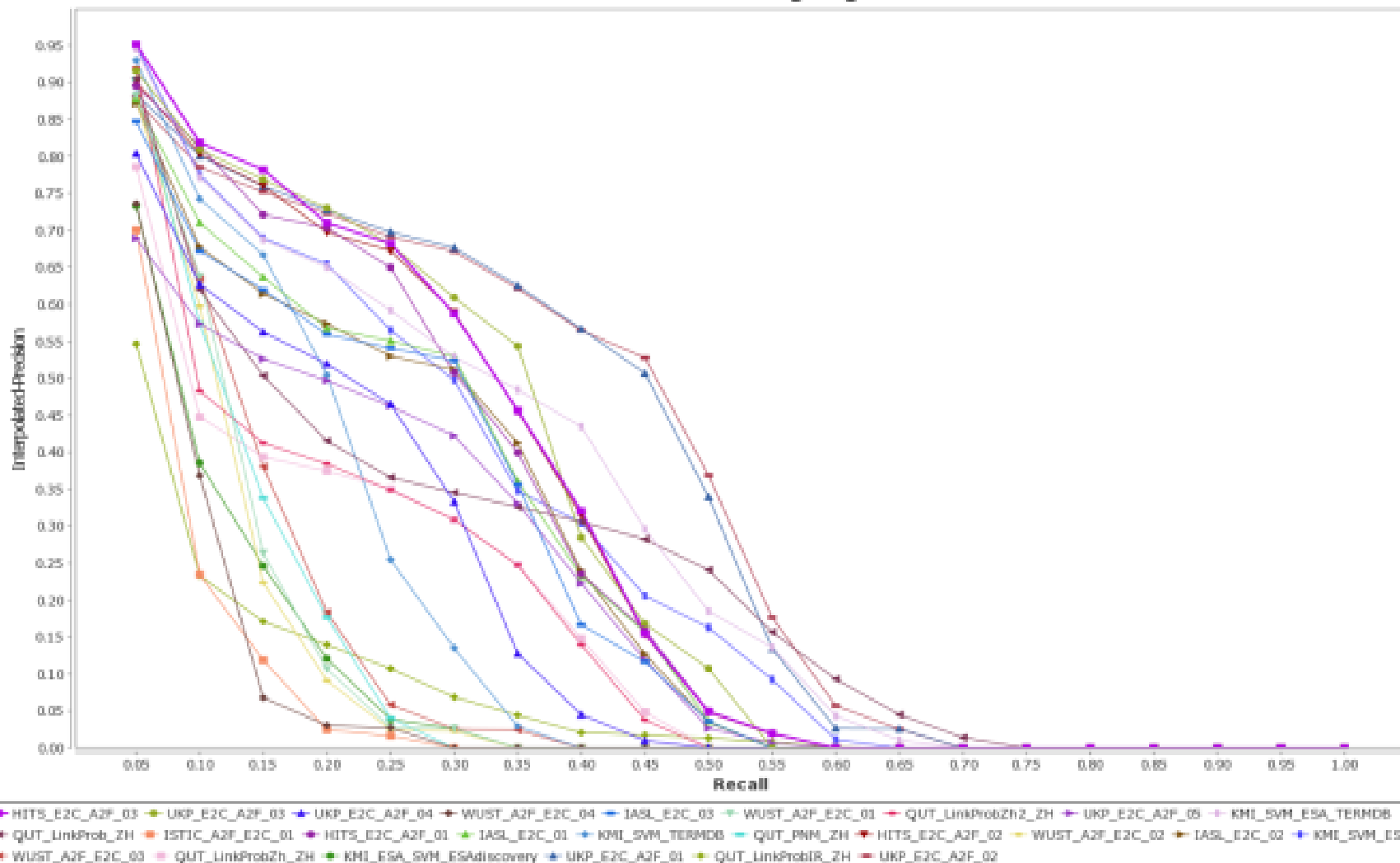
# F2F: Results Precision / Recall





# A2F: Results Precision / Recall

InteP-R Curve: Outgoing



# Unique Relevant Links

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

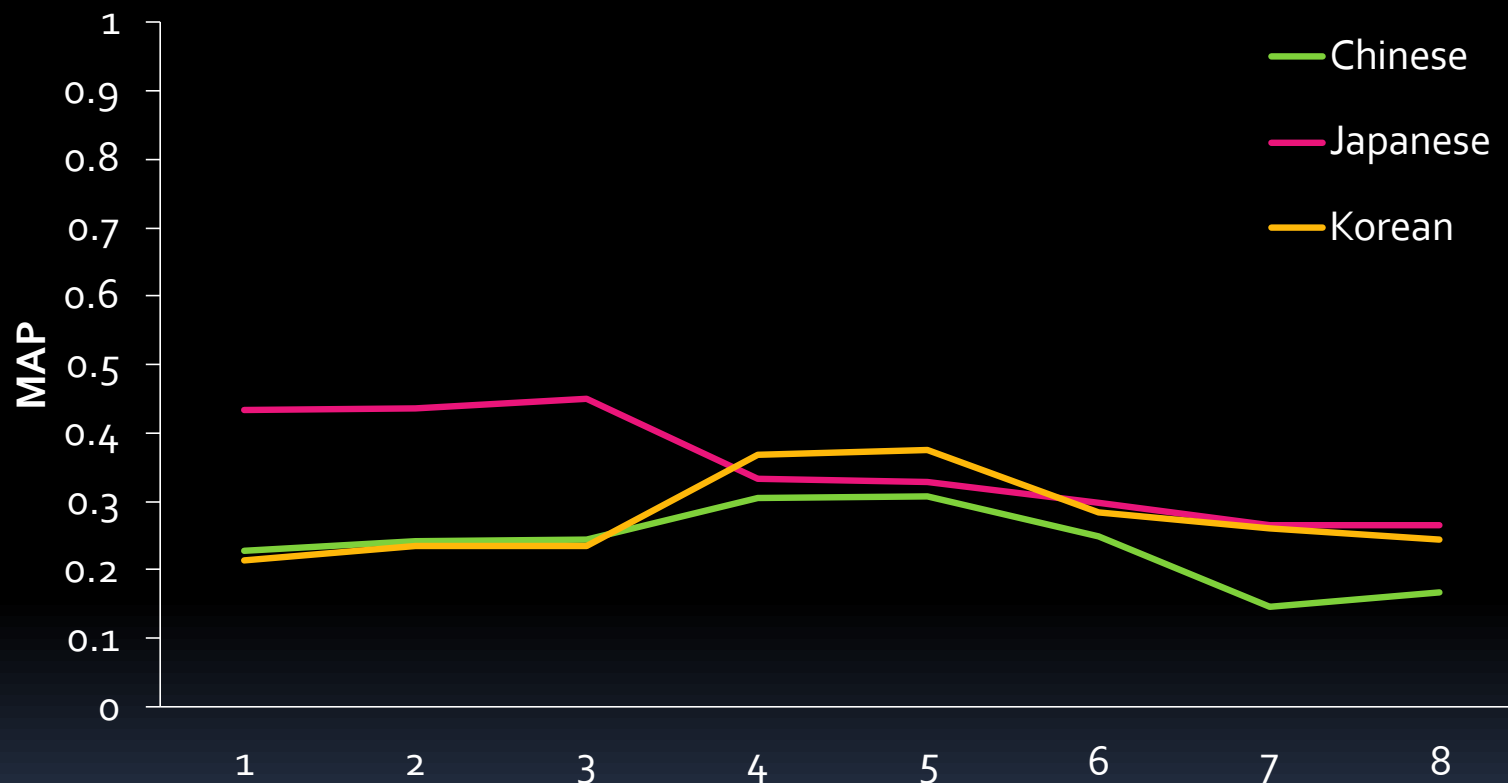
| Assessment | Total (%)    | Team | Rel  |
|------------|--------------|------|------|
| Automatic  | 245 (11.6%)  | UKP  | 97   |
| Manual     | 1397 (32.4%) | QUT  | 1103 |

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”?

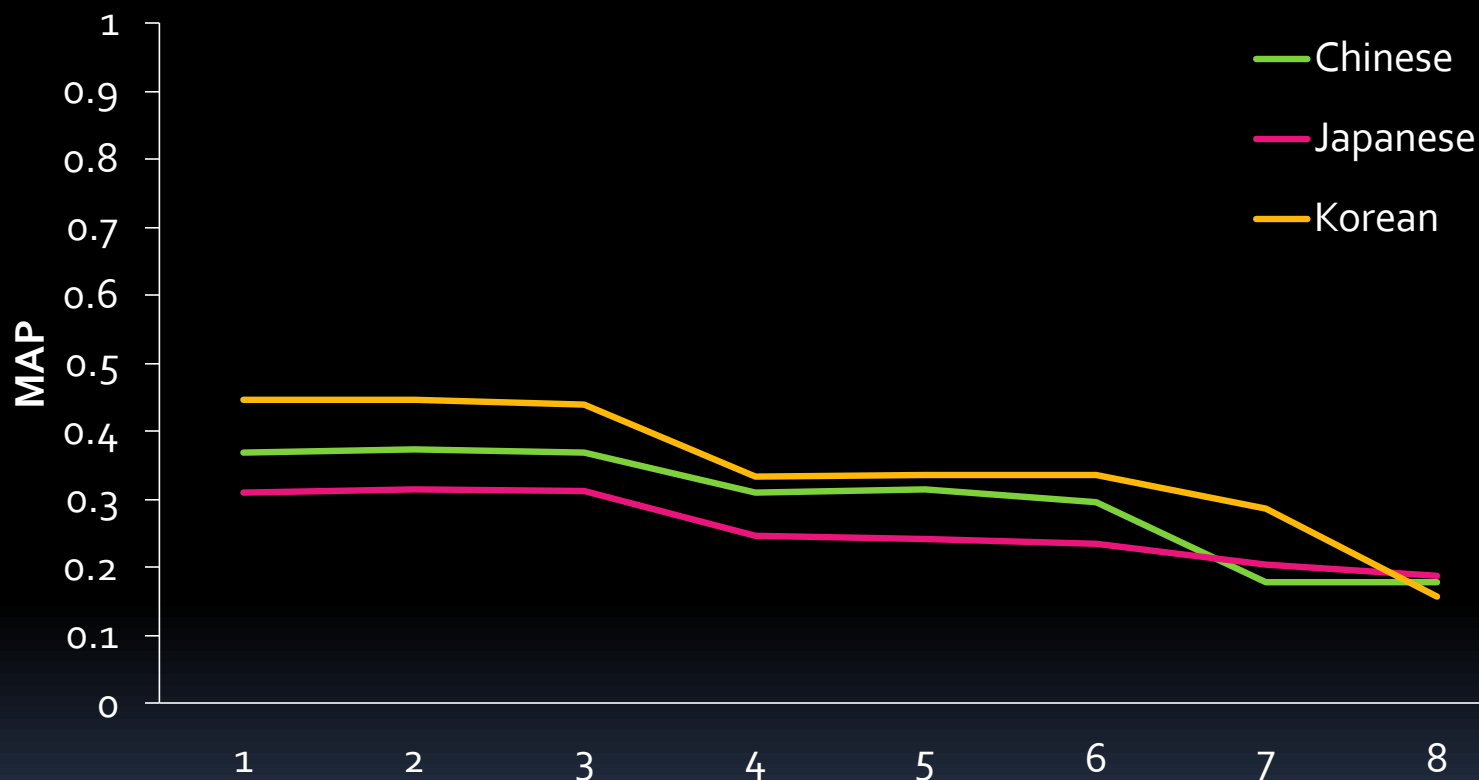
# Cross-language Agreement (Manual)



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?