Overview of the NTCIR-9 Crosslink Task: Cross-lingual Link Discovery

Ling-Xiang Tang¹, Shlomo Geva¹, Andrew Trotman², Yue Xu¹ and Kelly Y. Itakura^{1 1}Faculty of Science and Technology, Queensland University of Technology ²Department of Computer Science, University of Otago

ABSTRACT

This paper presents an overview of NTCIR-9 Cross-lingual Link Discovery (Crosslink) task. The overview includes: the motivation of cross-lingual link discovery; the Crosslink task definition; the run submission specification; the assessment and evaluation framework; the evaluation metrics; and the evaluation results of submitted runs.

1. WHAT IS CROSS-LINGUAL LINK DISCOVERY

Cross-lingual link discovery (CLLD) is a way of automatically finding potential links between documents in different languages. The goal of this task is to create a reusable resource for evaluating automated CLLD approaches. The results of this research can be used in building and refining systems for automated link discovery. The task is focused on linking between English source documents and Chinese, Korean, and Japanese target documents.

CONCLUSION

•The evaluations in both file-to-file and anchor-to-file levels show promising results of participating teams even this is the first year of this task. Particularly, team HITS and UKP achieved very high scores of different evaluation measures (*MAP, R-Prec, Precision-at-N*) in three language subtasks.

•With the standard test collections and evaluation data set developed for this task, systems or applications for realising CLLD can be built or further refined to provide a better automated cross-lingual linking for knowledge management and sharing. With the possible future deployment of techniques and approaches used by the participants, we hope user experience in viewing or editing document in different languages could be enhanced; language is then no longer a barrier for knowledge discovery.

7. EVAULATION RESULTS

InteP-R Curve: Outgoing

2. WHY CLLD?

Wikipedia is an online multilingual encyclopaedia that contains a very large number of articles covering most written languages and so it includes extensive hypertext links between documents of same language for easy reading and referencing. However, the pages in different languages are rarely linked except for the cross-lingual link between pages about the same subject. This could pose serious difficulties to users who try to seek information or knowledge from different lingual sources. The Figure on the right shows a snippet of Martial Art Wikipedia article in which anchors are only linked to related English articles about different types of martial links arts; direct to other related Chinese/Japanese/Korean articles do not exist in Wikipedia.



3. CLLD TASK DEFINITION

CLLD TASK

To submit a run for a given task, participants are required to choose the most suitable anchors in English topic documents, and for each anchor identify the most relevant documents in the target language corpus.

SUBTASKS

By technical focu

- · - · - · - · -

咏春拳

合氣道

柔道

泰拳

- English to Chinese CLLD (E2C)
- English to Korean CLLD (E2K)

THE TOP THREE TEAMS OF EACH MEASURE (MAP, R-PREC, P@N)

TOPICS

A set of 25 articles are randomly chosen from the English Wikipedia and used as test topics. All test topics are prepared in a form of XML file without *link* tags, which means the previously existing links in topics are removed.

WIKIPEDIA TEST COLLECTIONS

Language	# doc	Size	Dump Date
Chinese	318736	2.7G	27/06/2010
Japanese	716,088	6.1G	24/06/2010
Korean	201596	1.2G	28/06/2010

F2F Wiki-GT EVALUATION	F2F MANUL EVALUATION	F2F MANUL EVALUATION
English-2-Chinese	English-2-Chinese	English-2-Chinese
MAP: HITS, UKP, KMI	MAP: UKP, KMI, HITS	MAP: UKP, QUT, HITS
R-Prec: HITS, UKP, KMI	R-Prec: UKP, KMI, HITS	R-Prec: UKP, QUT, KMI
Precision-at-5: HITS, QUT, KMI	Precision-at-5: QUT, HITS, KMI	Precision-at-5: KMI, QUT, UKP
English-2-Japanse	English-2-Japanse	English-2-Japanse
MAP: HITS, UKP, QUT	MAP: HITS, UKP, QUT	MAP: HITS, UKP, QUT
R-Prec: HITS, UKP, QUT	<i>R-Prec:</i> HITS, UKP, QUT	R-Prec: HITS, UKP, QUT
Precision-at-5: HITS, UKP, QUT	Precision-at-5: HITS, UKP, QUT	Precision-at-5: HITS, UKP, QUT
English-2-Korean	English-2-Korean	English-2-Korean
MAP: HITS, DUIIS, UKP	MAP: UKP, DUIIS, HITS	MAP: UKP, HITS, KSLP
R-Prec: HITS, DUIIS, UKP	R-Prec: UKP, DUIIS, HITS	R-Prec: UKP, HITS, KSLP
Precision-at-5: HITS, DUIIS, KSLP	Precision-at-5: QUT, IISR, HITS	Precision-at-5: HITS, KSLP, UKP

4. ASSESSMENT AND EVALUTION

ASSESSMENT

- Wikipedia ground-truth (existing links)
- Manual Assessment

ASSESSMENT TOOL

🛃 NTCIR 9 Crosslink: Manual Assessment 💷 🔍					
File Utility Linking Language Help					
Source ID: 28271	Target ID: 698091	Completion:	Current subanchor:	Belongs to:	
Topic Title:	Target Intie: 促動制订任	0 / 2755	seaweed	seaweed	
Juan	\$6*5271/J/R				
		<u>^</u>			
	Sushi			鋸吻剃刀魚	
28271 375090162 2010-07-23720	0:06:43Z Nick Number 1526960	69	8091 6097982 2008-01-25T07:18:56Z Alan li 232717		
:Sushi :Japanese puisine	hic with	:棘	背魚目		
101		鋸	吻剃刀魚,又稱藍蜷剃刀魚、漂潮魚為 輻鰭魚綱 轅背魚目]海龍亞目 溝口魚科的其中一種。分布	
hig	hlighted	本	魚分布於 印度、西 太平洋。從 柬非到 馬紹爾群島,南 日本 度	即大堡礁,均可見。	
and	chors	水 特	案2公尺~20公尺。 徵		
Different types of night-zushi ready to be eaten 整特大,尾鳍类子头服件同形 成的鲜卵炎,體型也比公魚水 生態			則扁,無態線,體色多變化,由褐色至紛紅色或黃色均有,, 時大,尾鰭幾乎與腦砕同長,臀越圓形。第一背鰭有2枚大型 沖餅卵袋,體型也比公魚大,體長可達17公分。 態	曹帆布滿許多小黑點及白點;吻延長為扁管狀,無齒,具有一對鬚; 」 」始色斑,皮膚具有數列星狀突起;吻郎背面無細齒,平直而不彎曲,	第一背鳍無棘,胸鳍小,第二背鳍及月 母魚腹鳍的一部分下缘受形為相接而
Image		本, 經	魚體色適其棲地環境而定,多半生活或靠近海草或藻床,以 濟利用	疑態方式模仿海草,不易被發現。常成對出現,以吸食方式攜取浮游	主物,
	Une Antoho	3·	作為海水觀賞魚,不具食用價值。 考資料		
	Type of Sushi	4	兽鱼類資料		
is a Japanese dish consisting of cool fish or other seafood, 1 or put into rol is served rolled inside or around drier	ked vinegared rice which is commonly topped with other i lls. Sliced raw fish by itself is called sashimi, as distinct i d and pressed sheets of seaweed (or non) is makizushi (ngredients, such as rom sushi. Sushi that 巻き). Toppings	Target	document	
stuffed into a small pouch of fried tot	fu is inarizushi. A bowl of sushi rice with toppings scatter	ed over it is called			
Gindain Zuain (ウクし). Matury	100 10 100 100 and a second second		CIICK to	assess	
Image: Hiroshige Bowl of Sushi.jpg th sushi is fermented fish and rice, pres remains popular today.2 The term su literally "sushi" means "it's sour" 3	umb180px/right(Sushi by Hiroshige in Edo period)) The tr served with salt in a process that has been traced to Sou ishi comes from an archaic grammatical form no longer u a reflection of its historic fermonted mate	aditional form of theast Asia, where it sed in other contexts;	(Right a	lick irrolova	nt·
The selected behind it is addited by			(INBIIL)		,
The science behind the termentation the fish down into amino acids. This form of sushi in Japan, Narezushi, st Oshizushi and ultimately Edomae nig	of tish packed in noe is that the vinegar produced from the results in one of the five basic tastes, called umami in Ja ill very closely resembles this process. In Japan, Narezus prizushi, which is what the world today knows as "sushi."	ermenting rice breaks apanese.4 The oldest hi <mark>evolved</mark> into	Left clie	ck relevant)	
Contemporary Japanese sushi has lit the fermented fish was taken out of	ttle resemblance to the traditional lacto-fermented rice dis the rice, only the fish was consumed and the fermented r	th. Originally, when vice was discarded.			
	Anchor color lengend: Current anchor	Not assessed	complete Relevant Irrelevant	Previous	Next

VAL	.UAT	J	

• File-to-File evaluation Anchor-to-File evaluation

EVALUATION TOOL



The teams with the highest number of unique relevant links found with the Wikipedia ground-truth are: **UKP** for E2C, **QUT** for E2J, **HITS** for E2K. The teams with the highest number of unique relevant links found with the manual assessment results are: **QUT** for E2C, **HITS** for E2J, **KSLP** for E2K

6. SUBMISSIONS

GROUP	ORGANISATION
DUIIS	Daegu University
HITS	Heidelberg Institute for Theoretical Studies
IISR	Yuan Ze University
ISTIC	Institute of Scientific and Technical Information of China
KMI	The Open University
kslab_nut	Nagaoka University of Technology
KSLP	Kyungsung University
nthuisa	Academia Sinica
QUT	Queensland University of Technology
UKP	TU Darmstadt
WUST	Wuhan University of Science and Technology

Group	En-2-Zh	En-2-Ja	En-2-Ko
DUIIS	0	0	2
HITS	3	3	3
IISR	0	0	5
ISTIC	1	0	0
KMI	4	0	0
kslab_nut	0	1	0
KSLP	0	0	5
nthuisa	3	0	0
QUT	5	2	1
UKP	5	5	5
WUST	4	0	0
Sub-total	25	11	21
Total	57		

PRECISION AND RECALL IN F2F EVALUATION

 $Precision_{f2f} = \frac{number \ of \ correct \ links}{number \ of \ identified \ links}$ $Recall = \frac{number of correct links}{number of links in grels}$ Where *n* is the number of identified anchors; *N* is the number of anchors in *qrel*; k is the number of returned targets for anchor *i*; and k_i is the number of targets

recommended for this anchor.

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

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

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

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

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

where *n* is the number of topics (source articles used in evaluation); *m* is the number of identified items (articles for F2F or anchors in A2F); and P_{kt} is the precision of the top K items for topic t.

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

where *n* is the number of topics; and $P_t(a)R$ is the precision at *R* where *R* is the number of unique items in the *grels* of topic *t*.

Precision-at-N is computed using the average precision for all topics (source articles) at a pre-defined position N in the results list. Values of N were chosen as: 5, 10, 20, 30, 50, and 250.

