NTCIR-10
Crosslink Task

Crosslink Introduction

- Cross-lingual link discovery (Crosslink or CLLD) is concerned with automatically finding potential links between documents in different languages.

La Parisienne Almond Croissant

edit]

" in the Center and South and
onsidered to be the same, that is not
be crispy and contains a lot of butter,
ermore, the "cornetto vuoto" (Italian:
which include "crema pasticcera" iam
sugar.

crema pasticcera ?
What?

If there is such link

It means "custard"!!

奶黃

維基百科，自由的百科全书

關於美國內戰、北美印第安戰爭中聞名的美軍軍官，請見乔治·卡士達。

奶黃（**Custard**）也稱為**蛋奶凍**、**蛋漿**，或音譯為**卡士達**、**吉士**，泛指雞蛋與牛奶混合後加熱而凝固而成的食品，大多數是應用於製作甜品。製作時常會視乎需要，混入砂糖、奶油、香草如香子蘭（Vanilla，又譯雲呢拿）、栗粉或明膠等。

參見   [编辑]

- 吉士粉
- 梳芙厘

這是與食材或調味料相關的小作品。你可以通过編輯或修訂擴充其内容。

奶黃

Crosslink Introduction

- CLLD algorithms actively recommend a set of meaningful anchors in the context of a source document and establish links to documents in an alternative language.

- CLLD is helpful for complimentary knowledge discovery in different language domains.

Cross-lingual Link Discovery

Article: Australia

…
Ranked third in the [Index of Economic Freedom](#) (2010),[178] Australia is the [world's thirteenth largest economy](#) and has [the ninth highest per capita GDP](#); higher than that of the United Kingdom, Germany, France, Canada, Japan, and the United States. The country was ranked second in the United Nations 2010 [Human Development Index](#) and first in [Legatum](#)'s 2008 [Prosperity Index](#).[179] All of Australia's major cities fare well in global comparative livability surveys;[180] Melbourne reached first place on **The Economist**'s 2011 [World's Most Livable Cities](#) list, followed by Sydney, Perth, and Adelaide in sixth, eighth, and ninth place respectively.[181] Total government debt in Australia is about $190 billion.[182] Australia has among the highest house prices and some of the highest household debt levels in the world.
…

No link was created for this term, for finding articles in languages we prefer traditionally we do:

Search → **Translate**

经济学人
…

The Economist
…

エコノミスト
…

이코노미스트
…

How to automatically create cross-lingual links for a document if no links existing yet?

❑ Links in other languages?
❑ New articles?
❑ Missing links?
❑ Not what we are looking for?
❑ What about other relevant links?

Cross-lingual Link Discovery

❖ Cross-lingual Links
❖ New Links
❖ Better Links
❖ More options

✓

- All about *multi-lingual knowledge discovery* in knowledge bases (e.g. Wikipedia)
- All about *easy* and *efficient* information access

Link Level

Evaluate in both level

- File-to-File (F2F)

- Anchor-to-File (A2F)

File-to-File

Rice cake

From Wikipedia, the free encyclopedia

A **rice cake** may be any kind of food item made from rice that has been shaped, condensed, or otherwise combined into a single object. A wide variety of rice cakes exist in many different cultures in which rice is eaten, and are particularly prevalent in Asia. Common variations include cakes made with rice flour, those made from ground rice, and those made from whole grains of rice compressed together or combined with some other binding substance.The Japanese rice cake came from Southeast Asia in Jomon period. The rice cake evolved into Japanese sweets Wagashi by development of the Japanese tea ceremony[1].

Contents [hide]

1 Types of rice cakes by region

Link!!

餅

餅（日本語：もち，英語：**rice cake**）とは、穀物、特に米に水分と熱を加えた後に、外力を加えて練り合わせ、成形した食品の一種。粒状の米を蒸して、杵で搗いたものは、**つき餅**（搗き餅）という。穀物の粉に湯を加えて練り、蒸しあげたものは、**練り餅**（ねりもち）という。日本では餅といえばつき餅をさす場合が多い。

この項目に含まれる文字「**餅**」は、オペレーティングシステムやブラウザなどの環境により表示が異なります。

目次 [非表示]

1 概要
2 つき餅（搗き餅）
　　2.1 歴史
　　2.2 材料
　　2.3 形
　　2.4 餅つき（餅搗き）

丸餅

Anchor-to-File

La Parisienne Almond Croissant

Link!!

"crema pasticcera"

奶黃

維基百科，自由的百科全书

關於美國內戰、北美印第安战争、中聞名的美軍、軍官，請見乔治·卡士達。

奶黃（**Custard**）也稱為**蛋奶凍、蛋漿**，或音譯為**卡士達、吉士**，泛指雞蛋與牛奶混合後加熱而凝固而成的食品，大多數是應用於製作甜品。製作時常會視乎需要，混入砂糖、奶油、香草如香子蘭（Vanilla，又譯雲呢拿）、粟粉或明膠等。

參見   [编辑]

- 吉士粉
- 梳芙厘

這是與**食材**或**調味料**相關的小作品。你可以通过編輯或修訂擴充其内容。

奶黃

The Goal of Crosslink Task

- It is aimed to create a reusable resource for evaluating automated cross language link discovery approaches.
  - The results of this research will be used in building and refining systems for automated link discovery.

Crosslink task at NTCIR-9

- Cross-lingual link discovery (Crosslink) as a pilot task of NTCIR-9 has been successfully held in 2011
- Below is a list of participating teams with 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

Crosslink task at NTCIR-9 (Cont.)
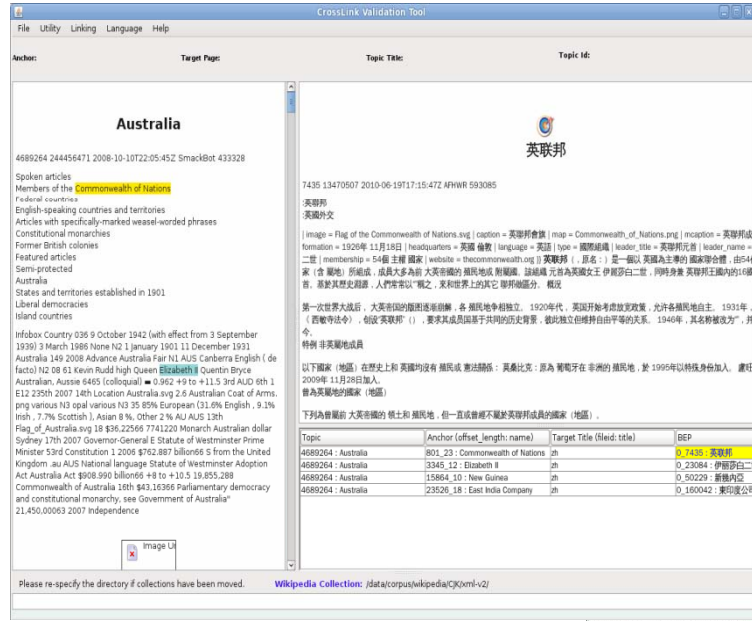
- There were in total 57 runs from 11 teams were received.

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		

Crosslink Task at NTCIR-10

- CLLD is a core task of NTCIR-10
- New Subtasks
  - Chinese to English CLLD (C2E)
  - Japanese to English CLLD (J2E)
  - Korean to English CLLD (K2E)
- New Document Collections
  - ECJK Wikipedia collections
- New Topics
  - 25 topics for each language

Evaluation Framework

- Assessment Types: Automatic (Wikipedia Ground Truth), Manual (human in the loop)
- Test data, training data, gold standard, ECJK Wikipedia collections, validation tool, assessment tool and evaluation tool (with system evaluation metrics).
- Snapshots of the validation, assessment and evaluation tools:
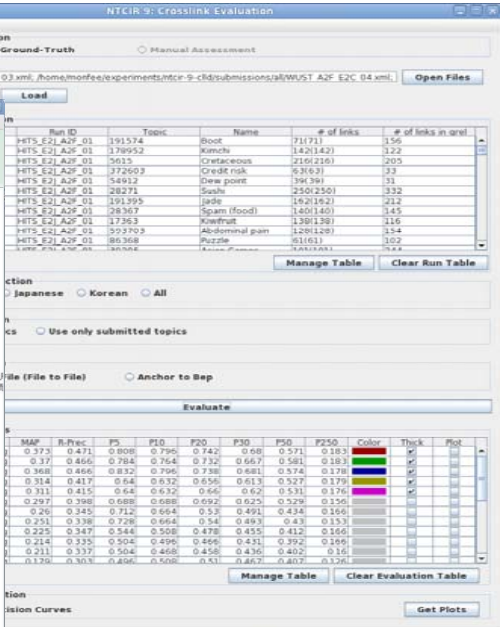
Validation Tool

Assessment Tool

Evaluatuion Tool

System Evaluation Metrics

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

where *n* is the number of topics; *m* is the number of identified items (links or anchors); P_{kt} is the precision at top *k* items (links or anchors) for topic *t*; $P_t@R$ (= *number of correct items (links or anchors) / number of items (links or anchors) in qrel*) is the precision calculated using number of links / anchors in *qrel* as denominator for topic *t*.

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

- *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.

Expected Outcomes

Evaluation Framework

- More good submissions

- More original and innovative approaches can be seen in identifying meaningful anchors and suggesting high quality cross-lingual links

- The research results can really help the cross-lingual knowledge discovery in knowledge bases
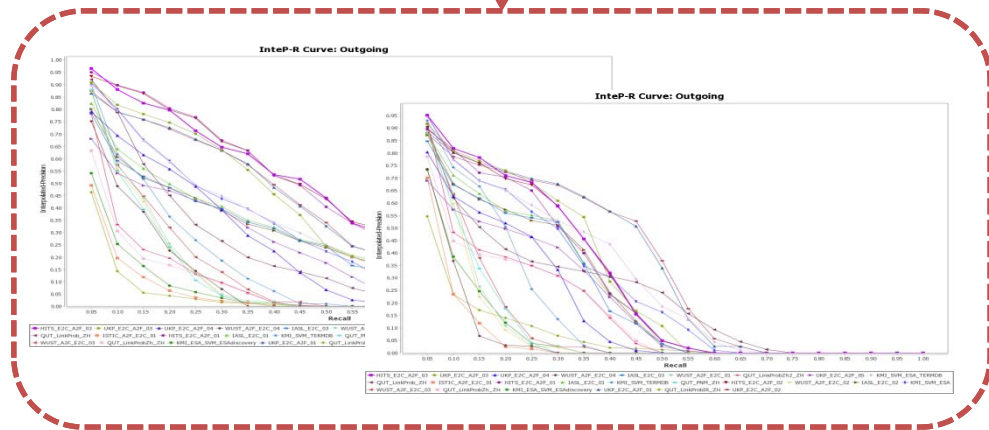
- The evaluation framework will be further refined

- The evaluation methods will be further perfected to distinguish the good and the bad CLLD algorithms for the new subtasks

Expected Participants

- Previous active participants of NTCIR-9 Crosslink task

- The registered participants of NTCIR-9 Crosslink task.

- Participants of CLIR or IR4QA task in previous NTCIR workshops.

- Participants of previous INEX Link-the-Wiki track

- Other researchers in the CLIR field

Contacts

- Organisers:

Shlomo Geva	Queensland University of Technology, Australia
Andrew Trotman	Universityof Otago, New Zealand
Yue Xu	Queensland University of Technology, Australia
Eric Tang	Queensland University of Technology, Australia
In-Su Kang	Kyungsung University, South Korea
Fuminori Kimura	Ritsumeikan University, Japan
Haitao Mi	Chinese Academy of Sciences, China

- Mailing list: crosslink@lists.otago.ac.nz