Cross-Lingual Knowledge Discovery: Chinese-to-English Article Linking in Wikipedia

Ling-Xiang Tang [1], Andrew Trotman [2], Shlomo Geva [1], Yue Xu[1]

[1] Science and Engineering Faculty, Queensland University of Technology,
Brisbane, Australia
{l4.tang, s.geva, yue.xu}@qut.edu.au
[2] Department of Computer Science, University of Otago,
Dunedin, New Zealand
andrew@cs.otago.ac.nz

Abstract. In this paper we examine automated Chinese to English link discovery in Wikipedia and the effects of Chinese segmentation and Chinese to English translation on the hyperlink recommendation. Our experimental results show that the implemented link discovery framework can effectively recommend Chinese-to-English cross-lingual links. The techniques described here can assist bi-lingual users where a particular topic is not covered in Chinese, is not equally covered in both languages, or is biased in one language; as well as for language learning.

Keywords: Wikipedia, Cross-lingual Link Discovery, Link Mining, Anchor Identification, Link Recommendation, Chinese Segmentation, Translation.

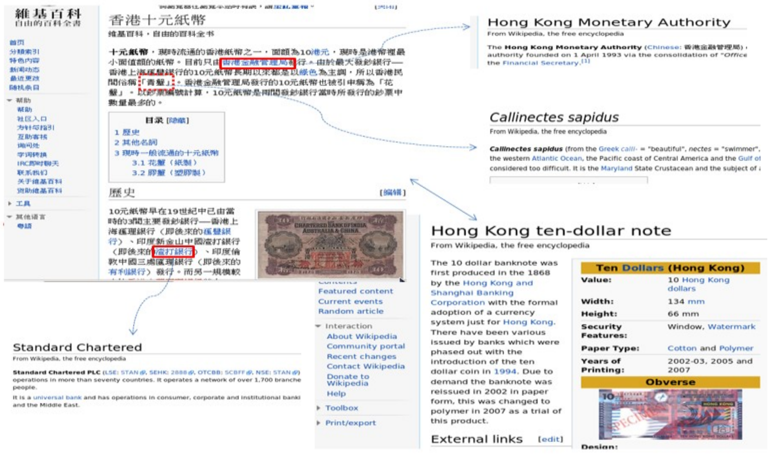
1 Introduction

Wikipedia is the largest multi-lingual encyclopaedia online with over ten million articles in almost every written language. However, knowledge in Wikipedia could have boundaries because of language barriers. The anchored links in Wikipedia articles are mainly created within the same language domain. Knowledge sharing and discovery are impeded by the absence of links between different language domains. Users are forced to use one language version of the resource and are not easily able to switch languages where appropriate. A user may prefer multiple explanations, or just the one in their preferred language, or the richer content, or to extend their understanding of a language through reading translations.

For example, in Hong Kong the word 花蟹 (“flower crab”) is colloquial for the ten-dollar note. There are, indeed, 花蟹 entries in both Chinese and English Wikipedia but they are not linked to each other. **Fig. 1** shows English and Chinese Wikipedia pages on the Hong Kong ten-dollar note. From the figure, it can be seen that there should be bi-directional language links, but that they have not yet been created. The boxed texts in the Chinese page could be used to further generate anchored links for bi-lingual users to explore those anchors' English counterparts.

Previous studies of link discovery between documents in different languages include the followings. Sorg & Cimiano [1] tackle the German and English Wikipedia language-link problem using a classification-based approach. Their study particularly examines missing language-links between Wikipedia articles on the same topic. Melo & Weikum [2] do the opposite, they examine incorrect Wikipedia language-links between articles on the same topic. In the NTICR Crosslink task, Fahrni *et al.* [3] implemented a CLLD system using a graph-based method for disambiguation and achieved very good results; the Kim & Gurevych approach performed the best in linking English documents to Chinese when measured with manual assessment results.

In this paper, we focus on the realisation of efficient and effective automated Chinese to English link discovery in Wikipedia and study the effects of Chinese segmentation and Chinese to English translation on the hyperlink recommendation.



維基百科
自由的百科全書

首页
分類索引
特色内容
新闻动态
最近更改
隨机条目

帮助
帮助
社区入口
方针与指引
互助客栈
询问处
字词转换
IRC即时聊天
联系我们
关于維基百科
资助維基百科

工具

其他语言
粵語

香港十元紙幣

維基百科，自由的百科全书

十元紙幣，現時流通的香港紙幣之一，面額為10港元，現時是港幣裡最小面值額的紙幣。目前只由香港金融管理局發行。由於最大發鈔銀行——香港上海滙豐銀行的10元紙幣長期以來都是以綠色為主調，所以香港民間俗稱「青蟹」。香港金融管理局發行的10元紙幣也被引申稱為「花蟹」。以鈔票編號計算，10元紙幣是兩間發鈔銀行當時所發行的鈔票中數量最多的。

目录 [隐藏]
1 歷史
2 其他名詞
3 現時一般流通的十元紙幣
3.1 花蟹（紙製）
3.2 膠蟹（塑膠製）

歷史 [编辑]

10元紙幣早在19世紀中已由當時的3間主要發鈔銀行——香港上海滙理銀行（即後來的滙豐銀行）、印度新金山中國渣打銀行（即後來的渣打銀行）、印度倫敦中國三處滙理銀行（即後來的有利銀行）發行。而另一規模較

Featured content
Current events
Random article

Interaction
About Wikipedia
Community portal
Recent changes
Contact Wikipedia
Donate to Wikipedia
Help

Toolbox

Print/export

Hong Kong Monetary Authority

From Wikipedia, the free encyclopedia

The **Hong Kong Monetary Authority** (Chinese: 香港金融管理局) authority founded on 1 April 1993 via the consolidation of "*Office* the Financial Secretary.[1]

Callinectes sapidus

From Wikipedia, the free encyclopedia

Callinectes sapidus (from the Greek *calli-* = "beautiful", *nectes* = "swimmer", the western Atlantic Ocean, the Pacific coast of Central America and the Gulf of considered too difficult. It is the Maryland State Crustacean and the subject of

Hong Kong ten-dollar note

From Wikipedia, the free encyclopedia

The 10 dollar banknote was first produced in the 1868 by the Hong Kong and Shanghai Banking Corporation with the formal adoption of a currency system just for Hong Kong. There have been various issued by banks which were phased out with the introduction of the ten dollar coin in 1994. Due to demand the banknote was reissued in 2002 in paper form, this was changed to polymer in 2007 as a trial of this product.

External links [edit]

Ten Dollars (Hong Kong)	
Value:	10 Hong Kong dollars
Width:	134 mm
Height:	66 mm
Security Features:	Window, Watermark
Paper Type:	Cotton and Polymer
Years of Printing:	2002-03, 2005 and 2007
Obverse	
Design:	

Standard Chartered

From Wikipedia, the free encyclopedia

Standard Chartered PLC (LSE: STAN, SEHK: 2888, OTCBB: SCBFF, NSE: STAN) operations in more than seventy countries. It operates a network of over 1,700 branche people.

It is a universal bank and has operations in consumer, corporate and institutional banki and the Middle East.

Fig. 1. The Wikipedia pages on "flower crab"

2 Chinese / English Wikipedia

2.1 Corpora Information

Dumps of the Chinese and English Wikipedia taken in June 2010 were converted into files marked up using the YAWN system [1]. After conversion, there were 3,484,250 properly formatted English articles and 316,251 properly formatted Chinese articles. In the collection, just over half of the Chinese articles (170,637), but only 5% of the English articles (169,974), were language cross-linked.

2.2 Links In Wikipedia

Language Link. Wikipedia of different languages is connected through links between articles on the same topic with a single page-to-page language link. Those language links can be used to produce Chinese / English title mapping table T_{lang}. This table can be utilised as a dictionary for translation which will be discussed in section 4.2.

Anchored Mono-Lingual Links. There are nearly 8 million (mostly mono-lingual) links in the Chinese corpus; and around 90 million links in the English corpus. From each corpus, a link table T_{link} ($T_{link\text{-}chinese}$ for Chinese and $T_{link\text{-}english}$ for English) can be mined. All T_{link} tables contain a list of linked documents each with a unique id, a link frequency (*lf*), and a document frequency (*df*). The usage of link information mined from the corpora will be discussed in the next section. Several entries taken from T_{lang} and $T_{link\text{-}chinese}$ are showed in **Table 1**.

Table 1. Extracts from T_{lang} and $T_{link\text{-}chinese}$

T_{lang}		$T_{link\text{-}chinese}$			
Title (zh)	Title (en)	Title (zh)	ID	*lf*	*df*
花旗银行	Citibank	花旗银行	53090	42	46
椰子蟹	Coconut crab	椰子蟹	536691	10	10
米高佐敦	Michael Jordan	英国	39793	5212	6866

3 Linking Chinese to English

3.1 Chinese Natural Language Processing

Study of both English mono-lingual and English-to-Chinese document linking has been covered by the recent research on link discovery [4, 5]. However, linking Chinese documents to English still has certain unique problems that need to be addressed. To the best of our knowledge, there are no published research papers that address this.

Chinese Wikipedia is a collaborative effort of contributors from different Chinese spoken geographic areas with different knowledge backgrounds and language variations. They cite modern and ancient sources combining simplified and traditional Chinese text, as well as regional variants. Therefore, in order to link Chinese documents to English documents while considering the linguistic complexity in the Chinese Wikipedia articles, it is necessary to break the Chinese text into separate words (to segment the text). Chinese segmentation breaks long strings of characters into n-gram words. It is presumed that this is a particularly critical step in Chinese-to-English cross-lingual link discovery because it affects not only the identification of the anchors but also the ability to translate them into English. The error rate of anchor translation, and translation in general, is dependent on the quality of the segmentation [6].

3.2 Article Linking

The state-of-the-art techniques for document linking have been seen in past studies. For mono-lingual link discovery there are: the Link mining (ML) method [7] and the Page Name Matching (PNM) method [8]. In this paper, it was intended to make use of these two techniques to build a comprehensive, effective Chinese-to-English link discovery framework that can recommend high quality link efficiently between two knowledge domains in different languages.

Link Mining. Link mining method mines the existing links in a single language version of Wikipedia to create a *link table*, T_{link}, of mono-lingual anchor-to-target ($a \rightarrow d$) pairs. From this link table, the probability of any sequence of terms being an anchor can be computed (for pre-existing anchors). Based on the existing link information that is extracted during the mining phase, the best target for an anchor can also be computed. Note that the same anchor text may be linked to different destinations in different instances where it appears and so it is necessary to identify the most likely link.

Itakura & Clarke [7] trawl English Wikipedia and extract all anchor target pairs. They then re-trawl the collection looking for the frequency of the anchor phrases used either as a link or in plain text. From this they compute an anchor weight, γ, the probability that a given phrase is an anchor and linked to a specific target document as follows:

$$\gamma = \frac{number\ of\ pages\ that\ have\ link(a \rightarrow d)}{number\ of\ pages\ that\ have\ text\ of\ anchor(a)} \qquad (1)$$

where the numerator is the link frequency, *lf*, of anchor *a* pointing to document *d*; and the denominator is the document frequency (*df*) of anchor *a* in the corpus.

To link documents within Wikipedia or any documents with Wikipedia, Mihalcea & Csomai [9] and Milne & Witten [10] also use a similar method to weight phrases.

Page Name Matching. An alternative approach for link discovery is title matching (also known as name-matching, and entity matching). For mono-lingual link discovery Geva [8] builds a *page title table,* a list of titles of all documents in Wikipedia. For a given document, a list of all possible n-gram substrings are built and then from the list the longest that are also in the page title table are chosen as the anchors. The targets are the documents with the given title.

To use this in Chinese to English link discovery, it is necessary to first construct a table of corresponding English and Chinese documents. Then, for a new Chinese document, identify all substrings that match Chinese document titles as the anchors. The targets are the corresponding English documents.

4 The Proposed Approach

4.1 Anchor Identification

In this work, we use both the anchor weight [7] and the page name matching [8] methods to identify anchors. The reasons are: first, they are very efficient methods, and anchors can be created easily on-the-fly because the title mapping table T_{lang}, and anchor weights (γ scores) of all possible anchor candidates can be pre-mined and pre-computed; second, the recommended anchors are mainly from the anchor pool that they are either article titles that readers might look up or ones that were previously linked by the human editors.

With the anchor weighting method, for a new previously unlinked document all possible n-gram substrings from the document are first computed. For each of these the γ score is looked-up and the anchors sorted by these values. An arbitrary number (based on a threshold, or alternatively a density) of highly ranked links are then chosen. In the case of overlapping anchors, the longest anchor is chosen.

Page name matching has a similar anchor identification process that from the document all possible n-grams that can be found in the Chinese title table are extracted and but then sorted based on the length of title. The rationale for choosing the longer titles – which also proves correct in experiments – is that longer phrase matches are less likely to be coincidental, and longer phrases in text are generally more specific than shorter ones.

The issue with these two anchor identification methods is that without Chinese segmentation anchors may be created for unrelated topics. For example, the following two sentences contain non-Chinese words (underscored) that could be mistakenly linked to the unrelated Chinese articles:

"胸甲骑兵在腓特烈大帝和拿破仑的军队中都扮演过非常重要的角色。"—taken from the Chinese Cuirassier article[1]. In this sentences, the two adjoining characters—中 and 都 means "in" and "both" separately, but together they (中都) are often used as place names (e.g. an old name for *Beijing* city).

4.2 Anchor Translation

Triangulation. One way to use page name matching and link mining approaches is to mine in one language and to identify target documents translated into the second language.

To do this, a table of documents existing in both languages could be used. Such a table, T_{lang}, can be generated from the page-to-page language links present in Wikipedia. This is similar to the translation memory approach that is commonly used in Machine Translation. This is a form of triangulation. An English page is a good target to

[1] http://zh.wikipedia.org/wiki/胸甲骑兵

a Chinese anchor if there exists a link from the anchor to the Chinese document and from the Chinese document to the English document. The relationship of the triangulation is illustrated in **Fig. 2**.

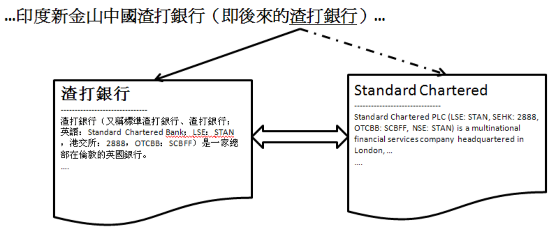

...印度新金山中國渣打銀行（即後來的渣打銀行）...
渣打銀行
渣打銀行（又稱標準渣打銀行、渣打銀行；英語：Standard Chartered Bank；LSE：STAN，港交所：2888，OTCBB：SCBFF）是一家總部在倫敦的英國銀行。
....
Standard Chartered
Standard Chartered PLC (LSE: STAN, SEHK: 2888, OTCBB: SCBFF, NSE: STAN) is a multinational financial services company headquartered in London, ...
....

Fig. 2. Cross-lingual triangulation

Machine Translation. As an addition to the translation with triangulation, Candidate anchors can be translated into English using Google Translate API[2]. Machine translation will be particularly helpful when triangulation fails to provide a proper translation for a high valued anchor and this will be often the case because table T_{lang} is an incomplete set of the mapping of Chinese / English article titles in Wikipedia.

4.3 Link Recommendation

Link recommendation is the final step of our link discovery approach. As in link mining and page name matching methods, all anchor candidates (either from T_{lang} or $T_{link\text{-}chinese}$) already have been associated with a specific target document. So with these two methods, once an anchor is identified, the target document is also determined. So Different anchor identification, translation and final document linking methods will lead to different discovered link sets.

5 Experiments

5.1 Anchor / Link Specification

Although there is no hard limit to the number of anchors that may be inserted into a document, a user will become overwhelmed if almost every term in an article is also an anchor. For evaluation purposes we impose a limit of 50 links per document.

5.2 Evaluation

To simulate Chinese-English cross-lingual linking, we create a set of 36 topics[3] (including 香港十元紙幣 (*Hong Kong ten-dollar note*)), then mine the remaining corpus

[2] http://code.google.com/apis/language/translate/overview.html

[3] http://crosslink.googlecode.com/files/zh-topics-36.zip

to generate the two kinds of tables, T_{link} and T_{lang}. With the Wikipedia ground-truth, the Precision-at-N and Link Mean Average Precision (LMAP) metrics employed in NTCIR-9 Crosslink task [5, 11] are used to quantify the performance of the different cross-linking methods.

Table 2. Experimental runs information

Run Name	Description
LinkProb	Anchor identified with the link table $T_{link\text{-}chinese}$ with link mining method, γ computed with $T_{link\text{-}chinese}$, and target links were identified trough triangulation
PNM	Page name matching through triangulation with T_{lang}
LinkProbEn	Anchor identified with the link table $T_{link\text{-}chinese}$, then with machine translation link probability taken from $T_{link\text{-}english}$
LinkProbEn2	Similar to *LinkProbEn* but final ranking with $T_{link\text{-}chinese}$
LinkProb_S	*LinkProb* run with segmentation
LinkProbEn_S	*LinkProbEn* run with segmentation

Table 3. Performance of experimental runs

Run ID	LMAP	P@5	P@10	P@20	P@50
LinkProb	0.168	0.800	0.694	0.546	0.386
PNM	0.123	0.667	0.567	0.499	0.351
LinkProbEn2	0.095	0.456	0.428	0.338	0.247
LinkProbEn	0.085	0.489	0.394	0.315	0.211
LinkProb_S	0.059	0.411	0.322	0.268	0.201
LinkProbEn_S	0.033	0.233	0.186	0.144	0.118

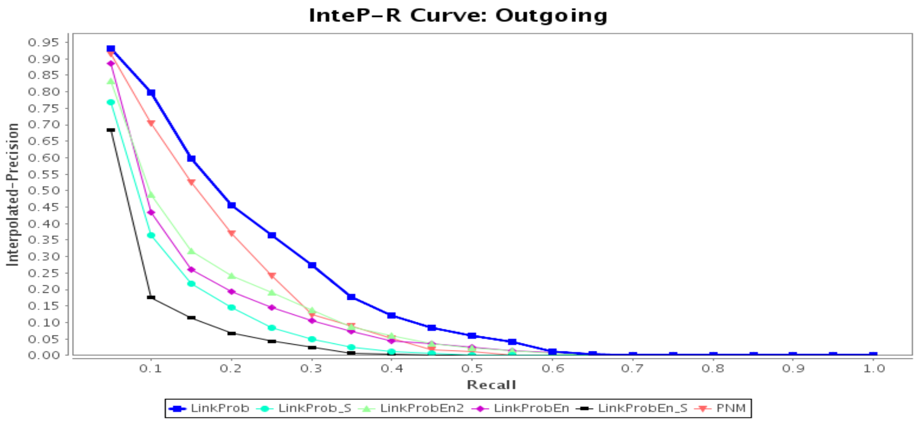


Fig. 3. The interpolated precision/recall curves for the different methods

5.3 Experimental Runs

By combining different translation methods (either triangulation or machine translation) and different anchor weighting strategy (γ score computed using either $T_{link\text{-}chinese}$ or $T_{link\text{-}english}$), the resulting discovered link sets are also different. The runs are outlines in **Table 2**. The segmentation approach proposed by Tang *et al.* [12] was used to complete the segmentation task.

6 Results and Discussion

The LMAP and P@N scores for the different runs are given in **Table 3**. Runs are scored on the extracted Wikipedia ground-truth and sorted on LMAP. Precision and recall curves are given in **Fig. 3**. All runs except *PNM* use the same anchor identification strategy. So, the difference in the performance of those runs can be attributed to the segmentation and translation. Overall, the best performing run, *LinkProb* has the best combination of strategies (and not a different method of choosing anchors).

6.1 Segmentation in CELD

In all cases non-segmented runs out performed the segmented variant of the run. Contrary to intuition, segmentation interferes with anchor identification. This reflects both the non-perfect performance of any segmentation algorithm, and the links themselves being unlikely to be ambiguous in context (because they are named-entities).

There is no doubt that segmentation can increase the accuracy of Chinese text processing, but for link discovery the problem lies in the difficultly of controlling the segmentation granularity for perfect anchor identification. Small granularity will result in small size of words and may help reducing errors of matching the anchors to unrelated topics but may miss out the named entities with compound words; large granularity in segmentation will however cause the exact opposite problem. The extra step for Chinese segmentation in link discovery will increase the computational complexity. Therefore, Chinese segmentation is not absolutely required for Chinese-to-English link discovery if the goal is set to achieve ultimate linking performance.

6.2 Translation in CELD

All runs that used machine translation performed worse than *LinkProb* and *PNM*. Run *LinkProbEn2* and run *LinkProb* indentified the same set of initial candidate Chinese anchors and used the same link ranking strategy. *LinkProbEn2*, however, performs worse than *LinkProb*. This suggests that the performance deteriorates as a consequence of the translation process. A failure analysis of the runs suggests that the problem is caused by translation error. **Error! Reference source not found.** lists some of the anchor candidates (column 1) that were incorrectly machine translated (column 2) and the preferred target document seen through link mining (column 3). The failure in translation is similar to that caused by segmentation. Without perfect knowledge of all entities the translation software cannot produce perfect results. Such results cannot be expected because the entity list cannot be closed.

Table 4. Example translation errors in the runs

Anchor	MT	Wiki
資治通鑑	Mirror	Zizhi Tongjian
社稷	Boat	Soil and grain
白骨精	White-Boned Demon	Bai Gu Jing

The result suggests that the mined mapping table T_{lang} used in runs *LinkProb* and *PNM,* is a better translation table than classical machine translation. This is hardly surprising as it is domain specific, and entity list (rather than phrasal text). An alternative we did not test was a combination of the two approaches – using machine translation if an entity could not be translated.

6.3 Chinese-to-English Document Linking

As can be seen from both **Table 3** and **Fig. 3**, run *LinkProb* performed best when scored using LMAP and P@N. Given that the number of candidate links in T_{lang} used by cross-lingual page name matching algorithm is much smaller than $T_{link\text{-}chinese}$ used by link mining method the good performance of *PNM* is surprising but encouraging.

Run *LinkProbEn2* ranked third performing better than *LinkProbEn*. The difference between the two runs was the source of the link probability γ score. In the former the probability came from the Chinese language corpus, but in the latter it came from the English corpus. This suggests that Chinese is a better predictor of which English documents to link to than is English. So the link mining was the best algorithm we tested for Chinese-English cross-language link discovery. As the experiments are the first reported for solving the Chinese-to-English document linking problem, the LMAP and P@N scores of run *LinkProb* are the best results to date.

7 Conclusion and future work

In this paper we presented a Chinese to English link discovery framework for automatically identifying anchors in Chinese document that should target documents in English. The experimented Chinese-to-English Cross-linking approach included the use of Chinese word segmentation, Chinese to English translation, and link mining.

Although Chinese segmentation and machine translation are two essential steps in Chinese to other-language information retrieval, our results suggest that they are not needed for link discovery. This is because segmentation is implicit in the anchor mining and the translation is implicit in cross-language triangulation.

The experimental results show that the implemented link discovery framework can effectively recommend Chinese-to-English cross-lingual links. This CELD framework can also be used as a Wikipedia article recommendation system to suggest articles for further reading. In future, to further improve our system performance we would like to explore other techniques such as linkage factor graph model used by Wang et al. [13] in their work of linking English Wikipedia to other online Chinese encyclopaedias.

References

1. Sorg, P., Cimiano, P.: Enriching the Crosslingual Link Structure of Wikipedia - A Classification-Based Approach. AAAI 2008 Workshop on Wikipedia and Artifical Intelligence. (2008)
2. Melo, G.D, Weikum, G.: Untangling the cross-lingual link structure of Wikipedia. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 844-853. (2010)
3. Fahrni, A., Nastase, V., Strube, M.: HITS' Graph-based System at the NTCIR-9 Cross-lingual Link Discovery Task. Proceedings of NTCIR-9, pp. 473-480. (2011)
4. Huang, W., Geva, S., Trotman, A.: Overview of the INEX 2009 Link the Wiki Track. Proceedings of INEX 2009, pp. 312-323. (2010)
5. Tang, L.-X., Geva, S., Trotman, A., Xu, Y., Itakura, K.Y.: Overview of the NTCIR-9 Crosslink Task: Cross-lingual Link Discovery. Proceedings of NTCIR-9, pp. 437-463. (2011)
6. Chang, P.-C., Galley, M., Manning, C.D.: Optimizing Chinese word segmentation for machine translation performance. Proceedings of the Third Workshop on Statistical Machine Translation. (2008)
7. Itakura, K., Clarke, C.: University of Waterloo at INEX2007: Adhoc and Link-the-Wiki Tracks. Proceedings of INEX 2007, pp. 417-425. (2008)
8. Geva, S.: GPX: Ad-Hoc Queries and Automated Link Discovery in the Wikipedia. Proceedings of INEX 2007, pp. 404-416. (2008)
9. Mihalcea, R., Csomai, A.: Wikify!: linking documents to encyclopedic knowledge. Proceedings of CIKM '07, pp. 233-242. (2007)
10. Milne, D., Witten, I.H.: Learning to link with wikipedia. Proceeding of CIKM 2008 pp. 509-518. (2008)
11. Tang, L.-X., Itakura, K.Y., Geva, S., Trotman, A., Xu, Y.: The Effectiveness of Cross-lingual Link Discovery. Proceedings of The Fourth International Workshop on Evaluating Information Access (EVIA), pp. 1-8. (2011)
12. Tang, L.-X., Geva, S., Trotman, A., Xu, Y.: A Boundary-Oriented Chinese Segmentation Method Using N-Gram Mutual Information. Proceedings of CIPS-SIGHAN Joint Conference on Chinese Language Processing, pp. 234-239. (2010)
13. Wang, Z., Li, J., Wang, Z., Tang, J.: Cross-lingual knowledge linking across wiki knowledge bases. Proceedings of the 21st international conference on World Wide Web, pp. 459-468. (2012)