Discovering Links by Context Similarity and Translated Key Phrases for NTCIR9 CrossLink

Yi-Hsun Lee, Chung-Yao Chuang, Cen-Chieh Chen, Wen-Lian Hsu
Institute of Information Science
Academia Sinica, Taiwan
{rog,cychuang,can,hsu}@iis.sinica.edu.tw

ABSTRACT

This paper describes our participation in the NTCIR-9 Cross-lingual Link Discovery (CrossLink) task. The task is focused on suggesting links between English Wikipedia and Chinese, Korean, and Japanese Wikipedia. In this event, we experimented our method on the English-to-Chinese subtask. Our method divides the link discovery process into three steps. First, we use a frequency-based anchor tagger to find phrases or pieces of text that may be viable for linking to other pages in the source language (English in our case.) Because there may be more than one page that an anchor can link to, we narrow down those pages by similarity in context overlapping. Next, we extract key phrases from remaining pages in the last step and translate those phrases using Google Translate[1] into target language. The translated phrases are then used as query to retrieve articles in Chinese indexed by Lucene[2]. Finally, we utilize a ranking algorithm based on pages' connection graph to sort the candidate articles. Our system achieved MAP score 0.225 when evaluating with Wikipedia ground truth, and 0.205 with manual assessment.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*text analysis*; I.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*linguistic processing*

General Terms

Experimentation

Keywords

Wikipedia, cross-lingual link discovery, link recommendation, cross-lingual information retrieval

Team Name: nthuisa (IASL)

Subtask: English to Chinese CrossLink

External Resources Used: Google Translate, Lucene

1. INTRODUCTION

In this paper, we describe our effort on NTCIR-9 Cross-lingual Link Discovery (CrossLink) task, which aims to promote research in automated methods for finding potential links between documents in different languages. The language differences between documents could easily stagnate the propagation of information and potentially create barriers between sharable knowledge. To mitigate such a problem, cross-lingual link discovery attempts to actively recommends a set of meaningful anchors in the source document, and link them to appropriate articles in other languages.

The CrossLink task was hold based on the documents of Wikipedia[3]. Wikipedia is an online multilingual encyclopedia that contains large number of articles in many languages. Despite that it has a richly linked structure between articles within each language, cross-lingual references are rare, except articles about exactly the same subject. Therefore, one way that we can facilitate the information access is to apply cross-lingual link discovery techniques to Wikipedia. With such an objective in mind, CrossLink comprises three independent subtasks:

- English to Chinese cross-lingual link discovery
- English to Japanese cross-lingual link discovery
- English to Korean cross-lingual link discovery

For each subtask, a set of English documents are provided, and each document is expected to be analyzed for potential anchors and recommended links to the articles in the target language.

In this event, we participated the English-to-Chinese subtask. The collection of Chinese Wikipedia used in this subtask comprises 318,736 articles, and was gathered on June 2010. A set of 25 articles were chosen from the English Wikipedia for evaluating the effectiveness of the proposed methods. All test articles are prepared in XML format without link tags, i.e. all the links and anchor information in the original pages were removed.

The rest of this paper will detail our approach for this subtask, and is organized as follows. Section 2 gives the background of this task. In Section 3, the proposed method is described. The empirical results are listed in Section 4. Section 5 discusses the results and Section 6 outlines some possible future works.

2. BACKGROUND

Wikipedia, a free on-line encyclopedia, contains lots of articles with rich linked information. Such linked information provide lots of extensive or related information of the current article. Because the open and voluntary nature of

[1] http://translate.google.com

[2] http://lucene.apache.org

[3] http://wikipedia.org

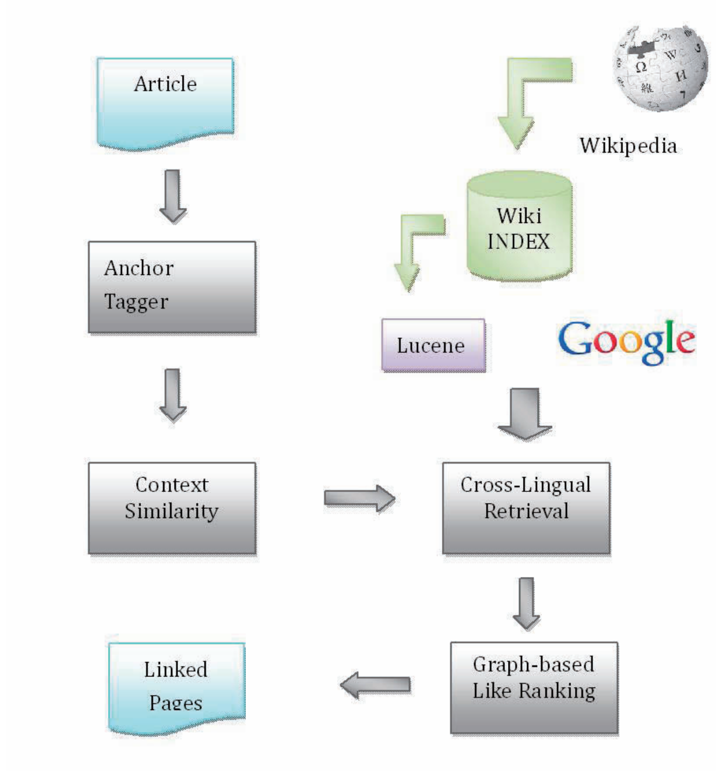
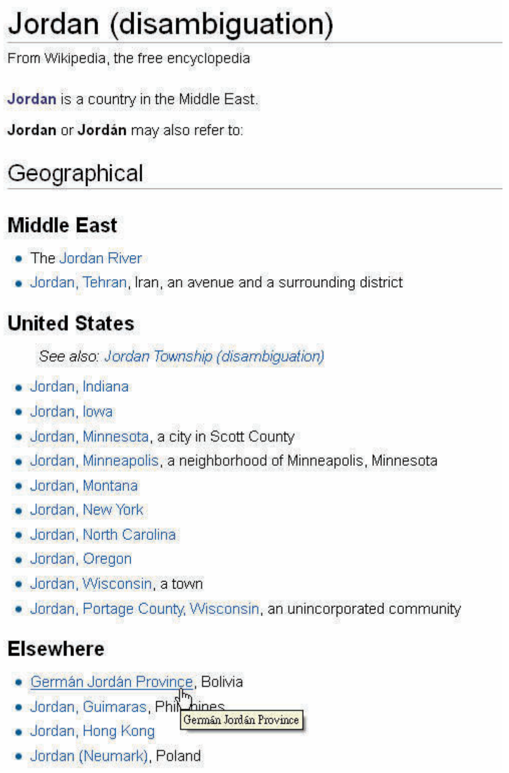


Figure 1: System Architecture



Jordan (disambiguation)

From Wikipedia, the free encyclopedia

Jordan is a country in the Middle East.

Jordan or **Jordán** may also refer to:

Geographical

Middle East

- The Jordan River
- Jordan, Tehran, Iran, an avenue and a surrounding district

United States

See also: Jordan Township (disambiguation)

- Jordan, Indiana
- Jordan, Iowa
- Jordan, Minnesota, a city in Scott County
- Jordan, Minneapolis, a neighborhood of Minneapolis, Minnesota
- Jordan, Montana
- Jordan, New York
- Jordan, North Carolina
- Jordan, Oregon
- Jordan, Wisconsin, a town
- Jordan, Portage County, Wisconsin, an unincorporated community

Elsewhere

- Germán Jordán Province, Bolivia
- Jordan, Guimaras, Philippines
- Jordan, Hong Kong
- Jordan (Neumark), Poland

Figure 2: Jordan's Disambiguation Page

Wikipedia allows editors to fill in information of different specificity, we would encounter the problem, missing links in current articles. Many approaches were successful adopted to find these links in mono-lingual environment. However, the pages in different languages are rarely linked except for *Interlanguage links*. This could pose serious difficulties to users who try to seek information or knowledge from different lingual source. Therefore, in CrossLink task, the system need to find possible anchors,a piece of text relevant to the topic, in source language and connect this anchor to corresponding pages in target languages [5]. Here, we propose a system to find cross-lingual related links by utilizing context-info similarity,described in Section 3.2 and graph-based like approach module to find outer linked pages in target language, described in Section 3.3.

3. METHOD AND SYSTEM DESCRIPTION

Cross-lingual link discovery aims to find potential links between documents in different languages. Methods for such a task actively recommends a set of words or phrases called *anchors* in a source document and suggests appropriate target pages to which those anchors can link. For the context of this report, we consider only English to Chinese link discovery.

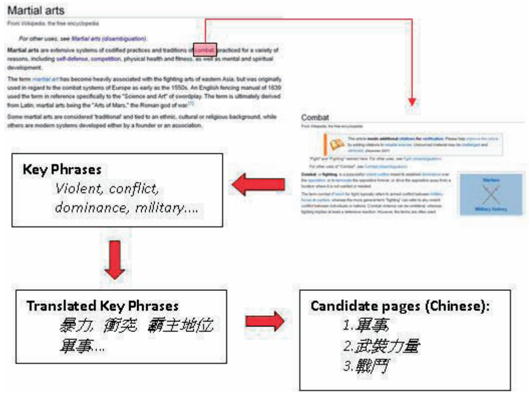
Our system divides CrossLink into several steps, shown in Figure 1. First of all, for a source page, we need an anchor tagger to find phrases or pieces of text appropriate for considering as anchors. Having decided candidate anchors, the next step is to link them to relevant pages in the source language (English in this case.) This step will later provide information for establishing cross-lingual links. However, there may be more than one possible and nontrivial choice that an anchor can link to. For instance, in Figure2, `Jordan`, may link to the page `Jordan (Country)` or `Michal Jordan` in different scenario. We assume that the context information would be quite different in different scenario. Hence, we need to utilize context information to disambiguate which pages are likely appropriate for a particular anchor.

After finding appropriate page for each anchor in source language , we need to find the translated link to the pages in target language, shown in Figure 3. First of all, we extract key phrases from those pages that the source document is linking to. Here, we utilize the anchors in current pages as key phrases. Next, we submit these key phrases to Google translator[4] to convert them into Chinese. These translated terms are then used as queries for retrieving Chinese relevant pages indexed by Lucene[5]. Finally, we use a graph-based like approach to re-rank relevant anchors and translated pages. Only top 250 anchors with corresponding Chinese pages are retained.

The operating environment of our system is as follows:

- OS:Windows Server 2003 (64-bit)
- CPU:Xeon E5310 1.60 GHZ
- Cores:8
- Memory:10 GB

[4] http://translate.google.com.tw
[5] http://lucene.apache.org



Martial arts

From Wikipedia, the free encyclopedia

For other uses, see Martial arts (disambiguation).

Martial arts are extensive systems of codified practices and traditions of combat, practiced for a variety of reasons, including self-defense, competition, physical health and fitness, as well as mental and spiritual development.

The term martial art has become heavily associated with the fighting arts of eastern Asia, but was originally used in regard to the combat systems of Europe as early as the 1550s. An English fencing manual of 1639 used the term in reference specifically to the "Science and Art" of swordplay. The term is ultimately derived from Latin, martial arts being the "Arts of Mars," the Roman god of war.[1]

Some martial arts are considered 'traditional' and tied to an ethnic, cultural or religious background, while others are modern systems developed either by a founder or an association.

Combat

Key Phrases
Violent, conflict, dominance, military....
Translated Key Phrases
暴力, 衝突 霸主地位 軍事...
Candidate pages (Chinese):
1.軍事
2.武裝力量
3.戰鬥

Figure 3: System Architecture

3.1 Anchor tagger

To begin CrossLink, we need to find words or phrases that are pertinent to the current document. In general, name entities and technical terms are ideal candidates to be marked as anchors. For example, consider an excerpt from page `Michael Jordan`: `Michael Jeffrey Jordan (born February 17, 1963) is a former American professional basketball player, active businessman, and majority owner of the Charlotte Bobcats.` In this case, the original Wikipedia page marks `American professional basketball player` and `Charlotte Bobcats` as anchors.

In this work, we only use a simple probabilistic measure [4] to evaluate whether a term or phrase should be selected as an anchor:

$$P(term) = \frac{freq(t_a)}{freq(t)} \tag{1}$$

where $freq(t_a)$ means the number of documents in which the term was tagged as anchors and $freq(t)$ is the number of documents in which the term has appeared. Then, we remove terms whose probability is lower than a threshold, *0.1*.

3.2 Measuring Context Similarity

Adafre and Rijke [1] propose the method to discover missing links by evaluating the concept of the refereed pages is similar to the current article or not. In this work, we utilize similar concept to find the links to pages that are closely related to the current document. Here, we suppose that the context in a linked page would be relevant to the current article. For instance, suppose that we have an anchor, `Jordan`, in the article `Chicago Bulls`. There are several candidate pages for which this anchor can link to, such as `Jordan (Country)` and `Michael Jordan`. Of course, the context of these two pages are quite different, and compared to current article, `Chicago Bulls`, the page `Michael Jordan` has many similar anchor texts like NBA or `Scottie Pippen`. Using this principle, we can measure the relatedness of current article, a and a particular page, p, as:

$$r_a(p) = \frac{|\mathbf{A} \cap \mathbf{P}|}{|\mathbf{A}|} \tag{2}$$

where $\mathbf{A}$ and $\mathbf{P}$ is the set of out-going links in article a and p. In this work, we only retain pages ranked top 5 for further computation. Next, we will utilize these pages to locate corresponding pages in target language.

3.3 Graph-ranking like ranked Approach

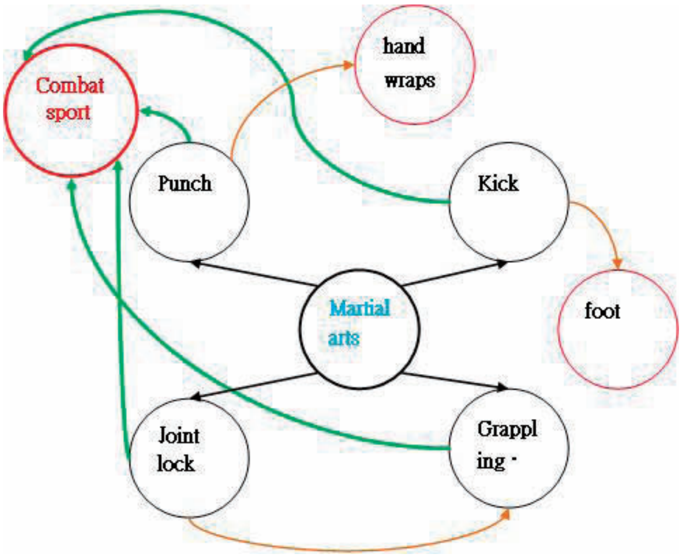


Figure 4: Material Art Linked Network

As mentioned previously, we hypothesize that the linked pages would have similar context information. From the process describe in Section 3.2, we create several links connecting pages in source language (English in this case.) For example, as shown in Figure 4, we can see the link structure centering at the article `Martial arts`. `Martial arts` has several outgoing links to the pages like `Punch`, `Jointlock`, `Grappling`, `Kick`. In these four different pages, they both link to the page `Combat sport` that could be considered as a related concept to `Martial art`. Therefore, if a candidate page that we think the source page, `Martial art`, could link to contains more occurrences of terms like *Combat sport*, the more likely that linking is appropriate. However, applying this idea in CrossLink requires more work, because we need to transform the context into the target language.

In Chinese Wikipedia, only a third of all pages contains interlanguage links[6] to English counterparts. This ratio may not allow us to establish a useful context by looking up only the interlanguage links. In order to solve this problem, we devise an approach to find relevant Chinese pages with Lucene and Google translate.

Based on the links in source language that we obtained in the last step, we want to transform those information into the context in the target language. For our purpose, we only translate the anchor texts gathered from pages obtained from last step that the source document might link to. Each of these translated terms is then submitted to Lucene for retrieving relevant Chinese pages. The retrieved Chinese pages are considered as candidates for which the original English anchor text may link to. Using the idea described above, we think that the tighter a Chinese page associate with English anchor texts and other Chinese pages, the more reliable that page may be. Thus, we need to find key terms or concepts that appeared in the Chinese pages. In Figure 5, we can find that the cross-lingual link structure at the article `Martial arts`. `Martial arts` has several outgoing links

[6] http://en.wikipedia.org/wiki/Help:Interlanguage_links

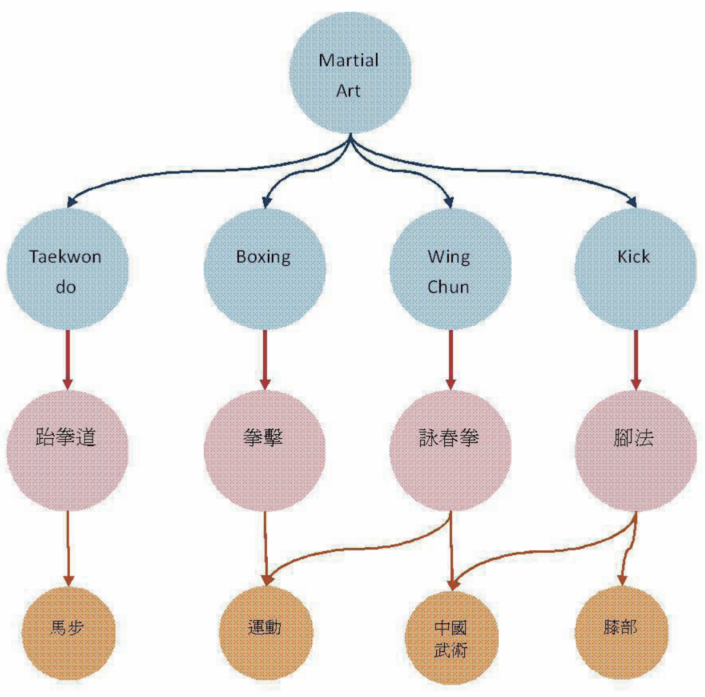


Figure 5: Material Art Cross-Lingual Linked Network

to the pages like `Taekwondo`, `Boxing`, `Wing Chun`, `Kick` and the interlanguage pages is 跆拳道, 拳擊, 詠春拳, 腳法. In these four outgoing links to the Chinese pages, we can find the terms 中國武術 and 運動 would appeared several times between these four pages. Here, we suppose the terms would be the concept terms or key terms relevant to the article `Martial Arts`. Therefore, We estimate the weight of each Chinese terms that appeared in the Chinese pages as anchors as follows:

$$GR(w) = \sum_{p \in \mathbf{P}} \begin{cases} o_s(p), & \text{if } binary(w,p) == 1 \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where $\mathbf{P}$ is the candidate Chinese pages and $binary(w,p)$ defined as follows:

$$binary(w,p) = \begin{cases} 1, & \text{if } w \text{ occurs in } d \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

Next, for each Chinese page, we sum up all weights of terms appeared in it as that page's weight, defined as follows:

$$W(p) = \sum_{w \in \mathbf{p}} GR(w) \tag{5}$$

In this work, we only keep the top 250 Chinese pages as our final output.

4. OFFICIAL RESULTS

The main metric to evaluate the performance of information retrieval is Mean Average Precision (MAP). Average precision is based on the whole list of documents returned by the system and emphasizes returning more relevant documents earlier. The Mean Average Precision is the mean value of the average precisions computed for each topic. Besides, *Precision-at-N* and *R-Prec*, are also good metrics which were provided by the evaluation result of NTCIR-9. *Precision-at-N* is the precision among the front of N anchors and R-Prec only considers the precision value returned by the system. There are two kinds of judgments: *Wikipedia Ground-Truth* and *manual assessment*. Table1 and 2 show the performance of our system and we also achieve 0.680 precision score at top 5 ranked anchors[5]. We submitted three CrossLink runs and focus on the runs as follows:

- IASL_E2C_01: a run using all features described in this paper.
- IASL_E2C_02: a run using all features described in this paper, only retained one English page linked from each anchor.
- IASL_E2C_03: a run without using *.Graph-ranking like ranked* Approach.

5. DISCUSSION

In this task, we only get the MAP score about 0.205 and we found several problems need to be solved. First is the out of vocabulary problem. In Section 3.1, we only consider the terms ever tagged as anchors. However, if the terms were tagged as anchors only in testing sets, we would miss these terms.

Run	MAP	R-Prec
IASL_E2C_01	0.225	0.347
IASL_E2C_02	0.214	0.335
IASL_E2C_03	0.211	0.337

Table 1: Evaluation Result By Wikipedia Ground-Truth

Run	MAP	R-Prec
IASL_E2C_01	0.205	0.308
IASL_E2C_02	0.200	0.312
IASL_E2C_03	0.194	0.301

Table 2: Evaluation Result By manually assessment

Second, in tagging module, we do not consider the types of anchors. For instance, the terms like a country name or a location name is easily tagged as anchors. However, the context in these outgoing pages would not be relevant to the article. But the ranking module described in Section 3.3 would easily influence on the English to Chinese anchors. In this task, we find that the terms like location and country name has a higher probability to tag as anchors. If we tag this terms, the ranking module would set the terms like 國家 or 人口 which is relevant to the concept *location* or *country*.

6. FUTURE WORKS

We still need to dissolve several issues. First of all, in anchor tagger module, we need to adopt a flexible mechanism to solve the out of vocabulary problem, like using SVM [2] or CRF [3] to tagged the terms as anchors or not. Second, in context-similarity module, we would try to adopt more information like category in Wikipedia to filter some anchors to avoid finding irrelevant terms. Finally, we need to ignore some terms like location or country into ranking module for reasons described in previous section.

7. ACKNOWLEDGMENTS

This research was supported in part by the National Science Council under grant NSC 100-2631-S-001 -001, and the research center for Humanities and Social Sciences under grant IIS-50-23.

8. REFERENCES

[1] S. F. Adafre and M. de Rijke. Discovering missing links in wikipedia. In *Proceedings of the 3rd international workshop on Link discovery*, LinkKDD '05, pages 90–97, New York, NY, USA, 2005. ACM.

[2] C.-C. Chang and C.-J. Lin. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2:27:1–27:27, May 2011.

[3] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[4] R. Mihalcea and A. Csomai. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 233–242, New York, NY, USA, 2007. ACM.

[5] L.-X. Tang, S. Geva, A. Trotman, Y. Xu, and K. Itakura. Overview of the ntcir-9 crosslink task: Cross-lingual link discovery. In *In Proceedings of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access*, 2011.