English-to-Traditional Chinese Cross-lingual Link Discovery in Articles with Wikipedia Corpus

Liang-Pu Chen[†*], Yu-Lun Shih[‡], Chien-Ting Chen[♭], Tsun Ku[†], Wen-Tai Hsieh[†], Hung-Sheng Chiu[†], Ren-Dar, Yang[†]

[†]IDEAS, Institute for Information Industry, Taiwan
[‡]CSIE, National Taipei Univeristy of Technology, Taiwan
[♭]ISA, National Tsing Hua Univeristy, Taiwan

[*]corresponding author

eit@iii.org.tw, t100598029@ntut.org.tw, s961441@gmail.com,
{cujing, wentai, bbchiu, rdyang}iii.org.tw

Abstract

In this paper, we design a processing flow to produce linked data in articles, providing anchor-based term's additional information and related terms in different languages (English to Chinese). Wikipedia has been a very important corpus and knowledge bank. Although Wikipedia describes itself not a dictionary or encyclopedia, it is if high potential values in applications and data mining researches. Link discovery is a useful IR application, based on Data Mining and NLP algorithms and has been used in several fields. According to the results of our experiment, this method does make the result has improved.

摘要

本篇論文中提出了一套自動化流程以發掘潛在的關鍵字連結，並且在找出文章關鍵字後能夠提供關鍵字於跨語言的相關資訊，而我們利用了Wikipedia做爲我們的知識庫，藉由Wikipedia的資料，系統能夠提供相關的關鍵字內容資訊，進而幫助使用者閱讀文章。論文中所提出的系統整合了相關的資訊檢索技術以及自然語言處理相關的演算法，以利於幫助我們進行關鍵字的識別以及相關的跨語言翻譯，同時系統整合了跨語連結發掘的技巧來幫助提供跨語言的關鍵字資訊。經過初步的實驗證實，相較於baseline方法，此方法確實能夠始數據有所提昇。

Keywords: Cross-lingual link discovery, Linked data, Wikipedia, Link Discovery
關鍵字: 跨語連結發掘, 資料連結, 維基百科, 連結發掘

1 Introduction

For our goal, we have to conquer some issues to find every potential linked data on articles. This paper focuses on Cross-lingual link discovery. Cross-lingual link discovery contains a lot of important tasks of NLP(Natural Language Processing) such as WSD(Word Sense Disambiguation) [1], NED(Named Entities Disambiguation) [2] or Machine Translation. The cross-lingual links in the Wikipedia[1] are established by the human contributors, and not all Wikipedia Pages have cross lingual links because no human editors established these links yet. Thus, when one visits English Wikipedia page which describes some special information, users cannot find any cross lingual link to visit the Wikipedia page whose language is the same as the user's mother tongue. This problem has been raised by many recent studies [3,4] , and recovering these missing links between two languages is the main goal of the CLLD (Cross-Lingual Link Discovery). In this paper, we propose a system which can automatically help users to tag potential links in

[1]http://wikipedia.org

their articles, and automatically find out the cross language link of the tag based on Wikipedia cross language links. As for cross lingual link discovery, our system is able to find the missing links between two related Wikipedia pages in two different language systems by exploiting and extracting data from Wikipedia dump files in two languages. In addition, we use two additional translation mechanisms to help find out the corresponding cross lingual translation , one is the *Pattern Translate* , the other one is *Google Translate*[2]. We further integrate the *Lucene*[3] software package to deal with the ambiguous phases in the articles. In order to find out the missing links between two pages, and automatically tagged this cross language link in users' articles.

The remainder of this paper is organized as follows: First, we described corresponding background of Wikipedia and cross-lingual link discovery in Section 2. In Section 3,The proposed WSD method and translation mechanism will be described in detail. Finally, the experiment and conclusion will be discussed in Section 4.

2 Background

2.1 Wikipedia

Wikipedia is a free, collaboratively edited, and multilingual Internet encyclopedia supported by the non-profit Wikimedia Foundation[4]. Recently, many researchers focus on developing data mining applications with Wikipedia's large-scale collaborative user data. Although Wikipedia describes itself not a dictionary, textbook or encyclopedia, exploiting its characteristics to develop new services is regarded as a promising method on auto text explanation.

One of the special feature of Wikipedia is that it contains many hypertext links to help users easily retrieve the information they need. These hypertext links might be embedded within the text content under the corresponding pages, and each of these links is linking to other pages related with different terms. Obviously, information flow is thus being traversed very easy and smoothing when the hypertext links are extensively tagged. Unfortunately, the hypertext links between different languages are mostly not being tagged because of the hypertext link is generated by human contributor, mostly monolingual ones. To solve this problem, we design a process flow trying to make it more completely.

2.2 Cross-lingual link discovery

The goal of cross-lingual link discovery(CLLD) is trying to find the potential links that are missing between the two different languages. There are three main challenges for the system to overcome. First, the system providing solution on CLLD can proactively recommends a set of words which called anchors. The set of words have higher chances to have their corresponding cross lingual links than other words in the same article. For example, considering different cases as following:

1. Let's go *dutch*.
2. A *Dutch* auction is a type of auction that starts with a high bid.

The system must determine the boundaries between anchor and rest of words, considering the first case above, the word "dutch" is meaning to share the money on something instead of meaning some behavior or something related to the country "Holland". In other words, the

[2]http://translate.google.com
[3]http://lucene.apache.org
[4]http://en.wikipedia.org/wiki/Wikipedia

word “dutch” should not be chosen as an anchor here and choosing the phase of “go dutch” is more significant. Considering the second case above, the word “Dutch auction” is an appropriate anchor rather than “Dutch”.

After the system identifies these anchors, there must exist many highly ambiguous cases in these anchors and this is the second challenge of CLLD, for example, the anchor *Apple* can be refer to the link which is related with *Apple(Computer Manufacturer)*, or the link which is related to *Apple(Fruit)*. The system must be able to choosing the most related corresponding links and also ensure the correctness of link discovery.

Once the system can return the most related links of each anchor, there is only one more problem need to solve. In the end of the CLLD flow, the system have to automatically discover the cross-lingual link based on the anchors which generated from previous steps. The system can just use simple parser or crawler to check the content of corresponding wikipedia page or combines some different mechanism to increase the accuracy of link discovery. In this paper, we implement these CLLD steps to help us find the corresponding cross-lingual links and we focus on anchor disambiguation and cross-lingual link discovery, which are both described in Section 3.

3 Method and System Description

In English-to-Chinese cross-lingual link discovery, the goal is to find every potential links in documents. At first, the system searches out potential terms as candidate terms. Overlapping problem happens in this stage, and adequate candidate term selection is required. We propose an *similarity-scoring* formula to calculate score of relevance. When candidate terms are selected, relevant pages in Chinese Wikipedia need to be linked with these terms. There are some cross-lingual articles in Wikipedia; however, many more links are still missed. (eg. *“Hundred Schools of Thought” with “*諸子百家*”*).

3.1 Candidates finding

To find cross-lingual links in a language, every potential term or phrase is to be listed in the beginning. Here we adopt n-gram tokenizer [5] and Maximum Matching algorithm [6] to segment. For example, assume a sentence “Magic Johnson is one of the best basketball player in NBA”, in our method , our approach will take “Magic Johnson” as an anchor rather than “Magic” or “John”. The system will first examine the longer term in the sentence and exploit the Wikipedia as a anchor look-up table to check whether this long term is meaningful or not.

3.2 Anchor decision

Many terms in Wikipedia have the same title but different meanings depending on their occurrences in contexts. To address this problem, Wikipedia has already define it as “Disambiguation”. In our system, we use redirect page, providing disambiguation information and candidate terms, to analysis and select one from terms for users by this system. For instance, a term “Virus” is shown in “A virus is a parasitic agent that is smaller than a bacterium and that can only reproduce after infecting a host cell.” and “Virus (clothing), an Israeli clothing brand”...etc. It indicates users may look out the clothing brand but Wikipedia gives him a virus’ definition in biology domain.

$$SimilarityScore(D_i, D_j) = \frac{TermRecog(D_i) \bigcap TermRecog(D_j)}{TermRecog(D_i) \bigcup TermRecog(D_j)} \quad (1)$$

Figure 1: Processing flow of our system.

$$Anchor = max(SimilarityScore(D_{current}, D_i)), \forall i \in candidates \quad (2)$$

In our work, we design a content-aware approach to perform auto selection among disambiguation terms. Our design principle is to analyze the input article, especially the source of terms, and use full-featured text search engine with a prepared index file. If a term has the disambiguation property, the system will extract the features from article and search the existed index to decide which term is more likely to the source article.

3.3 English-Chinese Link Discovery

In this section, we describe how we translate the anchor first and than how we find the cross lingual Wikipedia link after the translation. There are two main approaches of the translation mechanism, namely ***Cross-Lingual Link Dictionary and Google Translate***. We first use a ***Cross-Lingual Link Dictionary*** as the translation scheme, once if ***Cross-Lingual Link Dictionary*** can not provide any corresponding translation, ***Google Translate*** is then used by the system to discover the corresponding translation from the online *Machine Translation mechanism*. Google Translate is a state-of-the-art online commercial machine translation scheme, and it is exploited by our system to trying find out some possible translation when there doesn't have any corresponding translation which can be provided by the ***Cross-Lingual Link Dictionary***. With the support by the Google Translate, the system can provide higher translation coverage compared to using Cross-Lingual Link Dictionary only. We will describe the detail about the two translation mechanisms below and will also discuss the missing link recovery

approach in the end of this section.

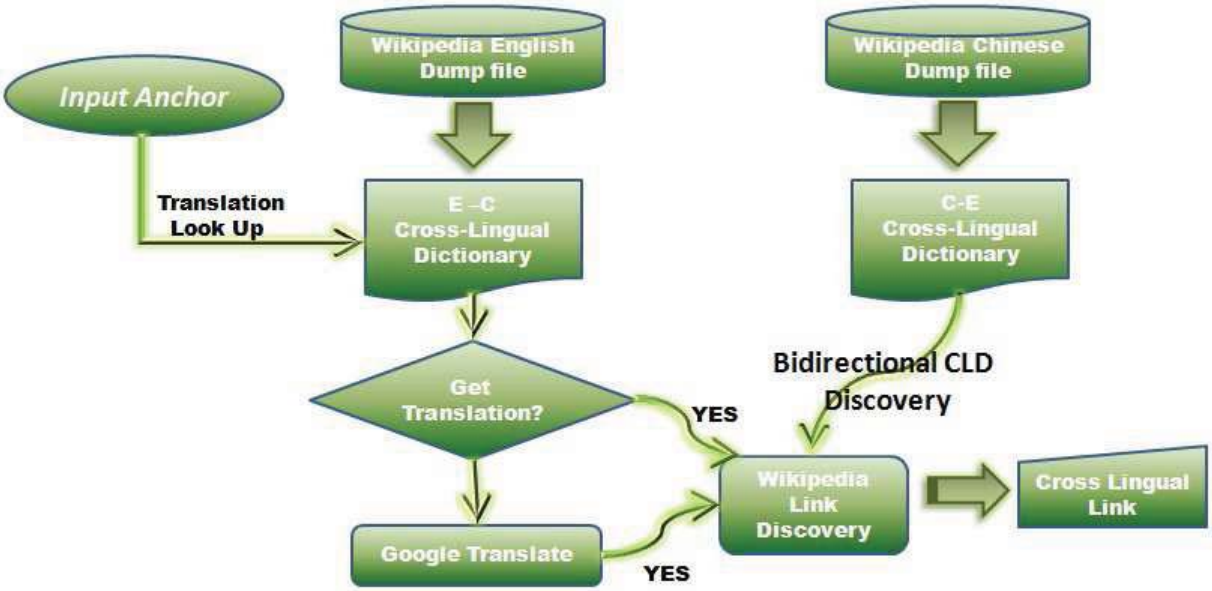

Figure 2: Flow of our English-Chinese Link Discovery system.

3.4 Anchor Translation Mechanisms and Missing Link Recovery

3.4.1 Google Translate

We first describe the Google Translate here because we are going to introduce the translation and missing link recovery within Cross-Lingual Dictionary in the end of this section together.

Google Translate has been a famous automatic translation mechanism, one distinguishing feature of this online translator is that it enables users to choose different languages that users want to translate. As for whole sentence translations, the users have a chance to modify the translation sentence once they find the output translation inadequate. As Google collects enough user data of modifying the translation sentence, Google Translator gets higher translation accuracy.

Although Google Translate has such special characteristic, it can not providing good accuracy at Anchor translation [7]. However, there is a special characteristic of Google Translate; that is, it can provide more possible translation candidates than previous methods such like Cross-Lingual Link Dictionary. The reason is that Google Translate is tends to adopt a best-effort approach, it aims to provide many translation candidates which enable users to understand what the untranslated sentence might be supposed to mean.

As a result, we put the lowest translation priority in Google Translate, namely, once the previous method(*Cross-Lingual Dictionary*) can not find out any possible translation candidates, we will try to get some translation suggested from Google Translate. The main reason is just what we describe above, we want to take a chance to find out the corresponding translation when we do not have any other translation candidate, only to use some anchor translation from Google Translate to find out the corresponding cross-language links.

For example, in our Cross-Lingual Link Dictionary, it does not contain the Chinese Translation of "***Freeway***". However, Google Translate can provide some useful Chinese translation like "**高速公路**", thus we can find the corresponding link of Chinese article page of Wikipedia page at "`http://zh.wikipedia.org/wiki/`".

3.4.2 Cross-Lingual Link Dictionary

Wikipedia provides a well formatted dump file for all languages. As a result, we can get the chinese translation from the english dump files and vise-versa. We exploit this property to construct both Chinese-English bilingual link dictionary and an English-Chinese bilingual link dictionary. Furthermore, once the translation in the dictionary has be found, there is a high probability that we can directly discover the link by adding the translated anchor after the specific wikipedia URL(e.g. `http://en.wikipedia.org/wiki/Computer_accessibility`), both in English and Chinese. We refer these two dictionaries as the translation dictionaries, one is the *English to Chinese (E-C) translation dictionary* and the other one is *Chinese to English (C-E) translation dictionary*. Once we use these two bilingual dictionaries as translation dictionaries, in our case, English-to-Chinese vise versa,we can have a chance to retrieve the link informations **bidirectional**. The reason is that we have noticed that links for Chinese-to-English are more than English-to-Chinese, because many Chinese editors will add English link for annotation or reference.

On link discovery part, we find out that some links may be missing in one translation dictionary, such as the term "Flag of Republic of China" is not able to found any corresponding Chinese translation in E-C translation dictionary. However, we can find the corresponding english translation of chinese term "諸子百家" in the C-E translation dictionary, which is the "*Hundred Schools of Thought*".

There is an additional problem about the English-Chinese dictionary with the Wikipedia disambiguation page. If the anchor which exist in the English-Chinese dictionary is a title of the Wikipedia disambiguation page, then we can not directly get the Chinese translation from the page content of disambiguation page. The reason is that a Wikipedia disambiguation page only contains the possible candidates that are referring to this title.

Fortunately, Wikipedia have a complete dump file format and it provide the *redirect information* of the disambiguation page. Therefore, we can using the redirect link information to find out the corresponding Chinese translation. The problem may also occur at Chinese Wikipedia disambiguation page, and it can be also solved by redirection information.

4 Results and Discussion

We use four articles as evaluation to see the performance of cross-lingual discovery, we first randomly choose four Bilingual news article from Yahoo! News, all terms in the Chinese articles are tagged by two human experts to generate correct answers. We apply two methods, the first method is tagging the English articles with English Wikipedia entries by means of long-term-first algorithm. Those tagged terms are then directly transformed into Chinese Wikipedia entries by original anchored links; the second method is to implement our proposed method, we then compare these two methods to see the coverage rates. As Figure 4 shows, the experiment result shows that our proposed method has 8% coverage rates higher than the that of direct anchor transformation method.

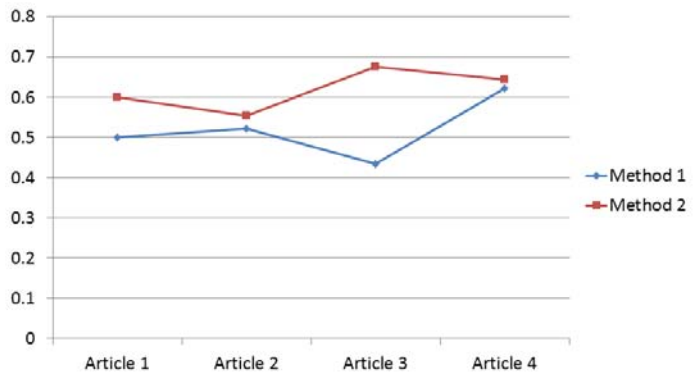

Figure 3: Results of English to Chinese link discovery.

5 Conclusion

In conclusion, we present a system to find potential cross-lingual linked data on articles, trying to discover miss cross-lingual links. The main contribution of our proposed method includes finding anchor and discovering missing cross-lingual links. We have successfully designed a practical system to perform tagging task on real-world articles. and proved that maximum match algorithm has a better performance than the original Wikipedia anchor links transformation. However, there are still issued to be improved for future work. First, the precision of WSD is still low, and second, we can apply machine learning approaches in our method, in which we are convinced that our proposed method might have higher performance in the future.

References

[1] Mihalcea and Csomai, "Wikify! linking documents to encyclopedic knowledge", in *sixteenth ACM conference*, 2007.

[2] Bunescu and Pasca, "Using encyclopedic knowledge for named entity disambiguation", in *EACL*, 2006.

[3] J. Kim and I. Gurevych, "Ukp at crosslink: Anchor text translation for cross-lingual link discovery", in *NTCIR-9*, 2011.

[4] A.F.V. Nastase and M. Strube, "Hits'graph-based system at the ntcir-9 cross-lingual link discovery task", in *NTCIR-9*, 2011.

[5] W. B. Cavnar and J. M. Trenkle., "N-gram based text categorization", in *Proceeding of the Symposium on Document Analysis and Information Retrieval*. University of Nevada, Las Vegas, 1994, pp. 161–175.

[6] Y. Shiloach Amos Israeli, "An improved parallel algorithm for maximal matching", in *Information Processing Letters*, 1986, vol. 22, pp. 57–60.

[7] Yu-Chun Wang Richard Tzong-Han Tsai Liang-Pu Chen, Chen-Ting Chen, "Exploit wikipedia and name-entity pattern as translation method on chinene-korean cross language multimedia information retrival", in *International Conference on Digital Contents*, 2009.