Cross-lingual Link Discovery Based on CRF Model for NTCIR-10 CrossLink

Liang-Pu Chen
IDEAS, Institute for Information Industry, Taiwan
eit@iii.org.tw

Yu-Lun Shih
CSIE, National Taipei Univeristy of Technology, Taiwan
t100598029@ntut.org.tw

Chien-Ting Chen
ISA, National Tsing Hua Univeristy, Taiwan
s961441@gmail.com

Ping-Che Yang
IDEAS, Institute for Information Industry, Taiwan
maciaclark@iii.org.tw

Hung-Sheng Chiu
IDEAS, Institute for Information Industry, Taiwan
bbchiu@iii.org.tw

Ren-Dar, Yang
IDEAS, Institute for Information Industry, Taiwan
rdyang@iii.org.tw

ABSTRACT

This paper described our participation in the NTCIR-10 Cross-lingual Link Discovery Task of Chinese-to-English(C2E). The task focuses on making sutiable links on terms between Chinese/Japanese/Korean lingual Wikipedia articles and English Wikipedia articles. In this event, we proposed a method on Chinese-to-English subtask. The method that we proposed have two stage. We divides this task into "Anchor Recognition' ' and "CrossLink' ' . The first one, we use conditional random field in machine learning method to recognize every potential anchors which could be linking to a article in target language. The second, we try to find candidate links of these anchors and then doing disambiguous with them. According to the offical result, our system achieved LMAP score 0.072 when evaluating with Wikipedia ground-truth, and 0.027 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

CRF model, Cross-lingual link discovery, Named-entity recognition, Word-sense disambiguation

Team Name

III-IDEAS

Subtasks

Chinese to English Crosslink

1. INTRODUCTION

This paper is for participating NTCIR-10 CrossLingual Link Discovery Task of Chinese-to-English(C2E). Our goal is to make suitable anchors on terms, and to link to related description page. Because one term may have many different meanings, how to choose the appropriate page is a big issue. Furthermore, anchors link to another language description page(CrossLingual). For making cross-lingual links, we have to find out terms mapping between two languages, and evaluate the more related result of ambiguous meanings.

In the task of CrossLink, the papers in Wikipedia are used as database. Wikipedia is a free multi-lingual encyclopedia. All the information is co-edited by the users worldwide, and it contains articles in various languages. Through the periodical published dumpfile, we can acquire different version of constructional paper information in different languages. At the same time, each version contains the notes of that language and the linking information. Our job is to find unique meaningful sentences and note the cross-language explanation link from original article. out competition this times contains three independent subtasks as follows:

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

Each items is expected to be found a unique meaningful substantial noun in each language and give a effective link to target language database.

In this project, we join the part Chinese to English (C2E) under CrossLingual Link Discover. The goal is to add Anchor link to related pages to certain nouns in the article. Furthermore, link to cross-language explanation page, and where to add Anchor has to be solved here. When there are multiple choices to link to the explanation page, the most likely item should be raised.

Table 1: Wikipedia document collections.

Language	Number of Article	File Size	Dump Date
English	3,581,772	33G	04/01/2012
Chinese	432,988	3.7G	11/01/2012

The details will be introduced later in the following chapters. Chapter 2 is about the background of the task. Chapter 3 introduces research method and system structure. Chapter 4 and 5 displays the result and discussion. Finally, chapter 6 and 7 are about future work and acknowledgement.

2. BACKGROUND

2.1 Wikipedia

Wikipedia is a free, collaboratively edited, and multilingual Internet encyclopedia supported by the non-profit Wikimedia Foundation[1]. 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)[7, 3] 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 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.

There are some reseachs about link discovery between documents in different languages. Sorg & Cimiano[6] aims language links problem between German and English Wikipedia approach based-on classification. In the NTICR-9, Nastase & Strube[5] developed a system for link discovery using a graph-based method for disambiguation and achieved good results. In this paper, we focus on anchor recognition using CRF approach.

3. METHODOLOGY

The aim of this section is suggesting good links in Chinese documents to English ones. In this paper, we design our system with many components: Anchor: (1)Process CRF training data, (2)CRF Training, CrossLink: (3)Translation and (5)Disambigus. Firstly, we use the document collection, NTCIR provided, as our training set. And next we would do some pre-processing for these raw data for we using CRF for marking the right anchor. After that data transformer to the right format, we use a specific CRF training pattern to train the data. Up to now, we have a CRF model for marking the good anchors in Chinese documents. On the other, we have to mapping these anchors to another language ones(English for this time). Our system specifications:

- OS: FreeBSD 9.0 amd64
- CPU: Xeon E5630 2.53GHz *2 (8 cores)
- RAM: 32GB

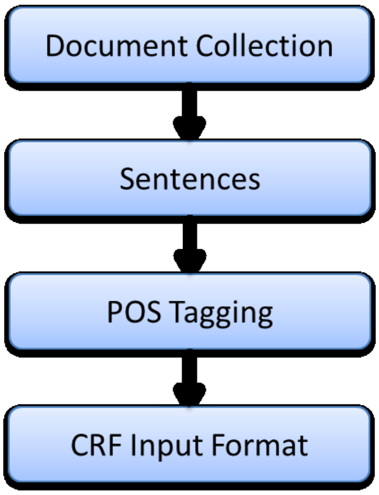


Figure 1: Pre-processing of CRF training.

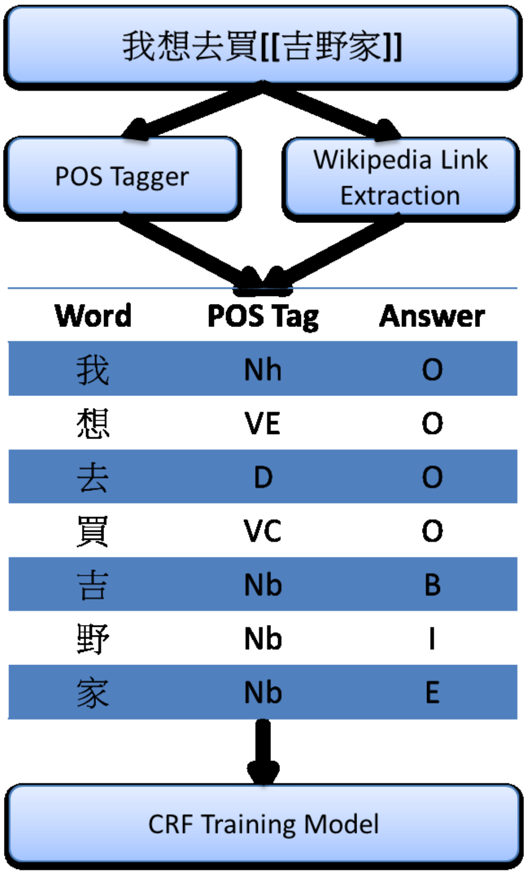
3.1 Anchor Recognition

Process CRF Training data

In this section, we described how we try to find every potential link by machine learning methods. In the traditional NER(Named Entities Recognition), SVM[2] has been used generally. However, here we try to employ CRF[4] as a new experiment. First we extract all sentence which contains article links from chinese corpus, and use them as our training set. Also, we use each word's part-of-speech as feature to conduct the training.

By document collection, the proposed method roughly diviedes articles to sentences using stopword list. In addtion,

[1] http://en.wikipedia.org/wiki/Wikipedia



Word	POS Tag	Answer
我	Nh	O
想	VE	O
去	D	O
買	VC	O
吉	Nb	B
野	Nb	I
家	Nb	E

Figure 2: CRF model training.

Table 2: Example of CRF input data.

Term	POS	Anchor Tag
今	Nd	x
天	Na	x
天	Na	b
氣	D	e
真	VH	o
好	VH	o

Table 3: Template of CRF model training.

```
(x)
# Unigram
U00:%x[-2,0]
U01:%x[-1,0]
U02:%x[0,0]
U03:%x[1,0]
U04:%x[2,0]
U05:%x[-1,0]/%x[0,0]
U06:%x[0,0]/%x[1,0]

U10:%x[-2,1]
U11:%x[-1,1]
U12:%x[0,1]
U13:%x[1,1]
U14:%x[2,1]
U15:%x[-2,1]/%x[-1,1]
U16:%x[-1,1]/%x[0,1]
U17:%x[0,1]/%x[1,1]
U18:%x[1,1]/%x[2,1]
```

keep anchor information from raw data, which is tagging by Wikipedia editors, for anwser column of CRF training. In our proposed method, CRF input data contains three columns: word, part-of-speech and answer. The word column is coming from previous processed sentences, and column of part-of-speech is using PosTag tool such as CKIP[2] , LTP[3], ICTCLAS[4]. Finally, the column of answer.

The start of anchor marks with B(Begin), the inter of anchor marks with I(Inter) ,the end of anchor marks with E(End) and others marks with O(Other). For this time, the proposed workflow is trying to build a CRF language model to produces well anchors in non edited articles.

In terms of CRF template, we used (x) as the training template in order to diagnose the location of anchor depending on the contextual clues.

3.2 Crosslink

So far, the research method has came out with a CRF model. Next, the cross lingual link has to be found. In Wikipedia, each page contains its own corresponding cross lingual link which can be edited artificially. By utilizing this feature, we use Chinese articles as a start point, and get cross lingual link through each of their corresponding English page. By doing this, we can find cross lingual link of every anchor in every article. However, this might cause the problem of Disambiguous. The solution for the Disambiguous problem will be explained in the Disambiguous Section.

3.2.1 *Translation*

According to Document Collection, each corresponding cross lingual link information can be extracted. For example, the manual cross lingual link of Apple Inc’ s company page contains Apple_Inc. (English). Therefore, we established the mapping table based on these information and cross lingual link works based the mapping table.

3.2.2 *Disambiguous*

From the previous section, the basic cross lingual link was found. However, sometimes we might come across situation like this: a word implies multiple meanings. The situation called Disambiguous. For example, apple can be referred as fruit, Apple Inc and Apple Bank and so on. And that is why “How to tag accurate cross lingual link on anchor?” is a challenging work. Adafre and Rijke[1] has proposed the method to discover missing links by evaluating the concept of the refered page is similar to the current article or not. In this research, our strategy is to compare the similarity of two articles. Two articles which indicates the original article and the target article candidate individually. According to algorithm, we calculate the similarity of any two of articles and make the ranking. The top 5 will be consider as possible cross lingual links.

As to similarity algorithm, our method is first, select any

[2] http://ckipsvr.iis.sinica.edu.tw/
[3] http://ir.hit.edu.cn/ltp/
[4] http://ictclas.nlpir.org/

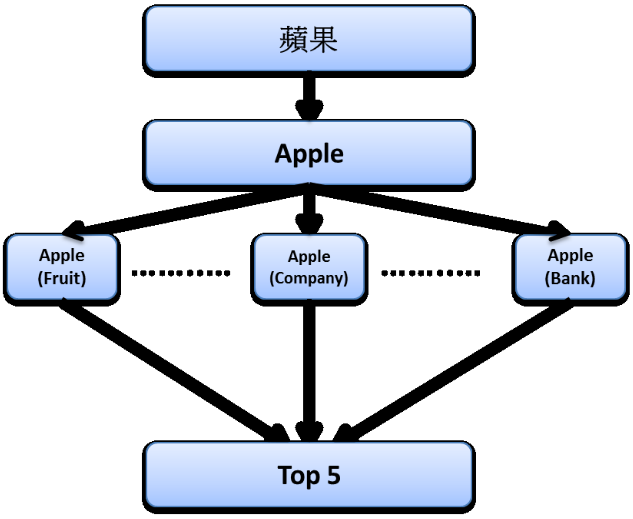


Figure 3: Architecture of CrossLink module.

two articles; second, extract the Named Entities; third, compare the Named Entities of two articles and mark those in common word of named-entities. By applying this concept, noises which affected results can be filtered out from the original data; in the same time, some possible meaningless words such as verb and adjectives can be filtered out from the data, too.

$$SS(D_i, D_j) = \frac{TR(D_i) \bigcap TR(D_j)}{TR(D_i) \bigcup TR(D_j)} \tag{1}$$

$$SS : SimilarityScore, TR : TermRecog$$

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

4. OFFICIAL RESULTS

The results that official released are evaluated at two levels (file-to-file and anchor-to-file); and system performance is evaluated with metrics: Link Mean Average Precision (LMAP), R-Prec and Precision-at-N (P@N). In the F2F (file-to-file) evaluation, performance is measured based on the ideology of what other relevant articles can be recommend without needing specifying relevant anchors. In A2F (anchor-to-file) evaluation, the relevance of specified anchors must however be considered. Precision and recall are the two key underlying fractions of the three system evaluation metrics (LMAP, P@N, and R-Prec). They are computed for each topic separately, and have to be treated differently for the different evaluation level (F2F or A2F). Precision-at-N is the precision among the front of N anchors and R-Prec only considers the precision value return by the system. There are two kinds of judgements: *Wikipedia ground-truth* and *manual assessment*. We submitted three runs as follows:

- Run 01 (III_C2E_A2F_01_PNM) : a run using all features described in this paper.
- Run 02 (III_C2E_A2F_02_PNM) : a run using all features described in this paper.
- Run 03 (III_C2E_A2F_03_PNM) : a run without using CRF model and alternated by Maximum Matching algorithm[8]

Table 4: CJK2E F2F evaluation with Wikipedia ground-truth: LMAP, R-PREC

Run-ID	LMAP	R-Prec
III_C2E_A2F_01_PNM	0.072	0.172
III_C2E_A2F_02_PNM	0.071	0.133
III_C2E_A2F_03_PNM	0.032	0.091

Table 5: CJK2E F2F evaluation with manual assessment results: LMAP, R-PREC

Run-ID	LMAP	R-Prec
III_C2E_A2F_01_PNM	0.011	0.061
III_C2E_A2F_02_PNM	0.027	0.090
III_C2E_A2F_03_PNM	0.009	0.037

Table 6: F2F evaluation with Wikipedia ground-truth: Precision-at-N (Chinese-to-English)

Run-ID	P5	P10	P20	P30	P50	P250
Run 01	0.272	0.272	0.254	0.227	0.184	0.043
Run 02	0.128	0.156	0.154	0.144	0.126	0.077
Run 03	0.168	0.188	0.158	0.131	0.103	0.022

5. CONCLUSIONS

From the result, it is found that the performance of C2E is clear; however, in the same time, it also found that some unsolved problems. First, in terms of anchor recognition, the method we used did not filter out special terms such as names of people and places; instead, it merely applied POS features to class. Second, since the current Answer Column of CRF model relies on links of Document Collection which edited by Wikipedia editors as answers, in this case, if a word has never been marked as a link by editors, the word could not been recognized by us as the result. Third, regarding to the translation, since the method merely adopt the mapping table to translate, situation such as “missing out local terms” or “disappearing on the mapping table” might happen. Finally, as to part of resolving WSD, we have compared the similarity of all articles’ words and links and we consider that the concept is feasible; however, the lack of data and the omitting of filtering procedure have caused the anchor we selected mixed with some common words such as “Father” , “English” and etc. Since these words could be found commonly in all categories of articles and therefore, the result has no actual help in terms of comparing the similarity. This is one of reasons of the decline in accuracy.

6. FUTURE WORKS

Currently, systems like this work is still immature. Dividing the system into three stages: tag anchor, crosslink candidate and WSD, it is found that the first and the third stage have relatively larger room to be improved. As to anchor, we can focus on improving the method of CRF model and considering to add more features or improve the accuracy of

Table 7: F2F evaluation with manual assessment results: Precision-at-N (Chinese-to-English)

Run-ID	P5	P10	P20	P30	P50	P250
Run 01	0.072	0.090	0.108	0.104	0.089	0.022
Run 02	0.056	0.056	0.096	0.105	0.104	0.077
Run 03	0.040	0.084	0.076	0.072	0.058	0.014

POS. In terms of WSD, the current method based merely on comparing the similarity of original article and target article and then calculate the degree of similarity between article of links. However, this method is apparently unsuitable to short articles and therefore, it lays a task about how to improve the performance. Besides above issues, back to the original data, since the dumpfile of Wikipedia contains extremely large amount of noises, try to figure out how to filter out these noises is also an important task.

7. ACKNOWLEDGMENTS

This study is conducted under the “Digital Convergence Service Open Platform” of the Institute for Information Industry which is subsidized by the Ministry of Economy Affairs of the Republic of China.

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*, pages 90–97. ACM, 2005.

[2] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[3] W. C. D. Huang, S. Geva, and A. Trotman. Overview of the inex 2009 link the wiki track. In *Focused Retrieval and Evaluation*, pages 312–323. Springer, 2010.

[4] J. Lafferty, A. McCallum, and F. C. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[5] A. F. V. Nastase and M. Strube. Hits¡¦ graph-based system at the ntcir-9 cross-lingual link discovery task, 2011.

[6] P. Sorg and P. Cimiano. Enriching the crosslingual link structure of wikipedia-a classification-based approach. In *Proceedings of the AAAI 2008 Workshop on Wikipedia and Artifical Intelligence*, pages 49–54, 2008.

[7] L.-X. Tang, I.-S. Kang, F. Kimura, Y.-H. Lee, A. Trotman, S. Geva, and Y. Xu. Overview of the ntcir-10 crosslink task: Cross-lingual link discovery task. *Proceedings of NTCIR-10*, 2013.

[8] P.-k. Wong and C. Chan. Chinese word segmentation based on maximum matching and word binding force. In *Proceedings of the 16th conference on Computational linguistics-Volume 1*, pages 200–203. Association for Computational Linguistics, 1996.