財團法人資訊工業策進會
INSTITUTE FOR INFORMATION INDUSTRY

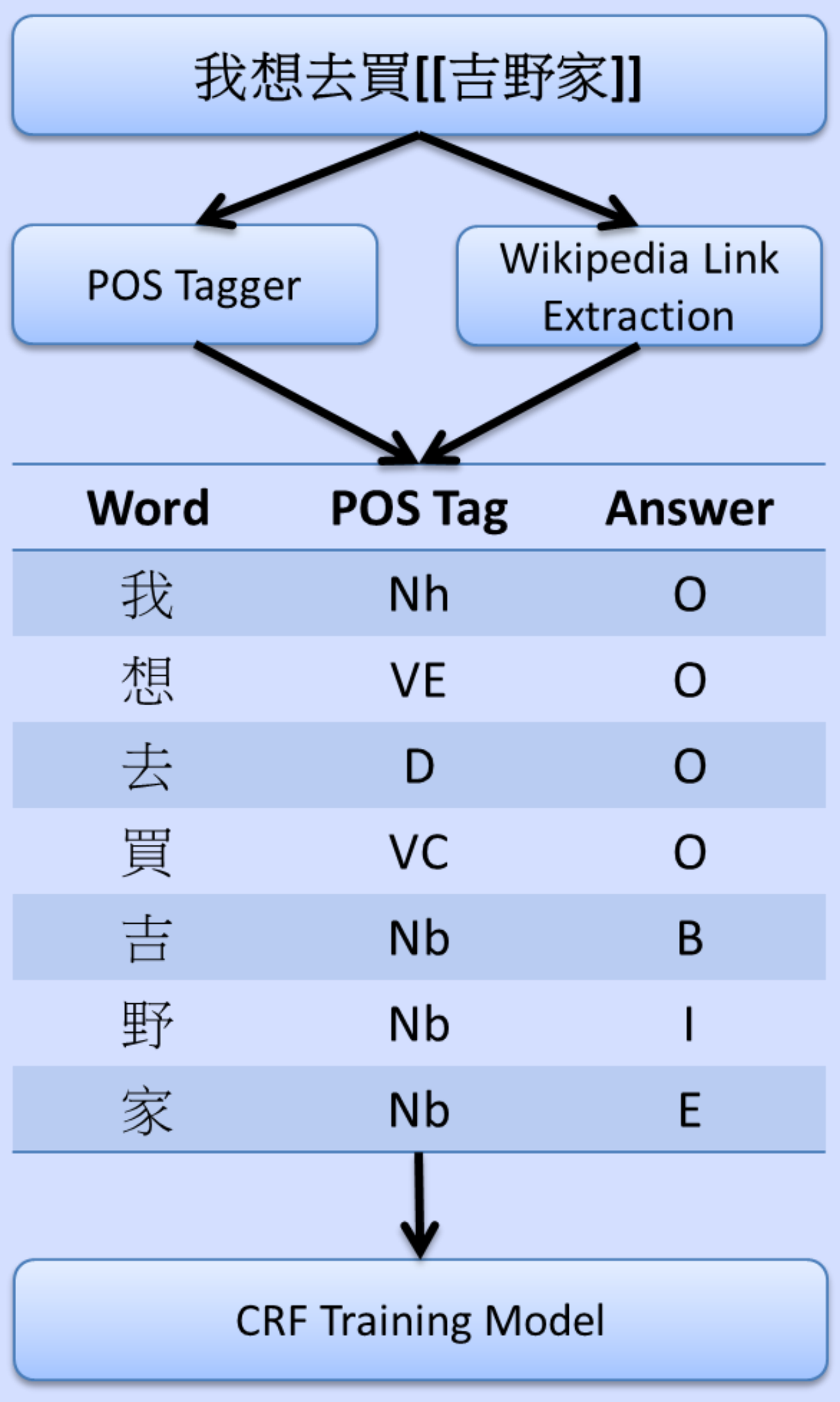
NTCIR-10 CrossLink

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

Liang-Pu Chen†*, Yu-Lun Shih‡, Chien-Ting Chen ♭ , Ping-Che Yang†, 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,
{maciaclack, bbchiu, 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 suitable 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 official result, our system achieved LMAP score 0.072 when evaluating with Wikipedia ground-truth, and 0.027 with manual assessment.



我想去買[[吉野家]]

POS Tagger

Wikipedia Link Extraction

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

CRF Training Model

Term	POS	Anchor Tag
今	Nd	X
天	Na	X
天	Na	B
氣	D	E
真	VH	O
好	VH	O

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]

Method

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).

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

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

Official Results

Table 1: Wikipedia groundtruth: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 2: 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 3: Wikipedia ground-truth results: Precision-at-N

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	0.141	0.103	0.022

Table 4: Manual assessment results: Precision-at-N

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

Conclusion

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.

IDEAS, Institute for Information Industry