Generating Patterns for Extracting Chinese-Korean Named Entity Translations from the Web

Chih-Hao Yeh[†], Wei-Chi Tsai[†], Yu-Chun Wang[‡], Richard Tzong-Han Tsai[†*]

[†]Department of Computer Science and Engineering, Yuan Ze University, Taiwan
[‡]Department of Electrical Engineering, National Taiwan Univeristy, Taiwan

[*]corresponding author

{s941539, s941537}@mail.yzu.edu.tw
r95921024@ntu.edu.tw
thtsai@saturn.cse.yzu.edu.tw

Abstract

One of the main difficulties in Chinese-Korean cross-language information retrieval is to translate named entities (NE) in queries. Unlike common words, most NE's are not found in bilingual dictionaries. This paper presents a pattern-based method of finding NE translations online. The most important feature of our system is that patterns are generated and weighed automatically, saving considerable human effort. Our experimental data consists of 160 Chinese-Korean NE pairs selected from Wikipedia in five domains. Our approach can achieve a very high MAP of 0.84, which demonstrates our system's practicability.

摘要

中韓跨語檢索上的困難之處，即在於query term中的專有名詞，由於人類語言的變化無法如同大多數的一般名詞一樣，能夠在雙語辭典中找到其對應的翻譯詞。本論文提出一種基於翻譯模板在Web中尋找翻譯詞的方式。由於Web的資料幾乎涵蓋人類到目前爲止的所有知識，且會隨時更新，因此能確保找到翻譯詞的recall。本研究的一大特點在於，所有用於擷取翻譯詞的模板，均爲自動生成，因此不需耗費大量人力來建構。此外，我們會利用訓練資料集來評估各模板的權重，藉以給與各候選詞適當的信心值。我們採用維基百科的中韓專有名詞pair做爲本方法所需之訓練集與測試集。經實驗過後，我們的方法可以達到MAP 0.84的高分，證明本論文提出方法的實用性。

Keywords: Chinese-Korean named entity translation, the Web, pattern
關鍵詞：中韓專有名詞翻譯, 網路語料, 模板

1 Introduction

In recent years, South Korean's entertainment industry has established itself as one of the most important emerging markets on the planet. In 2006, South Korean programs on Chinese government TV networks accounted for more than all other foreign programs combined [1]. South Korean actors such as Lee Young-ae (이영애, 李英愛), Bae Yong Joon (배용준, 裴勇俊), Rain, and Song Hye Gyo (송혜교, 宋慧喬) became very popular superstars in the great China area, making text and multimedia information related to them turned hot. Such information is firstly written or tagged in Korean. Unfortunately, most users in this area cannot directly specify queries in Korean. Therefore, it is necessary to translate queries from Chinese to Korean for Chinese-Korean (C-K) information retrieval systems. The main challenge involves translating named entities (such as names of shows, movies and albums) because they are usually the main concepts of queries.

Named entity (NE) translation is a challenging task because, although there are many online bilingual dictionaries, they usually lack domain specific words or NEs. Furthermore, new NEs are emerged everyday, but bilingual dictionaries cannot update their contents frequently. Therefore, it is necessary to construct a named entity translation (NET) system. In [2], the authors romanized Chinese NEs and selected their English transliterations from English NEs extracted from the Web by comparing their phonetic similarities with Chinese NEs. Yaser Al-Onaizan [3] transliterated an NE in Arabic into several candidates in English and ranked the candidates by comparing their counts in several English corpora. Chinese-Korean NET is much more difficult than NET considered in previous works because a Chinese NE may not have similar pronunciation to its Korean translation.

In this paper, we propose an effective pattern-based NET method which can achieve very high accuracy. All patterns are automatically generated and weighed, saving considerable human effort.

2 Difficulties in Chinese-Korean Named Entity Translation

To translate an NE originated from Chinese into Korean, we begin by considering the two C-K NET approaches. The older is based on the Sino-Korean pronunciation and the newer on the Mandarin. For example,"臺灣" (Taiwan) used to be transliterated solely as "대만" (Daeman). However, during the 1990s, transliteration based on Mandarin pronunciation became more popular. Presently, the most common transliteration for "臺灣" is "타이완" (Ta-i-wan), though the Sino- Korean-based "대만" is still widely used. For Chinese personal names, both ways are used. For example, the name of Chinese actor Jackie Chan ("成龍" Cheng-long) is variously transliterated as "성룡"Seong-ryong (Sino-Korean) and "청룽" Cheong-rung (Mandarin). Translating Chinese NEs by either method is a major challenge because each Chinese character may correspond to several different Hangul characters or character sequences that have similar pronunciations. This results in thousands of possible combinations of Hangul characters (e.g., "張韶涵' Zhang Shao-han can be transliterated to "장사오한" Jang-sa-o-han or "장샤오한" Jang-sya-o-han), making it very difficult to choose the most widely used one.

NEs originated from Japan may contain Hiraganas, Katakanas, or Kanjis. For each character type, J-C translation rules may be similar to or very different from J-K translation rules. Some of these rules are based on Japanese pronunciation, while some are not. For NEs composed of all Kanjis, their Chinese translations are generally exactly the same as their Kanji written forms. In contrast, Japanese NEs are transliterated into Hangul characters. Take "小泉純一郎" (Koitsumi Junichiro) for example. Its Chinese translation "小泉純一郎" is exactly the same as its Kanji written form, while its pronunciation (Xiao-quan Chun-yi-lang) is very different from its Japanese pronunciation. This is different from its Korean translation, "고이즈미 준이치로" (Ko-i-jeu-mi Jun-i-chi-ro). In this example, we can see that, because the translation rules in Chinese and Korean are different, it is ineffective to utilize phonetic similarity to find the Korean translation equivalent to the Chinese translation.

3 Pattern-Based Named Entity Translation

In this section, we describe our C-K NET method for dealing with the problems described in Section 2. We observed that an NE and its translation may co-exist in a sentence. Such sentences may have structural similarity. For example, for "李明博" and its Korean translation "이명박", the following sentences both contain the structure "NE (translation)":

"2007년 12월19일 밤 李明博 (이명박) 후보의 당선이 사실상"

NE translation

“ 李明博 (이명박) 대통령 중국 방문 의미와 과제”
NE translation

These local structures can be treated as surface patterns. In previous work, [4] employ five hand-crafted patterns to extract NE translations. However, it is time-consuming to manually create most patterns. Therefore, we aim to develop an approach that learns patterns automatically.

3.1 Learning Translation Patterns

To generate translation patterns (TP), we need to prepare sentences containing at least one Chinese NE and its Korean translation. For each pair of sentences x and y, we apply the Smith-Waterman local alignment algorithm [5] to find the longest common string. During the alignment process, positions where x and y share the same word are counted as a match. x's ith character and y's jth character are denoted as x_i and y_i, respectively. The algorithm firstly constructs an $|x| \times |y|$ matrix $\mathbf{S}$. Each element $S_{i,j}$ represents the similarity score of an optimal local alignment ending at x_i and y_j, which can be calculated by the following formula:

$$S_{i,j} = \max \begin{cases} 0 \\ S_{i-1,j-1} + s(x_i, y_i) \\ S_{i-1,j} - d \\ S_{i,j-1} - d \end{cases},$$

where $s(x_i, y_i)$ is the similarity function of x_i and y_i; d is the gap penalty. After $\mathbf{S}$ are calculated, we backtrack from the optimal element to generate the output TP τ. τ is initialed to be empty. The backtracking process iterates as follows. Suppose the current element is S_{ij}, our algorithm selects the largest one from $\{S_{i-1,j-1}, S_{i,j-1}, S_{i-1,j}\}$ as the next element. If $S_{i-1,j-1}$ is selected, that is, x_{i-1} and y_{i-1} are identical, x_{i-1} will be attached in front of τ. If either $S_{i,j-1}$ or $S_{i-1,j}$ are selected, a wild card will be attached in front of τ. This process stops until it arrives at the first zero-valued element and τ is output.

The following is an example of a pair of sentences that contains “言承旭” (Jerry Yen) and its Korean translation, “언승욱” (Eon Seung-uk) :

- 대만 배우 언승욱 (言承旭) 요약정보.
- 배우 언승욱(言承旭)이 취재진의 질문에 답하고 있다.

After alignment, the pattern is generated as:

배우 <Korean NE slot>(<Chinese NE slot>)

This pattern generation process is repeated for each NE-translation pair.

3.2 Weighting Translation Patterns

After learning the patterns, we have to filter out some ineffective patterns and determine each TP's weight for ranking translation candidates. Each TP τ is evaluated by employing it to extract all possible Korean translations for each training-set NE e in from the sentences used to generate τ. In extracting e's translations, τ's <Chinese NE slot> is replaced with e. τ's extraction F-score over all training-set NEs is treated as its weight and calculated as follows:

$$\text{Precision} = \frac{\text{\# of correctly extracted translations}}{\text{\# numbers of extracted translations}}$$

$$\text{Recall} = \frac{\text{\# of correctly extracted translations}}{\text{\# of correct translations}}$$

$$\text{F-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

3.3 Extracting and Ranking Translation Candidates

To find a Chinese NE's Korean translation, we can apply the TPs to extract possible Korean translations. For the given input Chinese NE e, our system sends e to AltaVista and limits the result pages to Chinese and Korean. The snippets are collected and break them into sentences. Only the sentences containing e are retained.

We then use all C-K TPs, whose <Chinese NE slot> are replaced with e, to match the retained sentences. The strings matched by <Korean NE slot> are extracted as e's translation candidates. Each candidate c is scored by summing up the weights of TPs extracting c. This score is used to rank all candidates.

4 Evaluation and analysis

In this section, we conduct an experiment to evaluate our pattern-based C-K NET system on different NE types.

4.1 Data Sets

In this section, we illustrate how to prepare the experimental data for learning and testing TPs. As mentioned in Section 3, TP learning requires sentences containing a Chinese NE and its Korean translation.To prepare them, we firstly collect a list of NEs in Chinese, which comprises of 120 NEs originated in Chinese such as 言承旭 (Jerry Yan) and 周杰倫 (Jay Chou) as well as NEs originated in Korea such as 裴勇俊 (Bae Young-Jun) and 張娜拉 (Jang Na-Ra). These NEs are divided into five types based on its Wikipedia page's categorization, including person, location, organization, architecture, and others. Secondly, to acquire these NEs' Korean translation, each NE is sent to the Chinese Wikipedia, and the title of the matched article's Korean version is treated as the NE's translation in Korean. Thirdly, each NE and its Korean translation are attached in a query and then the query is sent to AltaVista. For instance, "言承旭" (Jerry Yan) and its Korean translation "언승욱" are used to produce a query "+言承旭 +언승욱". The returned snippets in the top 20 pages are split into sentences. Only the sentences that contain at least one NE and its Korean translation are retained in the test set.

The preparation of test data is similar to that of training data. 40 NEs (other than the 120 training NEs) are collected from the Wikipedia. The distribution among the five categories are exactly the same as that of the training NEs. Then, each NE is directly sent to Altavista. The returned snippets in the top 20 pages are split into sentences. Only the sentences that contain the NE are retained.

4.2 Evaluation Methodology

We use the Mean Average Precision (MAP) [6] and Top-1 Precision at the Top-1 to measure our NET system's performance. For evaluating the performance of translating each NE, we can calculate the average precision (AP), which is the average of precisions computed after

Table 1: Evaluation Results

NE Type	MAP	Top-1 Precision
Location	0.7854	0.8571
Person	0.9458	1.0000
Organization	0.8333	0.6667
Architecture	0.7736	0.8333
Others	0.8702	0.7500
All	0.8402	0.8500

truncating the list after each of the correct translations in turn:

$$AP = \frac{\sum_{r=1}^{N}(P(r) \times rel(r))}{\text{number of correct translations}},$$

where r is the rank, N the number of extracted candidates, $rel()$ a binary function on the correctness of a given rank, and $P()$ precision at a given cut-off rank. For evaluating the performance of translating all NEs, we can calculate the mean average precision (MAP), which is mean value of all the average precisions. Top-1 precision, used for evaluating the quality of our system's Top-1 candidates, is computed as the ratio of numbers that the correct translation is the top 1 candidate divided by the total number of NE queries.

For evaluating the translation result, our Korean expert manually checked if the candidates are correct translations. A candidate is judged as correct regardless it is generated based on the Sino-Korean or Mandarin pronunciation.

4.3 Evaluation Result

Table 1 shows the categorical and overall performance of our pattern-based NET method. The results show that our method can achieve close to 0.85 in both metrics. The scores of translating person names are even higher. Both metrics are over 0.9 and the Top-1 precision is equal to 1, that is, all the Top-1 candidates output by our system are correct. This indicates that our system's ranking for translation candidates is very accurate and can be further exploited in other applications that incorporate our NET system.

5 Discussion

From the evaluation results, we find that our method can translate most Chinese NEs effectively. However, there are still some cases dropping the performance. In the following subsections, we discuss TP's effectiveness and analyze error cases.

5.1 Effectiveness of Pattern-based NET

For most test NEs, our method can extract their correct translations and put them in the forepart of the candidate list. For example, for "金泰熙" (Jin Tai-xi), the Top-1 candidate is "김태희" (Kim Tae-Hee), which is exactly the correct translation. As mentioned in section 2, Korean transliterates Chinese NEs according to their Sino-Korean or Mandarin pronunciation. Our method can extract translations based on both methods. For example, for "謝長廷" (Xie Chang-ting), both the Sino-Korean transliteration "사장정" (Sa-jang-jeong) as well as the Mandarin transliterations such "셰 창팅" (Sye-chang-ting), "세 창팅" (Se-chang-ting), and "시에츠앙

팅” (Si-e-cheu-ang-ting) are extracted. It shows that our method can extract most Chinese NEs’ Korean translations regardless of how the Korean translations are generated.

5.2 Error Analysis

5.2.1 Relevant Terms

Our method may extract an NE’s relevant phrases in addition to its translations. For instance, for “親民黨” (People First Party), both the correct translation “친민당” (Chin-min-dang) and the relevant phrase “쑹추위” (Ssung Chu-ui, 宋楚瑜, the chairman of People First Party) are extracted. Although relevant phrases are not exact the translations, they might improve the performance of some NET applications such as cross-language information retrieval (CLIR). This is because the effect of adding relevant phrases in queries is similar to that of query expansion.

5.2.2 Different Phraseology in Chinese and Korean

Different phraseology in Chinese and Korean might make our NET method extract false positives. For example, the First Sino-Japanese War (1894–1895) is called Jiawu Zhanzheng (甲午戰爭) by Chinese but is called Cheong-il-jeong-jaeng (淸日戰爭) by Koreans. The Top-1 candidate output by our system is “갑오전쟁” (kap-o-jeon-jaeng), which is only annotated for Koreans to understand its pronunciation. The most widely used Korean translations for the First Sino-Japanese War is “청일전쟁” (Cheong-il-jeong-jaeng, 淸日戰爭). The other example is the Chinese query “浪漫滿屋” (Lang-man-man-u), a Korean drama’s name. The Top-1 candidate is “랑만만우” (rang-man-man-u), which is the transliteration of the Chinese characters “浪漫滿屋” based on the Mandarin pronunciation. However, the correct Korean translation is “풀하우스” (Full House), which is the transliteration based on the English pronunciation. These two queries show that different phraseology may rank the incorrect translation candidates higher. However, the correct translations, such as “청일정쟁” and “풀하우스”, are also extracted by our method. For some applications, such as CLIR, the inaccurate ranking does not influence the performance a lot [7].

6 Conclusion

In this paper, we have demonstrated several advantages of our pattern-based NE translation method. Our pattern-based method achieves higher recall because it extracts NE translations from the Web, which contains most of human knowledge. Even translations of novel NEs can be found. Second, our method can extract most translations for each NE. This feature makes similar effects of query expansion and very helpful for cross-language information retrieval because documents containing frequent or infrequent translations can be retrieved. Finally, the high MAP over all five domains establishes our method’s generality.

In the future, we plan to apply our method to other language pairs. We also hope to extract not only the translations but also relevant information to them. We believe these new features can be applied to other applications, such retrieving multimedia contents on the Web 2.0 platform whose tags are written in different languages from queries.

References

[1] Cho Hae-Joang, "Reading the "Korean wave as a sign of global shift"", *Korea Journal*, pp. 167–172, 2005.

[2] Hsin-Hsi Chen, Sheng-Jie Huang, Yung-Wei Ding, and Shih-Cbung Tsai, "Proper name translation in cross-language information retrieval", *Proceedings of 17th COLING and 36th ACL*, pp. 232–236, 1998.

[3] Yaser Al-Onaizan and Kevin Knight, "Translating named entities using monolingual and bilingual resources", *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL)*, pp. 400–408, 2002.

[4] Dong Zhou, Mark Truran, Tim Brailsford, and Helen Ashman, "NTCIR-6 experiments using pattern matched translation extraction", *Proceedings of NTCIR-6 Workshop Meeting*, 2006.

[5] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences", *Journal of Molecular Biology*, vol. 147, pp. 195–197, 1981.

[6] Tefko Saracevic, Paul Kantor, Alice Y. Chamis, and Donna Trivison, "A study of information seeking and retrieving", *Journal of the American Society for Information Science*, vol. 39, no. 3, pp. 161–176, 1988.

[7] Yu-Chun Wang, Richard Tzong-Han Tsai, and Wen-Lian Hsu, "Learning patterns from the web to translate named entities for cross language information retrieval", *Proceedings of the third International Joint Conference on Natural Language Processing (IJCNLP)*, 2008.