Chinese-Japanese Search Query Translation System

Chenhui Chu
Graduate School of Informatics
Kyoto University
Yoshida-honmachi, Sakyo-ku
Kyoto, 606-8501, Japan
chu@nlp.ist.i.kyoto-u.ac.jp

Keiji Shinzato
Rakuten Institute of Technology
4-13-9, Higashi-shinagawa
Shinagawa-ku, Tokyo
140-0002, JAPAN
keiji.shinzato@mail.rakuten.com

Abstract

We describe a Chinese-Japanese search query translation system using search results of commercial web search API. Because a lack of available search results between Chinese and Japanese, our system uses intermediate language and do translation in a gradual manner, the intermediate language is not limited to English. Our system is domain independent and could be easily expanded to other language pairs. Experimental results show that our system is quite suitable for Chinese-Japanese search query translation task and can acquire high translation accuracy.

1 Introduction

Statistical machine translation (Brown et al., 1993) has become the main research field in machine translation, among which phrase-based statistical translation model (Koehn et al., 2003) and syntax-based statistical translation model (Yamada and Knight, 2001) have been developed significantly these years. However, these machine translation models are all for sentence level translation. Because in search query translation, the translation object is a combination of several words (usually less than three) which lacks of contextual information, these machine translation models are thought to be not suitable for search query translation task.

Search query translation task can be conducted by simple dictionary matching. This translation approach is effective for search queries consisting of common (or general) words such as ”自行车(bicycle)”, ”头盔(helmet)”. Dictionary-based translation, however, may not work well for product search queries which are frequently consisted of proper nouns and product names. For example, we can suppose the Chinese query ”拉菲庄园(シャトー　ラフィット, in Japanese)” which is a proper noun in wine domain and unknown word in an ordinary dictionary. Moreover, although both Chinese and Japanese come from French ”Chateau Lafite”, the Japanese is a transliteration of the original French, while in the Chinese, transliteration and translation are mixed together , where ”拉菲(Lafite)” is transliteration and ”庄园(Chateau)” is translation. In this case, using a transliteration model could not accomplish the translation task either.

In this paper, we focus on translation task of search queries used in product search. We propose a method of using search results of commercial web search API and develop a Chinese-Japanese search query translation system. We also report experimental results that show the effectiveness of our system on search query translation task. Because our system could be easily expanded to other language pairs, we think that it would be quite useful to connect Rakuten group’s services among different countries and improve the distribution.

2 System Description

Figure 1 is an overview of our system. The basic idea is using commercial web search API to find out Alphabet translation for Chinese search query, then search for Japanese translation using the Alphabet translation. It is unnecessary to construct specified parallel corpus, which makes our system domain independent, also it would be very easy to expand our

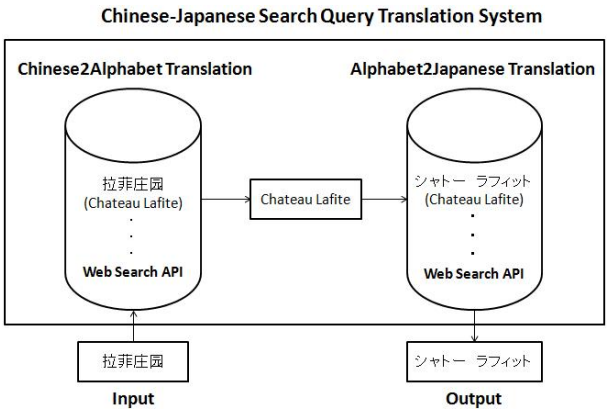

Figure 1: System overview.

translation system to other language pairs.

It would be straightforward if we can find the Japanese translation for Chinese search query directly using commercial web search API. However, because a lack of available Chinese-Japanese search results, we use intermediate language and do translation in two steps:

- Step 1: Chinese2Alphabet Translation
- Step 2: Alphabet2Japanese Translation

The intermediate language is fixed to be written in Alphabet, however, it is not limited to English.

2.1 Chinese2Alphabet Translation

In this step, we get Chinese search results from commercial web search API for the Chinese search query, and extract Alphabet translation candidates from the contents in ”Title” and ”Snippet” of the search results and obtain the best candidate by defined score calculation.

Take Chinese search query ”拉菲庄园” for example, by web search API we can get Chinese search results in ”Title” and ”Snippet” like this:

Title:

拉菲庄园【Chateau Lafite Rothschild】- 葡萄酒酒庄- 法国葡萄酒网...

Snippet:

2011年5月28日2011年7月9日... 拉菲庄园（Chateau Lafite Rothschild），是1855年波尔多葡萄酒评级时的顶级葡萄庄园之一...

Table 1: Examples of Alphabet n-gram candidate scores.

Candidates	n	freq	score
Chateau Lafite	2	40	43.94
Chateau Lafite Rothschild	3	31	42.97
Lafite Rothschild	2	33	36.25
Lafite	1	46	31.88
Chateau	1	43	29.81

We extract the nearest Alphabet sequence next to the search query. In this example, we can get ”Chateau Lafite Rothschild” twice. Because any n-gram of the extracted sequences may be the translation, which is called n-gram translation candidate, we use the following formula to calculate scores for all n-gram translation candidates extracted from search results:

$$score = \log(n+1) \times freq, \qquad (1)$$

where n is the n-gram length, $freq$ is the frequency of n-gram translation candidate appearing in search results. The one holding the highest score is regarded as the best candidate. Table 1 shows examples of Alphabet n-gram candidate scores of ”拉菲庄园” calculated using 200 search results from Yahoo search API, where ”Chateau Lafite” becomes the Alphabet translation for the next step.

2.2 Alphabet2Japanese Translation

This step is similar to Step 1, we get Japanese search results from commercial web search API for the Alphabet translation from Step 1, and obtain the best Japanese translation candidate by parsing the contents in ”Title” and ”Snippet” of the search results using particular rules.

Take Alphabet translation ”Chateau Lafite” for example, by web search API we can get Japanese search results in ”Title” and ”Snippet” like this:

Title:

シャトー ラフィット ロートシルト Chateau Lafite Rothschild ... - coneco.net

Snippet:

シャトー ラフィット ロートシルト Chateau Lafite Rothschild 1964 750mlを探して比べてお得に買う！価格比較サイトconeco.net...

Table 2: Examples of Katakana bi-gram candidate frequencies.

Candidates	freq
シャトー　ラフィット	70
ラフィット　ロートシルト	68
シャトー　マルゴー	8
シャトー　ラトゥール	7
シャトー　ル	5

Because translation for proper noun and product name search queries from Alphabet to Japanese can always be considered as a transliteration problem, we extract the nearest Katakana sequence next to the Alphabet translation. In this example, we get "シャトー　ラフィット　ロートシルト" twice. Also, in Alphabet to Japanese transliteration, the number of words is always the same, so we only consider Katakana n-gram holding the same number of words to the Alphabet translation as translation candidate. In this example, only bi-gram is regarded as translation candidate. By adding this constraint, the formula used to calculate n-gram candidate score in Step 1 could be simplified to only related to frequency. Table 2 shows examples of Katakana bi-gram candidate frequencies for "Chateau Lafite" in 200 search results from Yahoo search API, where "シャトー　ラフィット" becomes the final output of our system.

However, there may not exist Katakana n-gram with the same number of words to the Alphabet translation in search results. In such case, we use the following formula to calculate scores for all n-gram translation candidates extracted from search results:

$$score = \frac{1}{sim\_rank} + \frac{1}{n\_gram\_rank} + \frac{1}{|n - len| + 1}, \quad (2)$$

where sim_rank is the rank of similarity between Katakana n-gram and Alphabet translation, for similarity calculation we use AlphaBeta transliteration model (Brill and Moore, 2000), n_gram_rank is the rank of n-gram scores calculated using (1), n is n-gram length, and len is number of words in Alphabet translation.

Of course it is reasonable of only using (2) for candidates selection. However, the method we described above showed better performance in the preliminary experiments.

Table 3: Examples of test data.

Chinese	Japanese
华诗歌	ヴァスコス
都夏美伦	デュアール　ミロン
木桐**酒庄**	シャトー　ムートン　ロートシルト
拉高斯**城堡**	シャトー　グラン　ピュイラコスト
龙船将军	アミラル　ド　ベイシュヴェル

Table 4: Translation evaluation results.

	Full	Part	Null	No
Chinese2Alphabet	43	37	17	5
Alphabet2Japanese	68	5	17	12
Chinese2Japanese	35	32	14	21

3 Experiments

3.1 Settings

The test data we used is 102 manually extracted wine name pairs from Rakuten China and Rakuten Japan. Table 3 shows some examples of our test data, where the Chinese characters in bold are translation, the others are transliteration of original Alphabet. We also extracted the corresponding Alphabet translation. The web search API we used is Yahoo search API. Considering the tradeoff between translation accuracy and speed, we set the number of search results for each search query to 200. We did experiments for both the two separate steps and the system.

3.2 Evaluation

For machine translation evaluation, BLEU (Papineni et al., 2002) is the most popular metric being used. However, as mentioned in Section 1, search query translation task has its specificity that it is not sentence level translation, so we do not use BLEU. Instead, we use our own method for evaluation and define three types of translation matching:

- Full Match: if the translation output is exactly the same as golden translation
- Part Match: if the translation output is substring of golden translation, and vice versa

Table 5: Examples of Full Match.

System input	System output
拉菲	ラフィット
木桐嘉棣	ムートン　カデ
奥比昂**庄园**	シャトー　オー　ブリオン
布琅**兄弟**梅洛	ブラウン　ブラザーズ　メルロー
宝马挚友	アルテル　エゴ　ド　パルメ

Table 6: Examples of Null Match.

System output	Golden translation
ミヤルスト　**ルピコン**	ミヤルスト　**ルピコン**
バンフィ　**キアンティ**	**キャンティ**
リンデマン　カワラ	**リンデマンズ**

- Null Match: besides Full Match and Part Match

Besides these matching types, we also define ”No Result” for the ones that have no translation outputs.

Table 4 shows the evaluation results of our translation system, where Chinese2Alphabet and Alphabet2Japanese are translation accuracies of the two separate steps, and Chinese2Japanese is the final translation accuracy of our system. Table 5 shows some examples of ”Full Match”, we can see that even the Chinese search query that are mixed with translation and transliteration could be successfully translated by our system. Table 6 shows some examples of ”Null Match”, although the system outputs are evaluated as ”Null Match” based on our evaluation method, the Katakana in bold could be regarded as right translations according to human evaluation.

However, there exist many ”No Result”, which is a drawback of our system, we are thinking about using dictionary and transliteration models to translate these ”No Result” search queries as an auxiliary of our system. Also, we notice that the number of ”Full Match” in Chinese2Alphabet translation is less than Alphabet2Japanese, we think that doing segmentation for Chinese search query and adding the same number of words constraint could increase the number of ”Full Match” for Chinese2Alphabet translation. Moreover, because our system depends on search results of commercial web search API, and search results may change day by day, the system output may be not stable.

4 Conclusion

We developed a Chinese-Japanese search query translation system using search results of commercial web search API, which is quite suitable for Chinese-Japanese search query translation task and can acquire high translation accuracy. The system is domain independent and could be easily expanded to other language pairs.

However, there still exist some problems in our system, which we have to solve. Currently, the system only works for Chinese to Japanese search query translation task. In the future, we are planning to expand it to Japanese to Chinese search query translation task.

5 Acknowledgements

This work was done during my internship at Rakuten Institute of Technology (RIT). Thanks to all members of RIT, especially Dr. Masato Hagiwara who gave a lot of valuable advices for this work.

References

Eric Brill and Robert C. Moore. 2000. An improved error model for noisy channel spelling correction. In *Proceedings of 38th Annual Meeting of the Association for Computational Linguistics*, pages 286–293, Hong Kong, China, January. Association for Computational Linguistics.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Association for Computational Linguistics*, 19(2):263–312.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *HLT-NAACL 2003: Main Procedings*, pages 127–133.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Kenji Yamada and Kevin Knight. 2001. A syntax-based statistical translation model. In *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, pages 523–530, Toulouse, France, July. Association for Computational Linguistics.