English-Korean Named Entity Transliteration Using Substring Alignment and Re-ranking Methods

Chun-Kai Wu[†] **Yu-Chun Wang**[‡] **Richard Tzong-Han Tsai**[†]
[†]Department of Computer Science and Engineering,
Yuan Ze University, Taiwan
[‡]Department of Computer Science and Information Engineering,
National Taiwan University, Taiwan
s983301@mail.yzu.edu.tw d97023@csie.ntu.edu.tw
thtsai@saturn.yzu.edu.tw

Abstract

In this paper, we describe our approach to English-to-Korean transliteration task in NEWS 2012. Our system mainly consists of two components: an letter-to-phoneme alignment with m2m-aligner,and transliteration training model DirecTL-p. We construct different parameter settings to train several transliteration models. Then, we use two re-ranking methods to select the best transliteration among the prediction results from the different models. One re-ranking method is based on the co-occurrence of the transliteration pair in the web corpora. The other one is the JLIS-Reranking method which is based on the features from the alignment results. Our standard and non-standard runs achieves 0.398 and 0.458 in top-1 accuracy in the generation task.

1 Introduction

Named entity translation is a key problem in many NLP research fields such as machine translation, cross-language information retrieval, and question answering. Most name entity translation is based on transliteration, which is a method to map phonemes or graphemes from source language into target language. Therefore, named entity transliteration system is important for translation.

In the shared task, we focus on English-Korean transliteration. We consider to transform the transliteration task into a sequential labeling problem. We adopt m2m-aligner and DirecTL-p (Jiampojamarn et al., 2010) to do substring mapping and transliteration predicting, respectively. With this approach (Jiampojamarn et al., 2010) achieved promising results on NEWS 2010 transliteration tasks. In order to improve the transliteration performance, we also apply several ranking techniques to select the best Korean transliteration.

This paper is organized as following. In section 2 we describe the main approach we use including how we deal with the data, the alignment and training methods and our re-ranking techniques. In section 3, we show and discuss our results on English-Korean transliteration task. And finally the conclusion is in section 4.

2 Our Approach

In this section, we describe our approach for English-Korean transliteration which comprises the following steps:

1. Pre-processing
2. Letter-to-phoneme alignment
3. DirecTL-p training
4. Re-ranking results

2.1 Pre-processing

Korean writing system, namely *Hangul*, is alphabetical. However, unlike western writing system with Latin alphabets, Korean alphabet is composed into syllabic blocks. Each Korean syllabic block represent a syllable which has three components: initial consonant, medial vowel and optionally final consonant. Korean has 14 initial consonants, 10 medial vowels, and 7 final consonants. For instance, the syllabic block "신" (sin) is composed with three letters:

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 57–60,
Jeju, Republic of Korea, 8-14 July 2012.

a initial consonant "ㅅ" (s), a medial vowel "ㅣ" (i), and a final consonant "ㄴ" (n).

For transliteration from English to Korean , we have to break each Korean syllabic blocks into two or three Korean letters. Then, we convert these Korean letters into Roman letters according to Revised Romanization of Korean for convenient processing.

2.2 Letter-to-phoneme Alignment

After obtaining English and Romanized Korean name entity pair, we generate the alignment between each pair by using m2m-aligner.

Since English orthography might not reflect its actual phonological forms, it makes one-to-one character alignment between English and Korean not practical.

Compared with traditional one-to-one alignment, the m2m-aligner overcomes two problems: One is double letters where two letters are mapped to one phoneme. English may use several characters for one phoneme which is presented in one letter in Korean, such as "ch" to "ㅊ" and "oo" to "ㅜ". However, one-to-one alignment only allows one letter to be mapped to one phoneme, so it must have to add an null phoneme to achieve one-to-one alignment. It may interfere with the transliteration prediction model.

The other problem is double phonemes problem where one letter is mapped to two phonemes. For example, the letter "x" in the English name entity "Texas" corresponds to two letters "ㄱ" and "ㅅ" in Korean. Besides, some English letters in the word might not be pronounced, like "k" in the English word "knight". We can eliminate this by preprocessing the data to find out double phonemes and merge them into single phoneme. Or we can add an null letter to it, but this may also disturb the prediction model. While performing alignments, m2m aligner allows us to set up the maximum length substring in source language (with the parameter x) and in target language (with the parameter y). Thus, when aligning, we set both parameter x and y to two because we think there are at most 2 English letters mapped to 2 Korean letters. To capture more double phonemes, we also have another parameter set with $x = 1$ and $y = 2$.

As mentioned in previous section, Korean syllabic block is composed of three or two letters. In order to cover more possible alignments, we construct another alignment configurations to take null consonant into consideration. Consequently, for any Korean syllabic block containing two Korean letters will be converted into three Roman letters with the third one being a predefined Roman letter representing null consonant. We also have two set of parameters for this change, that is $x = 2$, $y = 3$ and $x = 1$,$y = 3$. The reason we increase both y by one is that there are three Korean letters for each word.

2.3 DirecTL-p Training

With aligned English-Korean pairs, we can train our transliteration model. We apply DirecTL-p (Jiampojamarn et al., 2008) for our training and testing task. We train the transliteration models with different alignment parameter settings individually mentioned in section 2.2.

2.4 Re-ranking Results

Because we train several transliteration models with different alignment parameters, we have to combine the results from different models. Therefore, the re-ranking method is necessary to select the best transliteration result. For re-ranking, we propose two approaches.

1. Web-based re-ranking
2. JLIS-Reranking

2.4.1 Web-based re-ranking

The first re-ranking method is based on the occurrence of transliterations in the web corpora. We send each English-Korean transliteration pair generated by our transliteration models to Google web search engine to get the co-occurrence count of the pair in the retrieval results. But the result number may vary a lot, most of them will get millions of results while some will only get a few hundred.

2.4.2 JLIS-Reranking

In addition to web-based re-ranking approach, we also adopt JLIS-Reranking (Chang et al., 2010) to re-rank our results for the standard run. For an English-Korean transliteration pair, we can measure if they are actual transliteration of each other by observing the alignment between them. Since

Table 1: Results on development data.

Run	Accuracy	Mean F-score	MRR	MAP_{ref}
1 ($x = 2$, $y = 2$)	0.488	0.727	0.488	0.488
2 ($x = 1$, $y = 2$)	0.494	0.730	0.494	0.494
3 ($x = 1$, $y = 3$, with *null* consonant)	0.452	0.713	0.452	0.452
4 ($x = 2$, $y = 3$, with *null* consonant)	0.474	0.720	0.474	0.473
Web-based Reranking	0.536	0.754	0.563	0.536
JLIS-Reranking	0.500	0.737	0.500	0.500

Table 2: Results on test data

Run	Accuracy	Mean F-score	MRR	MAP_{ref}
Standard (JLIS-Reranking)	0.398	0.731	0.398	0.397
Non-standard (Web-based reranking)	0.458	0.757	0.484	0.458

DirecTL-p model outputs a file containing the alignment of each result, there are some features in the results that we can use for re-ranking. In our re-ranking approach, there are three features used in the process: *source grapheme chain* feature, *target grapheme chain* feature and *syllable consistent* feature. These three feature are proposed in (Song et al., 2010).

Source grapheme chain feature: This feature can tell us that how the source characters are aligned. Take "A|D|A|M" for example, we will get three chains which are A|D, D|A and A|M. With this feature we may know the alignment in the source language.

Target grapheme chain feature: Similar to the above feature, it tell us how the target characters are aligned. Take "NG:A:n|D|A|M" for example, which is the Korean transliteration of ADAM, we will get three chains which are n|D, D|A and A|M. With this feature we may know the alignment in the target language. "n" is the predefined null consonant.

Syllable consistent feature: We use this feature to measure syllable counts in both English and Korean. For English, we apply an Perl module[1] to measure the syllable counts. And for Korean, we simply count the number of syllabic blocks. This feature may guard our results, since a wrong prediction may not have the same number of syllable.

[1] `http://search.cpan.org/~gregfast/Lingua-EN-Syllable-0.251/Syllable.pm`

Other than the feature vectors created by above features, there is one important field when training the re-ranker, performance measure. For this field, we give it 1 when we predict a correct result otherwise we give it 0 since we think it is useless to get a partially correct result.

3 Result

To measure the transliteration models with different alignment parameters and the re-ranking methods, we construct several runs for experiments as follows.

- Run 1: m2m-aligner with parameters $x = 2$ and $y = 2$.
- Run 2: m2m-aligner with parameters $x = 1$ and $y = 2$.
- Run 3: m2m-aligner with parameters $x = 1$ and $y = 3$ and add null consonants in the Korean romanized representation.
- Run 4: m2m-aligner with parameters $x = 2$ and $y = 3$ and add null consonants in the Korean romanized representation.
- Web-based reranking: re-rank the results from run 1 to 4 based on Google search results.
- JLIS-Reranking: re-rank the results from run 1 to 4 based on JLIS-rerakning features.

Table 1 shows our results on the development data. As we can see in this table, Run 2 is better than Run 1 by 6 NEs. It may be that the data in develop

set are double phonemes. And we also observe that both Run 1 and Run 2 is better than Run 3 and Run 4, the reason may be that the extra null consonant distract the performance of the prediction model.

From the results, it shows that our re-ranking methods can actually improve transliteration. Reranking based on web corpora can achieve better accuracy compared with web-based reranking. The JLIS-Reranking method slightly improve the accuracy. It could be that the features we use are not enough to capture the alignment between English-Korean NE pair.

Because the runs with re-ranking achieving better results, we submit the result on the test data with JLIS-Reranking as the standard run, and the result with the web-based re-ranking as the non-standard run for our final results. The results on the test data set are shown in table 2. The results also shows that the web-based re-ranking can achieve the best accuracy up to 0.458.

4 Conclusion

In this paper, we describe our approach to English-Korean named entity transliteration task for NEWS 2012. First, we decompose Korean word into Korean letters and then romanize them into sequential Roman letters. Since a Korean word may not contain the final consonant, we also create some alignment results with the null consonant in romanized Korean representations. After preprocessing the training data, we use m2m-aligner to get the alignments from English to Korean. Next, we train several transliteration models based on DirecTL-p with the alignments from the m2m-aligner. Finally, we propose two re-ranking methods. One is web-based re-ranking with Google search engine. We send the English NE and its Korean transliteration pair our model generates to Google to get the co-occurrence count to re-rank the results. The other method is JLIS-reranking based on three features from the alignment results, including source grapheme chain feature, target grapheme chain feature, and syllable consistent feature. In the experiment results, our method achieves the good accuracy up to 0.398 in the standard run and 0.458 in non-standard run. Our results show that the transliteration model with a web-based re-ranking method can achieve better accuracy in English-Korean transliteration.

References

Ming-Wei Chang, Vivek Srikumar, Dan Goldwas-ser, and Dan Roth. 2010. Structured output learning with indirect supervision. *Proceeding of the International Conference on Machine Learning (ICML)*.

Sittichai Jiampojamarn, Grzegorz Kondrak, and Tarek Sherif. 2007. Applying many-to-many alignments and hidden markov models to letter-to-phoneme conversion. *Association for Computational Linguistics*, pages 372–379.

Sittichai Jiampojamarn, Colin Cherry, and Grzegorz Kondrak. 2008. Joint processing and discriminative training for letter-to-phoneme conversion. *Association for Computational Linguistics*, pages 905–912.

Sittichai Jiampojamarn, Kenneth Dwyer, Shane Bergsma, Aditya Bhargava, Qing Dou, Mi-Young Kim, and Grzegorz Kondrak. 2010. Transliteration generation and mining with limited training resources. *Proceedings of the 2010 Named Entities Workshop, ACL 2010*, pages 39–47.

Yan Song, Chunyu Kit, and Hai Zhao. 2010. Reranking with multiple features for better transliteration. *Proceedings of the 2010 Named Entities Work-shop, ACL 2010*, pages 62–65.