NCU IISR English-Korean and English-Chinese Named Entity Transliteration Using Different Grapheme Segmentation Approaches

Yu-Chun Wang[†] **Chun-Kai Wu**[‡] **Richard Tzong-Han Tsai**[§*]

[†]Department of Computer Science and Information Engineering, National Taiwan University, Taiwan
[‡]Department of Computer Science, National Tsinghua University, Taiwan
[§]Department of Computer Science and Information Engineering, National Central University, Taiwan

d97023@csie.ntu.edu.tw s102065512@m102.nthu.edu.tw
thtsai@csie.ncu.edu.tw

Abstract

This paper describes our approach to English-Korean and English-Chinese transliteration task of NEWS 2015. We use different grapheme segmentation approaches on source and target languages to train several transliteration models based on the M2M-aligner and DirecTL+, a string transduction model. Then, we use two reranking techniques based on string similarity and web co-occurrence to select the best transliteration among the prediction results from the different models. Our English-Korean standard and non-standard runs achieve 0.4482 and 0.5067 in top-1 accuracy respectively, and our English-Chinese standard runs achieves 0.2925 in top-1 accuracy.

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. The vast majority of named entities (NE) such as person or organization names do not appear in bilingual dictionaries, and new NEs are being generated every day, making it difficult to keep an up-to-date list of NEs. One solution for NE translation is to use online encyclopedias like Wikipedia that contain pages in both the source and target language. However, coverage is spotty for many languages and/or NE categories.

Since the translations of many NEs are based on transliteration, a method of mapping phonemes or graphemes from a source language to a target language, researchers have developed automated transliteration techniques to add to the NE translation toolbox. NE transliteration has featured as a shared task in previous Named Entities Workshops (NEWS).

In the shared task for NEWS 2015, we focus on English-Korean and English-Chinese transliteration. We adopt the M2M-aligner and DirecTL+ to map substrings and predict transliteration results. Jiampojamarn et al. (2010) achieved promising results using this approach in the NEWS 2010 transliteration task. The Korean writing system, Hangul, is alphabetic, but Chinese characters are logograms. Because English and Korean use alphabetic writing systems, we apply different grapheme segmentation methods to create several transliteration models. For Chinese, we treat each distinct Chinese character as a basic unit for the alignment step. In order to improve the transliteration performance, we also apply two ranking techniques to select the best transliterations.

This paper is organized as follows. In Section 2 we describe our main approach, including how we preprocess the data, our alignment and training methods, and our reranking techniques. In Section 3 we show our results on the English-Korean and English-Chinese transliteration tasks and discuss our findings. Finally the conclusion is in Section 4.

2 Our Approach

Our approach for English-Korean and English-Chinese transliteration comprises the following steps:

1. Preprocessing
2. Alignment
3. DirecTL+ training
4. Re-ranking results

*corresponding author

Proceedings of the Fifth Named Entity Workshop, joint with 53rd ACL and the 7th IJCNLP, pages 83–87,
Beijing, China, July 26-31, 2015.

2.1 Preprocessing

2.1.1 English

Since English uses the Latin alphabet, we use three different segmentation methods for alignment: single letter, fine segmentation algorithm, and phonemic representation.

Single Letter (SINGLE) NEs are separated into single letters for further alignment. For example, the English name "ALEXANDER" is separated as four letters "A L E X A N D E R" for the alignment in the next step.

Fine-grained Segment Algorithm (FSA) Unlike English letters and words, each Hangul block or Chinese character corresponds to a syllable. Some previous approaches have used English letters and Chinese characters/Korean syllabic blocks as the basic alignment units for transliteration (Oh and Choi, 2006; Li et al., 2004; Jia et al., 2009). Other approaches have tried to segment English NEs into syllabic chunks for alignment with Hangul blocks or Chinese characters (Wan and Verspoor, 1998; Jiang et al., 2007; Zhang et al., 2012).

We adopt a heuristic syllable segmentation algorithm, namely Fine-grained Segment Algorithm (FSA), proposed by Zhang et al. (2012) with slight modification to syllabify English NEs. Our modified version of the FSA is defined as follows:

1. Replace '*x*' in English names with '*k s*'.

2. {'*a*', '*o*', '*e*', '*i*', '*u*'} are defined as vowels. '*y*' is defined as a vowel when it is not followed by a vowel.

3. When '*w*' follows '*a*', '*e*', '*o*' and isn't followed by '*h*', treat '*w*' and the preceding vowel as a new vowel symbol; Step 2 and 3 form the basic vowel set.

4. A consecutive vowels sequence which is formed by the basic vowel set is treated as a new vowel symbol, excepting '*iu*', '*eo*', '*io*', '*oi*', '*ia*', '*ui*', '*ua*', '*uo*'; Step 2, 3 and 4 form the new vowel set.

5. Consecutive consonants are separated; a vowel symbol(in the new vowel set) followed by a consonant sequence is separated from the sequence; if a vowel followed by a consonat sequence and the first consonat is { '*h*', '*l*', '*m*', '*n*', '*r*' }, the first consonat symbol is concatenated with the vowel into a syllable.

6. A consonant and its following vowel are treated as a syllable; the rest of the isolated consonants and vowels are regarded as individual syllables in each word.

   For example, the English term "ALEXANDER" is segmented as "A LE K SAN DER" by the FSA.

Phonemic Representation (PHONEME) In addition, since Korean is a phonological writing system, for non-standard runs, we also adopt phonemic information for English name entities. The English word pronunciations are obtained from the CMU Pronouncing Dictionary v0.7a. The CMU pronouncing dictionary provides the phonemic representations of English pronunciations with a sequence of phoneme symbols. For instance, the previous example *ALEXANDER* is segmented and tagged as the phonemic representation < AE L AH G Z AE N D ER >. Since the CMU pronouncing dictionary does not cover all the pronunciation information of the name entities in the training data, we also apply LOGIOS Lexicon Tool to generate the phonemic representations of all other name entities not in the CMU pronouncing dictionary.

2.1.2 Korean

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

We take two segmentation method for Korean: Hangul blocks and romanized letters.

Hangul Blocks (HANGUL) Hangul syllabic blocks of Korean words are separated into single blocks for further alignment. For example, the

http://www.speech.cs.cmu.edu./cgi-bin/cmudict

http://www.speech.cs.cmu.edu/tools/lextool.html

Korean word "녹스" is separated as two syllabic blocks "녹 스" for the alignment in the next step.

Romanized Letters (ROMAN) This segmentation method break each Hangul syllabic blocks into Korean letters and then convert these Korean letters into Roman letters according to Revised Romanization of Korean for convenient processing. For example, the Korean word "녹스" is first taken apart as "ㄴㅗㄱㅅㅡ", and then romanized as "n o k s eu".

2.1.3 Chinese

For Chinese, we treat each Chinese character as a basic alignment unit. Chinese chacters of a Chinese word are segment as each single Chinese character for further alignment processing. For example, the Chinese word "诺克斯" is separated as three character "诺 克 斯".

2.2 Alignment

After generating English, Korean, and Chinese segmented substrings in the previous step, we determine the alignment between each English-Korean and English-Chinese pair using the M2M-aligner (Jiampojamarn et al., 2007). The M2M-aligner is a many-to-many alignment method based on the expectation maximization (EM) algorithm. It allows us to create alignments between substrings of various lengths. During alignment, empty strings (*nulls*) are only allowed on the target side.

2.3 DirecTL+ Training

With aligned English-Korean and English-Chinese pairs, we can train our transliteration model. We apply DirecTL+ (Jiampojamarn et al., 2008) for training and testing. DirecTL+ is an online discriminative training model for string transduction problems. We individually train the transliteration models with different segmentation methods individually mentioned in section 2.1.

2.4 Reranking Results

Because we train several transliteration models with different alignment settings, we can combine the results from different models to select the best transliterations. Therefore, reranking is a necessary step to generate the final results. For reranking, we propose two approaches.

1. Orthography Similarity Ranking
2. Web-based Ranking

2.4.1 Orthography Similarity Ranking

For standard runs which are allowed to use the training data only, we measure the orthographic similarity between the term in the source language and the transliteration candidate. The transliteration candidates in target languages are all first Romanized into Latin alphabet sequences. Then, we rank the similarity between the source language term and the Romanized transliteration candidate according to the string edit distance.

2.4.2 Web-based Ranking

The second reranking method is based on the occurrence of transliterations in the web corpora. We send each transliteration pair generated by our transliteration models to the Bing web search engine to get the co-occurrence count of the pair in the retrieval results. We use mutual information between the source language term and the transliteration candidate as the similarity score for ranking.

3 Results

To measure the transliteration models with different segmentation methods and the reranking methods, we construct the following experimental runs:

English-Korean (EnKo) Runs:

- Run 1: SINGLE + HANGUL
- Run 2: SINGLE + ROMAN
- Run 3: PHONEME + ROMAN
- Run 4: FSA + HANGUL
- Run 5: FSA + ROMAN
- Run 6: Orthography Similarity Ranking with Run 1 to 5
- Run 7: Web-based Ranking with Run 1 to 5

English-Chinese (EnCh) Runs:

- Run 1: FSA + Chinese characters
- Run 2: SINGLE + Chinese characters

Table 1 and table 2 show the final results of our transliteration approaches on the English-Korean (EnKo) and the English-Chinese (EnCh) test data.

The EnKo results show that the alignment between single English letter and Romanized Korean letter (Run 2) achieves the best results among run 1

Table 1: Final results on the English-Korean (EnKo) test data

Run	NEWS 11				NEWS12			
	ACC	F-score	MRR	MAP_{ref}	ACC	F-score	MRR	MAP_{ref}
1	0.3186	0.6576	0.3186	0.3112	0.3276	0.7078	0.3276	0.3269
2	0.4483	0.7255	0.4483	0.4392	0.4457	0.7482	0.4457	0.4448
3	0.2742	0.6000	0.2742	0.2689	0.1457	0.5222	0.1457	0.1455
4	0.2151	0.5707	0.2151	0.2098	0.1743	0.5835	0.1743	0.1740
5	0.0427	0.3329	0.0427	0.0415	0.0562	0.3752	0.0562	0.0562
6	0.2085	0.5270	0.3432	0.2048	0.1952	0.5522	0.3349	0.1950
7	0.4992	0.7330	0.5395	0.4943	0.5067	0.7614	0.5317	0.5055

Table 2: Final results on the English-Chinese (EnCh) test data

Run	NEWS 11				NEWS12			
	ACC	F-score	MRR	MAP_{ref}	ACC	F-score	MRR	MAP_{ref}
1	0.2325	0.6303	0.2325	0.2199	0.2351	0.6237	0.2351	0.2242
2	0.2925	0.6719	0.2925	0.2772	0.2798	0.6455	0.2798	0.2652

to 5. The run with the alignment between English phonemic representation and Romanized Korean letter (Run 3) is not as good as Run 2. It might be due to two reasons: one is that the Korean transliteration is often based on the orthography, not the actual pronunciation; the second reason is that the pronunciation from LOGIOS lexicon tool may not be accurate to get the correct phonemic forms.

The FSA segmentation method (Run 4 and 5) does not perform well as other runs, especially, the Run 5 (FSA + ROMAN) has the worst result. The reason might be the unbalanced segment units between English and Korean. The M2M-aligner is originally designed to do letter-to-phoneme alignment. The FSA method grouping the consecutive English letter into syllables, but the Romanized Korean letters are all single characters. It might cause the M2M-aligner generate the incorrect alignment in this run. In EnCh runs, the FSA segmentation method (Run 1) also performs slightly worse than the single English letter segmentation method (Run 2).

The web-based ranking method (EnKo Run 7) significantly improves the transliteration performance. Because web corpora contains the actual usages of the transliterations, it is a good resource to rank and select the best transliterations. The orthography similarity ranking method (Run 6) does not improve but actually degrades the transliteration performance. This may be because the English orthography does not always reflect actual pronunciations; therefore, the similarity between English and Korean orthographies is insufficient to measure the quality of transliteration candidates.

4 Conclusion

In this paper, we describe our approach to English-Korean and English-Chinese NE transliteration task for NEWS 2015. We adopt different grapheme segmentation methods for the source and target languages. For English, three segmentation methods are used: single letter, fine-grained syllable algorithm, and phonemic representation. For Korean, we segment according to Hangul syllabic blocks and Romanized Hangul letters. For Chinese, we treat each Chinese character as a basic alignment unit. After segmenting the training data, we use the M2M-aligner to get the alignments from the source and target languages. Next, we train different transliteration models based on DirecTL+ with the alignments from the M2M-aligner. Finally, we use two reranking methods: web-based ranking using the Bing search engine, and the orthography similarity method based on the string edit distance of the orthographic forms in source and target languages. In experiments, our method achieves accuracy up to 0.4483 in the standard run and 0.5067 in the non-standard run for English-Korean. For English-Chinese standard run, it achieves an accuracy of 0.2925.

References

Yuxiang Jia, Danqing Zhu, and Shiwen Yu. 2009. A noisy channel model for grapheme-based machine transliteration. In *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration*, pages 88–91. Association for Computational Linguistics.

Sittichai Jiampojamarn, Grzegorz Kondrak, and Tarek Sherif. 2007. Applying many-to-many alignments and hidden markov models to letter-to-phoneme conversion. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 372–379, Rochester, New York, April. Association for Computational Linguistics.

Sittichai Jiampojamarn, Colin Cherry, and Grzegorz Kondrak. 2008. Joint processing and discriminative training for letter-to-phoneme conversion. In *Proceedings of ACL-08: HLT*, pages 905–913, Columbus, Ohio, June. Association for Computational Linguistics.

Sittichai Jiampojamarn, Colin Cherry, and Grzegorz Kondrak. 2010. Integrating joint n-gram features into a discriminative training framework. In *Proceedings of NAACL-2010*, Los Angeles, CA, June. Association for Computational Linguistics.

Long Jiang, Ming Zhou, Lee-Feng Chien, and Cheng Niu. 2007. Named entity translation with web mining and transliteration. In *IJCAI*, volume 7, pages 1629–1634.

Haizhou Li, Min Zhang, and Jian Su. 2004. A joint source-channel model for machine transliteration. In *Proceedings of the 42nd Annual Meeting on association for Computational Linguistics*, page 159. Association for Computational Linguistics.

Jong-Hoon Oh and Key-Sun Choi. 2006. An ensemble of transliteration models for information retrieval. *Information processing & management*, 42(4):980–1002.

Stephen Wan and Cornelia Maria Verspoor. 1998. Automatic english-chinese name transliteration for development of multilingual resources. In *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, pages 1352–1356. Association for Computational Linguistics.

Chunyue Zhang, Tingting Li, and Tiejun Zhao. 2012. Syllable-based machine transliteration with extrac phrase features. In *Proceedings of NEWS 2012*, pages 52–56.