A Hybrid Approach to English-Korean Name Transliteration

Gumwon Hong[*], **Min-Jeong Kim**[*], **Do-Gil Lee**[+] **and Hae-Chang Rim**[*]
[*]Department of Computer Science & Engineering, Korea University, Seoul 136-713, Korea
{gwhong,mjkim,rim}@nlp.korea.ac.kr
[+]Institute of Korean Culture, Korea University, Seoul 136-701, Korea
motdg@korea.ac.kr

Abstract

This paper presents a hybrid approach to English-Korean name transliteration. The base system is built on MOSES with enabled factored translation features. We expand the base system by combining with various transliteration methods including a Web-based n-best re-ranking, a dictionary-based method, and a rule-based method. Our standard run and best non-standard run achieve 45.1 and 78.5, respectively, in top-1 accuracy. Experimental results show that expanding training data size significantly contributes to the performance. Also we discover that the Web-based re-ranking method can be successfully applied to the English-Korean transliteration.

1 Introduction

Often, named entities such as person names or place names from foreign origin do not appear in the dictionary, and such out of vocabulary words are a common source of errors in processing natural languages. For example, in statistical machine translation (SMT), if a new word occurs in the input source sentence, the decoder will at best drop the unknown word or directly copy the source word to the target sentence. Transliteration, a method of mapping phonemes or graphemes of source language into those of target language, can be used in this case in order to identify a possible translation of the word.

The approaches to automatic transliteration between English and Korean can be performed through the following ways: First, in learning how to write the names of foreign origin, we can refer to a transliteration standard which is established by the government or some official linguistic organizations. No matter where the standard comes from, the basic principle of the standard is based on the correct pronunciation of foreign words. Second, since constructing such rules are very costly in terms of time and money, we can rely on a statistical method such as SMT. We believe that the rule-based method can guarantee to increase accuracy for known cases, and the statistical method can be robust to handle various exceptions.

In this paper, we present a variety of techniques for English-Korean name transliteration. First, we use a phrase-base SMT model with some factored translation features for the transliteration task. Second, we expand the base system by applying Web-based n-best re-ranking of the results. Third, we apply a pronouncing dictionary-based method to the base system which utilizes the pronunciation symbols which is motivated by linguistic knowledge. Finally, we introduce a phonics-based method which is originally designed for teaching speakers of English to read and write that language.

2 Proposed Approach

In order to build our base system, we use MOSES (Koehn et al., 2007), a well-known phrase-based system designed for SMT. MOSES offers a convenient framework which can be directly applied to machine transliteration experiments. In this framework, the transliteration can be performed in a very similar process of SMT task except the following changes. First, the unit of translation is changed from *words* to *characters*. Second, a *phrase* in transliteration refers to any contiguous block of character sequence which can be directly matched from a source word to a target word. Also, we do not have to worry about any distortion parameters because decoding can be performed in a totally monotonic way.

The process of the general transliteration approach begins by matching the unit of a source

Proceedings of the 2009 Named Entities Workshop, ACL-IJCNLP 2009, pages 108–111,
Suntec, Singapore, 7 August 2009.

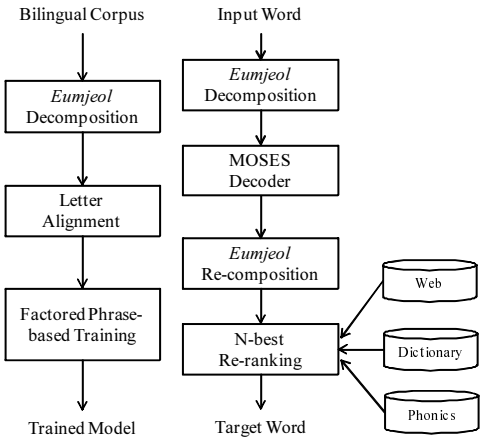

Figure 1: System Architecture

word to the unit of a target word. The unit can be based on graphemes or phonemes, depending on language pairs or approaches. In English-Korean transliteration, both grapheme-to-grapheme and grapheme-to-phoneme approaches are possible. In our method, we select grapheme-to-grapheme approach as a base system, and we apply grapheme-to-phoneme functions in pronouncing dictionary-based approach.

The transliteration between Korean and other languages requires some special preprocessing techniques. First of all, Korean alphabet is organized into syllabic blocks called *Eumjeol*. Korean transliteration standard allows each *Eumjeol* to consist of either two or three of the 24 Korean letters, with (1) leading 14 consonants, (2) intermediate 10 vowels, and (3) optionally, trailing 7 consonants (out of the possible 14). Therefore, Korean *Eumjeol* should be decomposed into letters before performing training or decoding any input. Consequently, after the letter-unit transliteration is finished, all the letters should be re-composed to form a correct sequence of *Eumjeols*.

Figure 1 shows the overall architecture of our system. The alignment between English letter and Korean letter is performed using GIZA++ (Och and Ney, 2003). We use MOSES decoder in order to search the best sequence of transliteration.

In this paper we focus on describing factored phrase-based training and n-best re-ranking techniques including a Web-based method, a pronouncing dictionary-based method, and a phonics-based method.

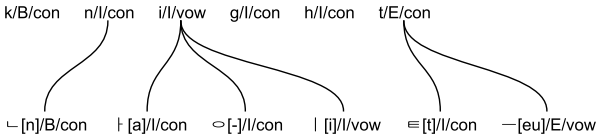

Figure 2: Alignment example between 'Knight' and '나이트 [*naiteu*]'

2.1 Factored Phrase-based Training

Koehn and Hoang (2007) introduces an integration of different information for phrase-based SMT model. We report on experiments with three factors: surface form, positional information, and the type of a letter. Surface form indicates a letter itself. For positional information, we add a BIO label to each input character in both the source words and the target words. The intuition is that certain character is differently pronounced depending on its position in a word. For example, 'k' in 'Knight' or 'h' in 'Sarah' are not pronounced. The type of a letter is used to classify whether a given letter is a vowel or a consonant. We assume that a consonant in source word would more likely be linked to a consonant in a target word. Figure 2 shows an example of alignment with factored features.

2.2 Web-based Re-ranking

We re-ranked the top *n* results of the decoder by referring to how many times both source word and target word co-occur on the Web. In news articles on the Web, a translation of a foreign name is often provided near the foreign name to describe its pronunciation or description. To reflect this observation, we use Google's proximity search by restricting two terms should occur within four-word distance. The frequency is adjusted as relative frequency form by dividing each frequency by total frequency of all n-best results.

Also, we linearly interpolate the n-best score with the relative frequency of candidate output. To make fair interpolation, we adjust both scores to be between 0 and 1. Also, in this method, we decide to remove all the candidates whose frequencies are zero.

2.3 Pronouncing Dictionary-based Method

According to "*Oeraeeo pyogibeop*[1]" (Korean orthography and writing method of borrowed for-

[1]http://www.korean.go.kr/08_new/data/rule03.jsp

Methods	Acc.$_1$	Mean F$_1$	Mean F$_{dec}$	MRR	MAP$_{ref}$	MAP$_{10}$	MAP$_{sys}$
BS	0.451	0.720	0.852	0.576	0.451	0.181	0.181
ER	0.740	0.868	0.930	0.806	0.740	0.243	0.243
WR	0.784	0.889	0.944	0.840	0.784	0.252	0.484
PD	0.781	0.885	0.941	0.839	0.781	0.252	0.460
PB	0.785	0.887	0.943	0.840	0.785	0.252	0.441

Table 1: Experimental Results (EnKo)

eign words), the primary principle of English-to-Korean transliteration is to spell according to the mapping table between the international phonetic alphabets and the Korean alphabets. Therefore, we can say that a pronouncing dictionary-based method is very suitable for this principle.

We use the following two resources for building a pronouncing dictionary: one is an English-Korean dictionary that contains 130,000 words. The other is the CMU pronouncing dictionary[2] created by Carnegie Mellon University that contains over 125,000 words and their transcriptions.

Phonetic symbols for English words in the dictionaries are transformed to their pronunciation information by using an internal code table. The internal code table represents mappings from each phonetic symbol to a single character within ASCII code table. Our pronouncing dictionary includes a list of words and their pronunciation information.

For a given English word, if the word exists in the pronouncing dictionary, then its pronunciations are translated to Korean graphemes by a mapping table and transformation rules, which are defined by "*Oeraeeo pyogibeop*".

2.4 Phonics-based Method

Phonics is a pronunciation-based linguistic teaching method, especially for children (Strickland, 1998). Originally, it was designed to connect the sounds of spoken English with group of English letters. In this research, we modify the phonics in order to connect English sounds to Korean letter because in Korean there is nearly a one-to-one correspondence between sounds and the letter patterns that represent them. For example, alphabet 'b' can be pronounced to 'ㅂ'(*bieup*) in Korean. Consequently, we construct about 150 rules which map English alphabet into one or more several Korean graphemes, by referring to the phonics. Though phonics cannot reveal all of the pronunciation of English words, the conversion from English alphabet into Korean letter is performed simply and efficiently. We apply the phonics in serial order from left to right of each input word. If multiple rules are applicable, the most specific rules are fist applied.

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

3 Experiments

3.1 Experimental Setup

We participate in both standard and non-standard tracks for English-Korean name transliteration in NEWS 2009 Machine Transliteration Shared Task (Li et al., 2009). Experimenting on the development data, we determine the best performing parameters for MOSES as follows.

- ***Maximum Phrase Length***: 3
- ***Language Model N-gram Order***: 3
- ***Language Model Smoothing***: Kneser-Ney
- ***Phrase Alignment Heuristic***: grow-diag-final
- ***Reordering***: Monotone
- ***Maximum Distortion Length***: 0

With above parameter setup, the results are produced from the following five different systems.

- ***Baseline System (BS)***: For the standard task, we use only given official training data [3] to construct translation model and language model for our base system.
- ***Expanded Resource (ER)***: For all four non-standard tasks, we use the examples of writing foreign names as additional training data. The examples are provided from the National Institute of the Korean Language[4]. The data originally consists of around 27,000 person names and around 7,000 place names including non-Ascii characters for English side words as well as duplicate entries. We preprocess the data in order to use 13,194 dis-

[3]Refer to Website http://www.cjk.org for more information

[4]The resource is open to public. See http://www.korean.go.kr/eng for more information.

tinct pairs of English names and Korean transliteration.

- ***Web-based Re-ranking (WR)***: We re-rank the result of *ER* by applying the method described in section 2.2.
- ***Pronouncing Dictionary-based Method (PD)***: The re-ranking of *WR* by combining with the method described in section 2.3.
- ***Phonics-based Method (PB)***: The re-ranking of *WR* by combining with the method described in section 2.4.

The last two methods re-rank the *WR* method by applying pronouncing dictionary-based method and Phonics-based method. We restrict that the pronouncing dictionary-based method and Phonics-based method can produce only one output, and use the outputs of the two methods to re-rank (again) the result of Web-based re-ranking. When re-ranking the results, we heuristically combined the outputs of *PD* or *PB* with the n-best result of *WR*. If the outputs of the two methods exist in the result of *WR*, we add some positive scores to the original scores of *WR*. Otherwise, we inserted the result into fixed position of the rank. The fixed position of rank is empirically decided using development set. We inserted the output of *PD* and *PB* at second rank and at sixth rank, respectively.

3.2 Experimental Results

Table 1 shows our experimental results of the five systems on the test data. We found that the use of additional training data (*ER*) and web-based re-ranking (*WR*) have a strong effect on transliteration performance. However, the integration of the *PD* or *PB* with *WB* proves not to significantly contribute the performance. To find more elaborate integration of those results will be one of our future work.

The MAP_{sys} value of the three re-ranking methods *WR*, *PD*, and *PB* are relatively higher than other methods because we filter out some candidates in n-best by their Web frequencies. In addition to the standard evaluation measures, we include the Mean F_{dec} to measure the Levenshtein distance between reference and the output of the decoder (decomposed result).

4 Conclusions

In this paper, we proposed a hybrid approach to English-Korean name transliteration. The system is built on MOSES with factored translation features. When evaluating the proposed methods, we found that the use of additional training data can significantly outperforms the baseline system. Also, the experimental result of using three n-best re-ranking techniques shows that the Web-based re-ranking is proved to be a useful method. However, our two integration methods with dictionary-based or rule-based method does not show the significant gain over the Web-based re-ranking.

For future work, we plan to devise more elaborate way to integrate statistical method and dictionary or rule-based method to further improve the transliteration performance. Also, we will apply the proposed techniques to possible applications such as SMT or Cross Lingual Information Retrieval.

References

Philipp Koehn and Hieu Hoang. 2007. Factored translation models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 868–876, Prague, Czech Republic, June. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *ACL 2007, Demo and Poster Sessions*, June.

Haizhou Li, A Kumaran, Min Zhang, and Vladimir Pervouchine. 2009. Whitepaper of news 2009 machine transliteration shared task. In *Proceedings of ACL-IJCNLP 2009 Named Entities Workshop (NEWS 2009)*, Singapore.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29.

D.S. Strickland. 1998. *Teaching phonics today: A primer for educators*. International Reading Association.