English-Korean Automatic Transliteration/Back-transliteration System and Character Alignment

Byung-Ju Kang
Department of Computer Science
Advanced Information Technology Research Center (AITrc)
Korea Terminology Center for Language and Knowledge Engineering
Korea Advanced Institute of Science and Technology
373-1 Kusong-dong, Yusong-gu,
Daejeon,
305-701, Korea
bjkang@world.kaist.ac.kr

Key-Sun Choi
Department of Computer Science
Advanced Information Technology Research Center (AITrc)
Korea Terminology Center for Language and Knowledge Engineering
Korea Advanced Institute of Science and Technology
373-1 Kusong-dong, Yusong-gu,
Daejeon,
305-701, Korea
kschoi@world.kaist.ac.kr

Abstract

Recently there is increasing concern about automatic transliteration and back-transliteration across two different languages, especially with radical differences in their alphabets and phoneme inventories such as English/Korean (Kang & Choi, 2000; Lee & Choi, 1998; Jeong et al., 1999), English/Japanese (Knight & Graehl, 1997), English/Arabic (Stalls & Knight, 1998), English/Chinese (Wan & Verspoor, 1998), etc. Transliteration is, given a source language word, to find its phonetic equivalent in target language. Back-transliteration is the backward process that finds the origin word from the transliterated word. For example, English word "internet" is generally transliterated into "intonet (romanization)" in Korean and the right back-transliteration of "intonet" should be "internet". Automatic transliteration and back-transliteration have several important applications in NLP systems such as machine translation, cross-lingual information retrieval, etc.

In this demo, we present English-to-Korean automatic transliteration system and Korean-to-English automatic back-transliteration system that are based on decision tree learning (Kang & Choi, 2000; 2000a). Our methodology is fully bi-directional, i.e. the same methodology is used for both transliteration and back-transliteration. For the machine learning of transliteration and back-transliteration rules, many phonetically aligned English word and Korean transliteration pairs are needed. However, such aligned pairs are not generally available so that automatic alignment is necessary. We developed a very effective English/Korean character alignment algorithm (Kang & Choi, 2000b). English/Korean character alignment is, given a source language word (English) and its phonetic equivalent in target language (Korean), to find the most phonetically probable correspondence between their characters. Generally English/Korean character alignment has many-to-many correspondence and also has null correspondence. These characteristics make the application of machine learning methods difficult or ineffective. So we constrained the alignment configuration such that in transliteration one and only one English character may be mapped to one or more Korean characters or null, on the other hand, in back-transliteration one and only one Korean character may be mapped to one or more English characters or null. This way of alignment reduces the number of decision trees to be learned to only 26 for each English alphabet in the case of English-to-Korean transliteration and only 46 for each Korean alphabet in the case of Korean-to-English back-transliteration. Once the 26 or 46 decision trees are independently learned, the transliteration of an English word or the back-transliteration of a Korean word is very

straightforward. Given an input English word, in the case of transliteration, each alphabet is mapped to Korean strings using its decision tree and then concatenating all the Korean strings produces final Korean transliteration. Back-transliteration is done in the similar way.

Acknowledgements

This work was supported by the Korea Science and Engineering Foundation (KOSEF) through the Advanced Information Technology Research Center (AITrc).

References

Jeong, K. S., S. H. Myaeng, J. S. Lee, and K. S. Choi (1999) Automatic identification and back-transliteration of foreign words for information retrieval. *Information Processing and Management*, 35(4):523-540.

Kang, B. J. and K. S. Choi (2000) Automatic transliteration and back-transliteration by decision tree learning. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, Athnes, Greece.

Kang, B. J. and K. S. Choi (2000a). Automatic transliteration and back-transliteration. TR 00-31-004, Advanced Information Technology Research Center, Korea Advanced Institute of Science and Technology.

Kang, B. J. and K. S. Choi (2000b) Character alignment. TR 00-31-005, Advanced Information Technology Research Center, Korea Advanced Institute of Science and Technology.

Knight, K. and J. Graehl, (1997) Machine Transliteration. In *Proceedings. of the 35th Annual Meetings of the Association for Computational Linguistics (ACL)*, Madrid, Spain.

Lee, J. S. and K. S. Choi (1998) English to Korean Statistical transliteration for information retrieval. *Computer Processing of Oriental Languages*, 12(1):17-37.

Stalls, B. and K. Knight (1998) Translating Names and Technical Terms in Arabic Text. In *Proceedings of the COLING/ACL Workshop on Computational Approaches to Semitic Languages*.

Wan, S. and C. M. Verspoor (1998) Automatic English-Chinese name transliteration for development of multilingual resources. In *Proceedings of COLING-ACL'98, the joint meeting of 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics*, Montreal, Canada.