Direct Combination of Spelling and Pronunciation Information for Robust Back-Transliteration

Slaven Bilac and Hozumi Tanaka

Tokyo Institute of Technology
Ookayama 2-12-1, Meguro, 152-8552 Tokyo, Japan
{sbilac,tanaka}@cl.cs.titech.ac.jp,

Abstract. Transliterating words and names from one language to another is a frequent and highly productive phenomenon. For example, English word *cache* is transliterated in Japanese as キャッシュ "kyasshu". Transliteration is information losing since important distinctions are not always preserved in the process. Hence, automatically converting transliterated words back into their original form is a real challenge. Nonetheless, due to its wide applicability in MT and CLIR, it is an interesting problem from a practical point of view.
In this paper, we demonstrate that back-transliteration accuracy can be improved by directly combining grapheme-based (i.e. spelling) and phoneme-based (i.e. pronunciation) information. Rather than producing back-transliterations based on grapheme and phoneme model independently and then interpolating the results, we propose a method of first combining the sets of allowed rewrites (i.e. edits) and then calculating the back-transliterations using the combined set. Evaluation on both Japanese and Chinese transliterations shows that direct combination increases robustness and positively affects back-transliteration accuracy.

1 Introduction

With the advent of technology and increased flow of goods and services, it has become quite common to integrate new words from one language to another. Whenever a word is adopted into a new language, pronunciation is adjusted to suit the phonetic inventory of the language. Furthermore, the orthographic form of the word is modified to allow representation in the target language script. For example, English word *cache* is transliterated in Japanese as キャッシュ "kyasshu".[1] In similar fashion, a proper noun Duncan is transliterated as 桔刃 "deng4ken3" in Chinese.[2] This process of acquisition and assimilation of a new word into an existing writing system is referred to as transliteration [1].

[1] We use *italics* to transcribe the English words, while Japanese transliterations (e.g. キャッシュ) are given with romaji (i.e. roman alphabet) in "typewriter" font (e.g. "kyasshu"). The romanization used follows [1], thus closely reflecting English–like pronunciation with long vowels transcribed as "aa" rather than "ā".

[2] Chinese transliterations are given with the PinYin romanization with numerical tone codes.