SOME LINGUISTIC ISSUES IN THE MACHINE TRANSLITERATION OF CHINESE, JAPANESE, AND ARABIC NAMES

Keynote address at 6th NEWS Named Entities Workshop

Jack Halpern
CEO, The CJK Dictionary Institute, Niiza, Japan

ABSTRACT

The romanization of non-Latin scripts is a complex computational task that is highly language dependent. This presentation will focus on three of the most challenging non-Latin scripts: Chinese, Japanese, and Arabic (CJA).

Much progress has been made in personal name machine-transliteration methodologies, as documented in the various NEWS reports over the last several years. Such techniques as phrase-based SMT, RNN-based LM and CRF have emerged, leading to gradual improvements in accuracy scores. But methodology is only one aspect of the problem. Equally important is the high level of ambiguity of the CJA scripts, which poses special challenges to named entity extraction and machine transliteration. These difficulties are exacerbated by the lack of comprehensive proper noun dictionaries, the multiplicity of ambiguous transcription schemes, and orthographic variation.

This presentation will clear up the differences between three basic concepts -- transliteration, transcription, and romanization -- that are a source of much confusion, even among computational linguists, and will focus on (1) the major linguistics issues, that is, the special characteristics of the CJA scripts that impact machine transliteration, and (2) the important role played by lexical resources such as personal name dictionaries.

A major issue in romanizing Simplified Chinese (SC) is the one-to-many ambiguity of many characters (*polyphones*), such as /lè/ and /yuè/ for 乐. To disambiguate accurately, the names must be looked up in word-level (not character-level) name mapping tables. This is complicated by (1) the presence of orthographic variants in traditional Chinese (TC), and (2) the need to for cross-script conversion between (SC) and (TC), Transcription into Chinese is even more ambiguous, since some phonemes can correspond to dozens of characters.

A major characteristic of Japanese, a highly agglutinative language, is the presence of countless orthographic variants. The four Japanese scripts interact in a complex way, resulting in *okurigana* variants (取り扱い, 取扱い, 取扱 etc. for /toriatsukai/), cross-script variants (猫, ねこ, ネコ for /neko/), kanji variants (大幅 and 大巾 for /oohaba/), kana variants (ユーザー and ユーザ for /yuuza(a)/), and more. Another issue is the numerous *kun* and *nanori* readings (some kanji have dozens) and the various romanization systems in current use, such as the Hepburn, Kunrei and hybrid systems.

The Arabic script poses a different set of challenges to developers of NLP tools in general, and to machine transliteration of names in particular. This includes but is not limited to a high level of morphological and orthographical ambiguity, many ambiguous transcription schemes, and name variant expansion. For example, the string كاتب can represent distinct words such as /kaatib/, /kaataba/ and /kaatiba/, while long /aa/ can be written as ا ,ى and آ. Automatically romanizing unvocalized Arabic without resorting to mapping tables is a complex task fraught with pitfalls.

These linguistic and orthographic difficulties are exacerbated by the lack of good lexical resources, especially of comprehensive personal name mapping tables. We will introduce several large-scale CJA name databases designed to support accurate romanization, transcription, and cross-script conversion (a subset of which has been used in the NEWS shared tasks over the last few years) and explain how these resources can be used to enhance the accuracy of name transliteration systems.