Exploring Syllables, Romanization, and Analogy in Names

Deryle Lonsdale
BYU Linguistics and English Language

This paper discusses the application of a language modeling approach to perform Romanization of Arabic-script names. Romanization involves taking a word (in this case a name) and rendering it into the Roman alphabet, the script used for the English language (among many others).

Prior work in Romanizing names has involved a wide range of approaches involving variable combinations of linguistic, statistical, corpus-based, lexical, and computational methods (Lonsdale, 2005). Romanization of personal names is inherently a human issue that must also take into consideration language structure (of at least the source and target languages), cultural context, and orthographic conventions.

This talk reports on recent work to predict and score possible English Romanizations of names (given names and surnames) from Arabic script. The investigation was carried out using analogical modeling (Skousen et al., 2002), a machine learning approach that has been widely applied to model language use at the phonological, morphological, and lexical levels. The system is exemplar-based and finds symbolic structural analogies between prior feature vectors and similarly vectorized queries.

The raw sponsor-supplied data included several thousand names and their Romanizations. We mention interesting variants in both the source language orthography and in the corresponding English Romanizations. We then discuss our development of an automatic procedure for labeling the original data using a finite-state transducer. Once the labeled instances were collected and manipulated to create a training set, the system was run and results collected. Sample scored inputs and outputs include the following:

"حافظ"
450.000 Hafiz
450.000 Hafeez
"بهنام "
436.044 Bahnaam
402.424 Behnaam

"جمشيد"
399.414 Jamsheed
396.716 Jamshid
394.940 Jamshaid
384.322 Jamasheed

For example, we see two equiprobable Romanizations for the name "حافظ ", whereas the name "بهنام " has two variants with different scores. In the last name, variants even disagree in the number of syllables (the last having an extra syllable not observed in its previous three alternatives). Such variation is common since Arabic-script languages typically do not represent vowels overtly in the orthography.

One of the interesting issues that (perhaps predictably) arose involved syllable structure. Disagreement (or at least hesitation) about syllable boundaries does influence orthography, and the crosslinguistic discrepancies inherent this project resulted in appreciable variation in names. The analogical nature of the system's processing was able to reflect such differences in producing Romanized forms that were viable, though not even present in the sponsor data. In addition, though not directly representing statistical information, the system is able to derive scores based on recurrent patterns observed in the training data.

We conclude with a discussion of possible applications and further work.

References

Skousen, R., D. Lonsdale & D.B. Parkinson (Editors). (2002) *Analogical Modeling: An exemplar-based approach to language*, Vol. 10 of *Human Cognitive Processing Series.* Amsterdam: John Benjamins Publishing Company.

Lonsdale, D. (2005). Transcription, transliteration, transduction, and translation: a typology of crosslinguistic name representation strategies; Fifth Family History Technology Workshop; Brigham Young University, Provo, UT, March 2005.