Statistical Transliteration of Judeo-Arabic Texts

Gitit Kehat and Nachum Dershowitz

Judeo-Arabic is a group of Arabic-based languages used by Jews for many centuries, which like some other Jewish languages (for example,Yiddish), is written in the Hebrew alphabet. Many of the great Jewish literary works of the Middle Ages were written in Judeo-Arabic. A large quantity of additional Judeo-Arabic text has become available with the digitization of manuscripts found in the Cairo Genizah.

Processing Judeo-Arabic involves several difficulties. First, the Hebrew transliteration of Arabic words is often not in one-to-one correspondence with Arabic orthography, due to language restrictions, nor need it be consistent within a single text, frequently causing some loss of the original Arabic information. In addition, Judeo-Arabic texts were not bound by the classical written Arabic standardization rules, and, like spoken Arabic, could vary in different countries or during different periods. These two facts make it hard for Islamic and Arabic researchers to utilize these texts.

We present our current work on statistical transliteration of Judeo-Arabic into the Arabic alphabet. In this ongoing work, we apply several methods in order to recover the Arabic, as well as its diacritics. Our first attempts involved methods that are similar to post-OCR correction, when the uncertainty of the manually-transliterated word is treated as in the case of inaccurate or missing characters within an OCR'ed word. We created word patterns from the non-ambiguous characters, and gave certain degrees of freedom to Hebrew characters that can be transliterated into several different Arabic characters. The words were then compared to an Arabic lexicon and were given the transliteration with the minimal error rate according to the Levenshtein edit distance. But, while in OCR, the errors are mainly due to geometric similarities between different characters, in our case, the so-called errors in the manual transliteration into Arabic are mainly for phonetic reasons or due to grammatical constraints caused by the transformation between Arabic and Hebrew characters.

In addition to the purely statistical approach, we employed in our work morphological analyzers of Modern Standard Arabic, to solve the natural ambiguity that appears in non-diacritized Arabic text. This analysis was used in both directions: as part of the statistical transliteration process to generate a larger lexicon of Arabic words, and as a feedback system to test the transliteration validity in a repeating process until convergence. In this process, suggested transliterations of each word were subject to morphological analysis, which checked their existence as Arabic word inflections.

Initial results show a slight improvement in the accuracy of transliteration, especially for Judeo-Arabic text which is relatively similar to Standard Arabic. Future work includes combining local syntactic analysis so as to minimize ambiguity. We also consider generating Judeo-Arabic texts from a given text of the same period that was originally written in Arabic, and perform a learning process in the opposite direction, basing ourselves on the saved characteristics of each transliterated character.