Automatic Transliteration of Judeo-Arabic Texts into Arabic Script

Kfir Bar, Tel Aviv University
Nachum Dershowitz, Tel Aviv University
Yaacov Choueka, The Friedberg Genizah Project

The Judeo-Arabic languages comprise a set of dialects spoken and written by Jewish communities living in Arab countries, mainly during the middle ages. Judeo-Arabic is typically written in Hebrew letters, enriched with various diacritic marks. The Judeo-Arabic spoken and written by any particular Jewish community is similar to the Arabic dialect used by their local Muslim community. In addition, Judeo-Arabic dialects borrow words from Aramaic and Hebrew, sometimes modified according to Arabic morphological rules. Since the Arabic alphabet is larger than the Hebrew one, additional diacritic marks are added to some Hebrew letters when rendering Arabic consonants that are lacking in the Hebrew alphabet. Judeo-Arabic authors often use different letters and diacritic marks to represent the same Arabic consonant. For example, some authors use ג to represent ج and ג' to represent غ, while others reverse the two. This inconsistency increases the level of ambiguity of a given word, making the reading of Judeo-Arabic texts a challenging task even for an Arabic speaker. Such inconsistencies may be observed even in the same document. For instance, the letter י sometimes represents the letter ي, such as in the word פי ("in", في), sometimes represents the letter ئ, such as in the word סילת ("I was asked", سُئِلْتُ), and sometimes the letter ى, as in עלי ("on"/"to"/"at", in Arabic على).

Currently, many works in Judeo-Arabic are being made available on the web. However, most Arabic speakers are unfamiliar with the Hebrew script, let alone the way it is used to render Judeo-Arabic. Therefore, there is a crucial need for automatic tools capable of transliterating Judeo-Arabic texts in Arabic letters. Since Judeo-Arabic texts usually contain some non-Arabic words, transliteration is only one step toward providing a full Arabic translation for a given Judeo-Arabic input. We are focusing mainly on the transliteration process, leaving Hebrew and Aramaic words in their original Hebrew script.

As mentioned above, Judeo-Arabic is not a single language, but rather a set of dialects, each used by the local Jewish community in some of the Arab countries. Some texts are similar to Classical Arabic, the ancestor of the Modern Standard Arabic (MSA), which is widely used today in formal settings, while other texts are more similar to local Muslim dialects. We focus, for now, on the Judeo-Arabic version that is similar to MSA more than on the colloquial versions.

We model the transliteration process using the noisy-channel approach, by employing a phrase-based statistical translation system (Koehn et al., 2003) trained on the character level. We use three models:

- A *translation* model, along with a character-based reordering model, generated from pairs of Judeo-Arabic words, one in Hebrew script and one in transliterated Arabic form. The parallel words are extracted from the original Judeo-Arabic version of the medieval book, the *Kuzari*, composed by Judah Halevi, alongside its Arabic translation, prepared by Nabih Bashir (Bashir, 2012).
- An Arabic *word* model, trained on the character level over a large number of Arabic texts extracted from Arabic Gigaword (4th ed.) (Parker et al., 2011).
- A word-based Arabic *language* model, trained on a large portion of Arabic Gigaword.

Following the machine-translation phrase-based approach, the translation and word models are combined using a log-linear formula, and their weights are automatically tuned using a relatively small development set.

Given a Judeo-Arabic word for transliteration, the statistical translation process, as described above, does not consider the preceding words as part of the context for deciding how to transliterate it; therefore, we have experimented with an additional re-ranking component, applied to the *n* best transliterations that are generated by the translation system for every input word (*n* is a variable parameter). The re-ranking decision is made using the word-based language model, trained on a large portion of Arabic Gigaword.

In this paper, we report on our experiments and results obtained with different values for the translation parameters. All our experiments are performed using the Moses (Koehn et al., 2007) implementation of phrase-based statistical translation. We evaluate our system on unseen words from two different sources. One includes parts of the *Kuzari* that were not used for training, and another includes parts of the seventh section of the *Book of Beliefs and Opinions* (in Hebrew, אמונות ודעות; better known in Arabic as כתאב אלאמאנאת ואלאעתקאדאת), completed in 933 C.E. by Sa'adia Gaon. The Arabic version of this treatise was also created by Nabih Bashir.

It turns out that most Judeo-Arabic Hebrew characters can be transliterated deterministically into Arabic; hence, a simple baseline algorithm that is based on some deterministic rules, was able to correctly transliterate 93.4% of the test set. With our system's best settings, trained on circa 8,000 parallel words, we were able to improve on this, correctly transliterating 96.9% of the test-set letters.

The following is an example of an original text written in Judeo-Arabic (left) and its corresponding Arabic transliteration (right), as generated by our algorithm. This section is taken from the seventh chapter of Sa'adia's work. Transliteration mistakes are highlighted in red. Among others, we consider a missing *shadda* (Arabic geminate indicator) and an incorrect placement of a *hamza* (glottal stop indicator) on an *alif* as mistakes.

אלמקאלה אלסאבעה	المقالة السابعة
פי אחיא אלמותי פי דאר אלדניא	في إحياء الموتى في دار الدنيا
קאל צאחב אלכתאב, אמא אחיא אלמותי	قال صاحب الكتاب, اما إحياء الموتى
אלד'י ערפנא רבנא אנה יכון פי דאר אלאכ'רה	الذي عرّفنا ربنا انه يكون في دار الآخرة
ללמג'אזאה פד'לך ממא אמתנא מג'מעה עליה	للمجازاه فذلك مما امتنا مجمعة عليه
ואצל אג'מאעהם ללמעני אלמקדם ד'כרה פי	واصل اغماعهم للمعنى المقدم ذكره في
אלמקאלאת אלאואיל, לאן אלמקצוד מן ג'מיע...	المقالات الإوائل, لأن المقصود من جميع...

As part of our investigation of the use of automatic tools for providing a complete Arabic translation for Judeo-Arabic texts, we have begun working on identifying code-switching points between Judeo-Arabic and Hebrew. As mentioned above, Judeo-Arabic texts are often peppered with Hebrew citations and borrowings, which cannot be transliterated into Arabic, but rather need to be *translated* from Hebrew to Arabic. To identify language boundaries, we have built a sequential classifier that works at the word level, considering both character-level and word-level features calculated for the surrounding words. The classifier is supervised by a relatively large set of Judeo-Arabic sentences, extracted from various sources, in which Hebrew words have been marked accordingly. This dataset was recently made publicly available online by the Friedberg Jewish Manuscript Society (http://www.jewishmanuscripts.org). This aspect of our project is still in its initial stages; however, already, for some of the sources, we have been able to predict its Hebrew words with precision and recall above 95%.

References

Judah Halevi. ca. 1140. *The Kuzari - In Defence of the Despised Faith*; transliterated, translated and annotated by Nabih Bashir, Beirut and Baghdad: Al-Kamel Verlag, 2012.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin and Evan Herbst. 2007. **Moses**: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL)*, demonstration session, Prague, Czech Republic. http://www.statmt.org/moses/

Philipp Koehn, Franz Josef Och and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In

Proceedings of the Human Language Technology Conference (HLT-NAACL 2003), pp. 48-54, Edmonton, Canada.

Robert Parker, David Graff, Ke Chen, Junbo Kong and Kazuaki Maeda. 2011. Arabic Gigaword, 4th ed. Linguistic Data Consortium, LDC2009T30, Phila., PA.

Acknowledgement

We gratefully thank Nabih Bashir and Yonatan Belinkov for their advice and help in this project.