Towards Automatic Cataloguing of Hebrew Manuscripts

Miller, Hadar, Haifa University; Philips, Yoav, Haifa University; Prof. Kuflik, Tsvi, Haifa University; Dr. Lavee, Moshe, Haifa University; Prof. Dershowitz, Nachum, Tel Aviv University; Londner, Samuel, Tel Aviv University.

Abstract

Hebrew manuscripts contain the treasures of Jewish culture from biblical times to the present day. More than 1.2 million images representing almost 90,000 different manuscripts, and above 130,000 different identified works, as well as numerous unidentified texts are currently available through National Library Ktiv website. Many manuscripts contain content that has not yet been published, studied, and sometimes not even cataloged at all, or are cataloged very generally, leaving large space for the exploration of unknown texts and textual witnesses.

The proposed lecture will describe another step in a long process aimed at full textual accessibility of Hebrew manuscripts. Following the successful digitization of manuscripts (i.e., Ktiv, the Friedberg Project), and the development of tools aimed at crowdsourced, or AI based transcription (i.e., Scribes of the Cairo Genizah, Tikun Sofrim, eScriptorium) we turn to implementation of text reuse detection tools on imperfect machine/crowdsourced texts for the sake of automatic manuscript cataloging and/or feedback for improving machine readings and models.

Such automatic cataloging allows for accurate mapping of the manuscripts down to the line level, identifying the compositions and specific paragraphs included in the manuscript. In the case of unfamiliar texts, it is expected to indicate related texts, thereby classifying the genre and realm of the manuscript.

We will present the current state in the development of a text reuse detection framework, aimed at allocating text reuse of large texts units, from a few sentences long to several passages or even an entire manuscript. (Other features of the framework, such as identifying short text, as in the case of biblical citations will not be discussed at length.)

We defined the manuscript line, 5-15 words long, as a granular unit for text reuse to be detected. On the one hand, it is big enough to quickly detect suspected sources. On the other hand, it is short enough to identify multiple sources that might be integrated into the same manuscript.

We designed a three-phase framework: The first phase locates all possible suspect that share a single word or more with the tested manuscript row while allowing minor orthographic alterations between the texts. We utilize a positional inverted index for a quick search for all suspects and use edit distance to grade the suspects and choose the most probable candidates; the second phase aligns the score set for the tuple (source, row) compared to the sources located for neighboring rows. The underlying idea is that

extended text reuse increases the probability of consecutive rows to share the same source; The third phase generates a full synopsis of the three most probable suspects and the manuscript line. The synopsis algorithm measures the probability of two words to be aligned even if they are not identical, aimed at assessing whether the difference reflects a genuine variant of the text or a mistake in the automated reading of the manuscript.

The process we are developing is expected to provide detailed catalogs of manuscripts, to reveal unknown witnesses of texts, and to optimize the processes of automated transcription of Hebrew manuscripts.