Exploiting Synonym Choice to Identify Discrete Components of a Document

Navot Akiva, Dept. of Computer Science, Bar-Ilan University, Ramat Gan
Idan Dershowitz, Dept. of Bible, Hebrew University, Jerusalem
Moshe Koppel, Dept. of Computer Science, Bar-Ilan University, Ramat Gan
{navot.akiva,dershowitz,moishk}@gmail.com

When studying ancient texts, scholars often contend with documents that appear to be composite. A key challenge is to tease apart the various constituents.

One notable example of such a text is the Pentateuch, in which many scholars have found what they think are discrete narrative threads. The most prominent theory relating to the literary history of the Pentateuch is known as the "Documentary Hypothesis."

In some cases, scholars have at their disposal several widely divergent manuscripts, providing them with valuable data to better approach the primary text or texts. But when the available manuscripts are less obliging, the work of analyzing composite texts is generally done in an impressionistic fashion. Factors such as repetitions, contradictions, or possible interruptions in narrative flow, play a large part in scholars' considerations. But what is for one scholar an intolerable repetition or contradiction, is for another an instance of sophisticated literary variation. We propose to set this work on a firm algorithmic basis by identifying an optimal stylistic sub-division of a given manuscript. We do not concern ourselves with how or why such distinct threads might exist.

The most straightforward way to divide a potentially composite document is to represent segments of text as numerical vectors reflecting the frequencies of lexical features and to use clustering algorithms to find natural clusters. However, this method tends to divide texts topically rather than stylistically. Limiting features to function words is inadequate, and in tests on the books of Jeremiah and Ezekiel, we find that clustering on function words fails to separate out the two books.

Our main innovation is the use of synonym choice. Our hypothesis is that different literary works should differ in the proportions with which different synonyms in the same synset are found. By focusing our attention only on words that have synonymous counterparts in the same set of books, we can be relatively confident that the resulting division will not be according to topic. If one author speaks of a "big" house and another of a "large" one, the difference between the two is not subject matter, but personal preference or style.

We leverage very precise translations of the Bible as well as manual sense tagging for the Bible to automatically identify sets of synonyms. The automatically generated synonym set list is then manually cleaned of obvious errors. We also use a specially designed similarity measure that captures the extent to which different passages make similar/different synonym choices. This method separates Jeremiah and Ezekiel very well.

There is one additional hurdle that must be handled. Initially, we used the standard chapters as our natural units. But these units may not be pure; a single chapter might be a mix of two or more literary strands. Thus, we develop several new algorithms for automatically identifying literary boundaries.

Results show that optimal separation of the Pentateuch into two clusters roughly correlates with the portions identified by Bible scholars as P and non-P.