Modeling the Scholars: Detecting Intertextuality through Enhanced Word-Level N-Gram Matching

Neil Coffee, Christopher Forstall, Thomas Buck, Katherine Roache, Sarah Jacobson

Forthcoming *LLC* 2014.

Abstract

The study of intertextuality, or how authors make artistic use of other texts in their works, has a long tradition, and has in recent years benefited from a variety of applications of digital methods. This paper describes an approach to detecting the sorts of intertexts that literary scholars have found most meaningful, as embodied in the free Tesserae website http://tesserae.caset.buffalo.edu/. Tests of Tesserae Versions 1 and 2 showed that word-level n-gram matching could recall a majority of parallels identified by scholarly commentators in a benchmark set. But these versions lacked precision, so that the meaningful parallels could only be found among long lists of those that were not meaningful. The Version 3 search described here adds a second stage scoring system that sorts found parallels by a formula accounting for word frequency and phrase density. Testing against a benchmark set of intertexts in Latin epic poetry shows that the scoring system overall succeeds in ranking parallels of greater significance more highly, allowing site users to find meaningful parallels more quickly. Users can also choose to adjust recall and precision by focusing only on results above given score levels. As a theoretical matter, these tests establish that lemma identity, word frequency, and phrase density are important constituents of what make a phrase parallel a meaningful intertext.