Modelling the Interpretation of Literary Allusion with Machine Learning Techniques

N. Coffee[1], J. Gawley[1], C.W. Forstall[1], W.J. Scheirer[2,3]
D. Johnson[4], J.J. Corso[4], and B. Parks[2]

1. Dept. of Classics, State Univ. of New York at Buffalo
2. Dept. of Computer Science, Univ. of Colorado, Colorado Springs
3. Dept. of Molecular and Cellular Biology, Harvard Univ.
4. Dept. of Computer Science and Engineering, State Univ. of New York at Buffalo

A Computational Perspective on Allusion

Most literary allusion, the deliberate evocation by one text of a passage in another, is based upon text reuse. Yet most instances of textual similarity are not meaningful literary allusions. The goal of the Tesserae project (http://tesserae.caset.buffalo.edu) is to automatically detect allusion in a corpus of literary texts, primarily Classical Latin poetry. We begin with a large set of textual parallels, and then attempt to model which of these instances of text reuse are meaningful literary allusions and which are not, according to a group of human readers. While initial attempts with a few basic textual features have proven surprisingly effective, here we employ a more complex feature set and machine learning techniques drawn from the field of computer vision in an attempt to improve the results. Novel applications of machine learning, beyond the well known but constrained textual classification tasks of attribution and categorization, have the potential to be transformative for complex analysis tasks in the Digital Humanities.

Benchmark Data

As an illustration, we consider textual parallels between Book 1 of Lucan's *Bellum Civile* and the entirety of Vergil's *Aeneid* [2]. Our benchmark dataset comprises a list of 3,400 pairs of sentences that share at least two different words. Each of these pairs has been read and graded for its literary significance by a group of students and faculty working in small teams. These annotator rankings range from 1 (no literary significance) to 5 (pointed literary allusion).

*Work supported by NEH Start-Up Grant Award No. HD-51570-12

Learning Relevant Features

Earlier work showed that high-ranked parallels could be distinguished from the others with modest accuracy using only word frequency, distance between words, and the presence of exact form matching versus differently-inflected forms of the same word [3]. Nevertheless, others have recommended more sophisticated approaches to this problem [1]. Here we consider an expanded feature set including bi-gram frequency, frequency of individual words, character-level n-grams and edit distances. Our goal is to learn relevant combinations of features in the presence of often incomplete data.

Recent work by members of our team has developed new methods for tuning machine-learning using support vector machines [4] and random forests [6]. Random forest is of particular interest, providing robust feature selection that shows promise for literary analysis [5]. The problem of missing data is prevalent in all areas of literary study, but is not well addressed by existing algorithms in common use by digital humanists. This is especially true for ancient texts, where we often find a significant gap in the manuscript tradition. Using principled strategies for imputation and marginalization, we reduce the impact on the results.

Results and Implications

Our ability to learn the difference between high-ranked parallels (ranks 4 & 5) and low-ranked parallels (ranks 1 & 2) for *Bellum Civile* and the *Aeneid* is strong: random forest achieves an average AUC score between 82% and 83%, while linear SVMs yield an average score of 81.5%. This suggests that quantifiable patterns do exist across allusions, which can be captured algorithmically. In this ongoing research we seek a more successful model of literary significance that will allow our software to put interesting allusions at the top of the list; at the same time, we hope it will also cast new light on the underlying structures of our experience of literature.

An interactive demonstration of the Tesserae allusion detection tool accompanies this poster.

References

[1] D. Bamman and G. Crane. The Logic and Discovery of Textual Allusion. *LaTeCH*, 2008.

[2] N. Coffee, J.-P. Koenig, S. Poornima, C.W. Forstall, R. Ossewaarde, and S. Jacobson. Intertextuality in the digital age. *Transactions of the American Philological Association*, 142(2), 2012.

[3] J. Gawley, C.W. Forstall, and N. Coffee. Evaluating the literary significance of text re-use in latin poetry. *DHCS*, 2012.

[4] W.J. Scheirer, A. Rocha, J. Parris, and T.E. Boult. Learning for meta-recognition. *IEEE T-IFS*, 7, August 2012.

[5] T. Tabata. Approaching Dickens' style through random forests. *DH*, 2012.

[6] C. Xiong, D. Johnson, R. Xu, and J. J. Corso. Random forests for metric learning with implicit pairwise position dependence. *ACM SIGKDD*, 2012.