Tracking Literary Reputation with Text Analysis Tools*

Maite Taboada
Associate Professor
Department of Linguistics
Simon Fraser University
8888 University Dr.
Burnaby, BC V5A 1S6
mtaboada@sfu.ca

Mary Ann Gillies
Associate Professor
Department of English
Simon Fraser University
8888 University Dr.
Burnaby, BC V5A 1S6
gillies@sfu.ca

Paul McFetridge
Associate Professor
Department of Linguistics
Simon Fraser University
8888 University Dr.
Burnaby, BC V5A 1S6
mcfet@sfu.ca

Robert Outtrim
Post-Graduate Research Assistant
Department of English
Simon Fraser University
8888 University Dr.
Burnaby, BC V5A 1S6
outtrim@alumni.sfu.ca

Abstract

We describe the initial stages of a project tracking the literary reputation of authors. The objective of our research is to extract information on the reputation of different authors, based on writings concerning the authors. The project, a collaboration involving researchers in the Departments of English and Linguistics at Simon Fraser University, aims to create a database of texts, and computational tools to extract content automatically. It is in its initial stages, dealing with data collection and the construction of tools to extract sentiment from texts. In this paper, we will present the current state of the system, and illustrate it with some examples.

Research on opinion and subjectivity in text has grown considerably in the last few years. New methods are being created to distinguish objective from subjective statements in a text, and to determine whether the subjective statements are positive or negative with respect to the particular subject matter. We believe that the methods currently being used to extract subjective opinion, or sentiment, from movie and consumer product reviews (e.g., Hu and Liu, 2004; Turney, 2002) can be applied to literary reviews and other texts concerning an author's works.

The quantitative aspects of the project are based on research in information retrieval and text categorization. We are scanning documents pertaining to the authors in this study into a computer database designed to store them, and we will then analyze these documents automatically for positive and negative content, i.e., the document's overall *sentiment*. This problem has been characterized as one of determining whether the text is "thumbs up" or "thumbs down" (Turney, 2002), although a continuum analysis is probably more suited to literary reputation.

* To be presented at the Meeting of the Society for Digital Humanities. Vancouver, June 2008.

The question of why writers' works, and by extension their literary reputations, fall in and out of critical and popular favour has long fascinated literary critics. In 1905, Marie Corelli was the best-known and most successful novelist in Britain. By 1950 she had been consigned to literary obscurity and few read her books. D.H. Lawrence did not enjoy wide recognition during his lifetime, yet he is now part of the English literary canon. How do we account for such shifts in literary reputation? These two questions form the core of our project, on literary reputation in Britain between 1900 and 1950.

We are currently conducting a pilot project with two authors: John Galsworthy and D.H. Lawrence. We have in mind a larger project, with more authors. For the larger project, we have selected six writers: three who were very successful in the public discourse (financial and/or critically) in the early years of the 20th century and who had largely been consigned to the margins of literary study by 1950—John Galsworthy, Arnold Bennett, and Marie Corelli; and three who were less well known at that time but who came to occupy central places in the literary canon by 1950—Virginia Woolf, Joseph Conrad, and D.H. Lawrence.

Our specific concern will be to create a database of English language published material on each of the six writers in the period 1900-1950. We are not concerned with "creative" or "imaginative" literature written by the six, but with reviews, newspaper articles, magazine or periodical press articles (critical or scholarly) either written by the six or on the six. We will enter/scan all items into the database thereby creating a very large data set of information. The database will also house the bibliographical information on each item we obtain. This information will then be mounted on the Simon Fraser University Library's Electronic Document Centre where it will be available for use by other scholars. This part of the project will require that the text already scanned into the database be coded—using either HTML or XML—so that it can be made available on the web.

The computational aspect of the project involves processing documents and extracting sentiment from them. A number of techniques have been proposed for the problem of automatic sentiment classification, based on adjective classification (Hatzivassiloglou and McKeown, 1997), subjective content (Wiebe, 2000), or machine learning methods (Pang et al., 2002). In all cases, the most difficult problem consists of finding the relevant parts of the text, those that contain subjective evaluation. We propose to apply our knowledge of text structure, and to use discourse parsing, a method that parses the discourse structure of the text, to establish main and secondary parts.

An accurate identification of semantic orientation requires analysis of units larger than individual words; it requires understanding of the context in which those words appear. To this end, we intend to use Rhetorical Structure Theory to impose on the text a structure that indicates the relationships among its rhetorical units. In particular, we want to distinguish units that are nuclei (most important) from those that are satellites (secondary or contributing parts) so that their respective contributions can be appropriately calculated. Our current efforts involve building a system that can automatically extract the structure of a text. In this paper, we will describe how such a system can contribute to extracting sentiment. A pilot project on extracting sentiment from movie reviews shows that this method of extracting sentiment only from relevant parts performs better than a simple aggregation method (Voll and Taboada, 2007). The results of this pilot project show that we can extract information about discourse structure and topical sentences relatively easily. The method is now being applied to the literary reviews in the Galsworthy-Lawrence pilot project.

The final goal of our project is to be able to determine what in a reviewer's text seems to influence the literary reputation of a particular author, and whether what reviewers say can be mapped to the author's reputation trajectory. The most novel aspect of the project it the use of objective tools to extract highly subjective material.

In the process of constructing the database and creating tools we need to address a number of issues relating to the nature of the project itself and to the application of existing methodologies. First of all, we focus on quantifiable aspects of reputation, and how it varies according to audience. Each text is annotated with the reviewer's name, source, place of publication, and audience type. Thus, we can measure the impact of specific reviews.

Secondly, we are applying tools created for present-day texts (e.g., part-of-speech taggers and parsers). Our impression is that some modifications are necessary to account for the slightly different style of a different time and register. A related problem is the heavy use of irony and external references as indicators of evaluation.

Finally, the most important challenge lies with the evaluation of the system. The current evaluation of our present-day movie review system relies on whether the author recommended the movie or not. In literary reviews, the overall evaluation is much more subtle. We need to determine whether the system is flagging each review correctly as positive or negative, and also the review's contribution to the author's literary reputation.

References

Hatzivassiloglou, Vasileios, and Kathleen McKeown, 1997. Predicting the semantic orientation of adjectives, *Proceedings of 35th Meeting of the Association for Computational Linguistics*, Madrid, Spain, pp. 174-181.

Hu, Minqing, and Bing Liu, 2004. Mining and summarizing customer reviews, *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004)*, Seattle, WA.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan, 2002. Thumbs up? Sentiment classification using Machine Learning techniques, *Proceedings of Conference on Empirical Methods in NLP*, pp. 79-86.

Turney, Peter, 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews, *Proceedings of 40th Meeting of the Association for Computational Linguistics*, pp. 417-424.

Voll, Kimberly, and Maite Taboada, 2007. Not all words are created equal: Extracting semantic orientation as a function of adjective relevance, *Proceedings of the 20th Australian Joint Conference on Artificial Intelligence*, Gold Coast, Australia, pp. to appear.

Wiebe, Janyce, 2000. Learning subjective adjectives from corpora, *Proceedings of 17th National Conference on Artificial Intelligence (AAAI)*, Austin, Tx, pp. 735-740.