Sentiment Classification Techniques for Tracking Literary Reputation*

Maite Taboada[1], Mary Ann Gillies[2] and Paul McFetridge[1]

[1]Department of Linguistics, [2]Department of English

Simon Fraser University

8888 University Dr., Burnaby, BC, V5A 1S6, Canada

E-mail: mtaboada@sfu.ca, gillies@sfu.ca, mcfet@sfu.ca

Abstract

The initial stages of a project tracking the literary reputation of authors are described. The critical reviews of six authors who either rose to fame or fell to obscurity between 1900 and 1950 will be examined and we hope to demonstrate the contribution of each text to the evolving reputations of the authors. We provide an initial report on the use of the semantic orientation of adjectives and their rough position in the text to calculate the overall orientation of the text and suggest ways in which this calculation can be improved Improvements include further development of adjective lists, expansion of these lists and the consequent algorithms for calculating orientation to include other parts of speech, and the use of Rhetorical Structure Theory to differentiate units that make a direct contribution to the intended orientation from those that are contrastive or otherwise make an indirect contribution.

1. Introduction

The objective of our research is to extract information on the reputation of different authors, based on writings concerning the authors. The project aims to create a database of texts, and computational tools to extract content automatically.

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., Gamon, 2004; Hu & Liu, 2004; Turney, 2002) can be applied to literary reviews and other texts concerning author's works.

In this paper, we describe some of the methods currently being used to extract sentiment from text, and explain how we are applying those methods to literary reviews, letters to the editor, newspaper articles, and critical and scholarly publications concerning six authors who were active in the 1900-1950 period. Section 2 provides some background on literary reputation, and how we plan to quantify it. Section 3 discusses sentiment detection, as it has been applied to movie reviews and other present-day reviews of consumer reports. In Section 4, we address the issue of document structure: how important it is to identify the most important parts of the text, and what methods we can use to that end. This project is in its initial stages, and we do not have conclusive results yet. We present, however, the current state of the system in Section 5, and illustrate it with two examples in Section 6. Finally, conclusions and a discussion of future work are found in Section 7.

2. Background

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. In 1910, T.S. Eliot was an unknown American poet in Paris, dreaming of "belonging in a great centre of artistic and intellectual innovation" (Gordon, 1977: 33). By 1950 Eliot, a Nobel Laureate, stood at the very centre of Western aesthetic and intellectual culture. Why had these two writers' reputations suffered such dramatically opposite fates? 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.

Scholarly discussions of publishing, readership, canon construction, and the various institutions of literature have proliferated in recent years, most of which attempt to map out how "our experience of the work" (Herrnstein Smith, 1988: 16) relates to its critical or popular value (Fromm, 1991; Guillory, 1993; Lecker, 1991; Remplin, 1995). And yet in literary studies, few of these discussions attempt to combine a quantitative analysis of data with a qualitative analysis. An exception is Gaye Tuchman & Nina Fortin's *Edging Women Out* which sets out to answer the question "Why does some literature supposedly transcend the ages and so constitute 'culture' while other once-popular books languish in disuse?" (Tuchman & Fortin, 1989: 1). Tuchman & Fortin focus on one publisher, Macmillan, from 1867-1917. They designed a quantitative study of Macmillan's records, identifying four distinct data sets and applying a

* In Proceedings of LREC 2006 Workshop "Towards Computational Models of Literary Analysis", pp. 36-43.

systematic analysis of the records in order to derive conclusions about the "literary opportunities" of women at the turn of the century. Tuchman & Fortin admit, however, "Although our data about the literary opportunities of most women novelists are substantial, our conclusions are based on inferences." (Tuchman & Fortin, 1989: 18). Our project asks similar questions to Tuchman & Fortin and Herrnstein Smith, but we have designed it so that it permits us to combine the aesthetic and evaluative concerns raised by the former with the kinds of quantitative methodology employed by the latter.

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).

A number of techniques have been proposed for the problem of automatic sentiment classification, based on adjective classification (Hatzivassiloglou & McKeown, 1997), extraction of subjective content (Wiebe et al., 2004), or through the use of machine learning methods (Bai et al., 2004; Gamon, 2004; 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, establishing main and secondary parts.

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.

We selected the time period 1900-1950 for two reasons. First, the advent of mass market publications around the turn of the century created new ways of producing and disseminating literature—for example, cheap paperback novels and tabloid newspapers helped transform the very definition of literature; at the same time, they focused ever greater attention on individual authors. Writers and readers came to view literature as something very different than had their Victorian parents thus making 1900 a marker of a crucial sea change in literary studies. Second, another major shift occurred around 1950. Here technology also played a leading role: the advent of television and vinyl recordings brought writers into people's homes in ways never before possible, thereby solidifying the celebrity status of authors. The influence of the educational establishment in post war society is also important; university syllabi, designed by writers and critics whose vested interests were served through creating a canon that fit their definitions of what "great" literature was, created a publishing demand for these very writers. The result was a wholesale shift away from the writers who were prominent at the beginning of the century towards those who were notable for their marginal status in the 1900-1920 period.

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 next few sections describe how we process the texts once they have been scanned, and how we are extracting information from the texts that we hope will shed light on how literary reputation is built or destroyed.

3. Sentiment Classification: Semantic Orientation of Words

The problem of extracting the semantic orientation (SO) of a text (i.e., whether the text is positive or negative towards a particular subject matter) often takes as a starting point the problem of determining semantic orientation for individual words. The hypothesis is that, given the SO of relevant words in a text, we can determine the SO for the entire text. We will see later that this is not the whole or the only story. However, if we assume that SO for individual words is an important part of the problem, then we need lists of words with their corresponding SO, since such information is not typically contained in a traditional dictionary. The expressions "semantic orientation", "sentiment", and "opinion" are used in this paper to refer to the subjective evaluation conveyed by a word, a phrase, a sentence, or an entire text.

One approach is to manually compile a list of words that are known to express sentiment, and annotate them according to whether the sentiment is positive or negative. One such list is the one contained in the General Inquirer, a content analysis program (Stone, 1997; Stone et al., 1966). The General Inquirer contains lists of words, classified according to specific categories, such as "self-reference", "strong", "active", or abstract concepts (words relating to objects, places, institutions, etc.). Of interest to sentiment detection are two tags that indicate whether the word is positive or negative. These have been used to determine whether the majority of words in a text are either positive or negative.

Whitelaw et al. (2005) use a semi-automatic method to create a dictionary of words that express appraisal.

Appraisal is a functional framework for describing evaluation in text: how personal feelings, judgement about other people, and appreciation of objects and art are expressed (Martin & White, 2005; White, 2003). Whitelaw and colleagues compiled a list of appraisal words from the literature on appraisal, and extended it automatically by extracting synonyms and related words from WordNet (Fellbaum, 1998) and on-line thesauri. Other researchers have explored this avenue, extracting synonyms using either Pointwise Mutual Information (Turney, 2001) or Latent Semantic Analysis (Landauer & Dumais, 1997). It is unclear which method provides the best results; published accounts vary(Rapp, 2004; Turney, 2001). Word similarity may be another way of building dictionaries, starting from words whose SO we already know. For this purpose, WordNet is a valuable resource, since synonymy relations are already defined (Kamps et al., 2004). Esuli and Sebastiani (2005) also use synonyms, but they exploit the glosses of synonym words to classify the terms defined by the glosses.

Manual and semi-automatic methods, although highly accurate, are not ideal, given that it is time-consuming and labour-intensive to compile a list of all the words that can possibly express sentiment. Researchers have turned to automatic methods to "grow" dictionaries of sentiment words, out of a few words. Most research in this area has focused on adjectives. Adjectives convey much of the subjective content in a text, and a great deal of effort has been devoted to extracting SO for adjectives. Hatzivassiloglou & McKeown (1997) pioneered the extraction of SO by association, using coordination: the phrase *excellent and X* predicts that *X* will be a positive adjective. Turney (2002), and Turney & Littman (2002; 2003) used a similar method, but this time using the Web as corpus. In their method, the adjective *X* is positive if it appears mostly in the vicinity of other positive adjectives, not only in a coordinated phrase. "Vicinity" was defined using the NEAR operator in the Altavista search engine, which by default looked for words within ten words of each other. The contribution of Turney & Littman was to find a way to not only extract the sign (positive or negative) for any given adjective, but also to extract the strength of the SO, expressed in a number (e.g., 2.2 is more positive than 1.3). They use Pointwise Mutual Information (PMI) for that purpose. PMI calculations do not have to be limited to adjectives. In fact, Turney (2002) used two-word combinations that included Adjective+Noun, Adverb+Noun, and Adverb+Verb.

Pang et al. (2002) propose three different machine learning methods to extract the SO of adjectives. Their results are above a human-generated baseline, but the authors point out that discourse structure is necessary to detect and exploit the rhetorical devices used by the review authors. Machine Learning methods have also been applied to the whole problem, i.e., the classification of whole text as positive or negative, not just the classification of words (Bai et al., 2004; Gamon, 2004).

We have tested a number of methods for creating SO dictionaries, in part motivated by the fact that Altavista no longer allows searches with the NEAR operator (Taboada et al., 2006). We tested whether an AND search, where the two words can be found anywhere in a document, not just close to each other, would be useful for the task. The AND searches were performed using the Google search engine. Our results show that NEAR-created dictionaries outperform AND-based ones in the task of extracting sentiment. The tests were performed on reviews of movies and other consumer products. However, our results indicate that variability in the number of hits returned by Google (since it indexes a dynamic space) affects the quality of the dictionary.

In summary, SO dictionaries are actively being created. Although no perfect method for compiling one exists, progress is being made, and we can expect better methods and larger dictionaries in the near future.

4. Document Structure

Research in subjective evaluation of text has not taken into account text structure, most of it relying on the content of adjectives, such as *great* or *poor* (e.g., Turney, 2002). However, adjectives have different meanings according to their linguistic context, whether immediate: *a huge disaster* vs. *a huge success*, or more remote: *The movie is great, if you're looking for reasons to be depressed*. In the latter example, it is important to know that the positive evaluation (*the movie is great*) is hedged by a condition on it. Previous work on movie reviews has revealed a common argumentation device, whereby authors list a number of positive aspects, to end with a negative summary. Example (1) illustrates the strategy[1]: the author lists a number of positive qualities for the movie "The Last Samurai". He or she, however, finishes with a clear negative evaluation. The concession structure ("good in some aspects, but overall bad") is very common in reviews, especially those found on-line.

(1) [1] It could have been a great movie. [2] It could have been excellent, and to all the people who have forgotten about the older, greater movies before it, will think that as well. [3] It does have beautiful scenery, some of the best since Lord of the Rings. [4] The acting is well done, [5] and I really liked the son of the leader of the Samurai. [6] He was a likeable chap, [7] and I hated to see him die. [8] But, other than all that, this movie is nothing more than hidden rip-offs.

It is obvious that we need to understand the overall structure of the text, and especially the concessions and conditions that authors attach to their opinions. For that purpose, we need to parse the entire structure of the text. Discourse parsing is analogous to sentence parsing: elements of the text are tagged, and incorporated into a tree that captures the dependencies found in the text.

Discourse parsing in this project is based upon Rhetorical Structure Theory (Mann & Thompson, 1988). RST is one of the most successful theories of discourse structure, in part because it lends itself well to computational implementations: it has been used in parsing and natural language generation, and in text summarization. A rhetorical, or discourse, relation is one that holds between two non-overlapping text spans, called nucleus and

[1] From the website Epinions.com. The text is reproduced verbatim. We have only added unit numbers (in square brackets).

satellite. Some relations are also multinuclear, consisting of two spans that are equal in importance. The nucleus contains the most important information, whereas the satellite supports or enhances that information. Spans are typically clauses in their minimal composition, but they are also built incrementally, so that a span may consist of different clauses, with their own internal structure. Multinuclear relations are analogous to paratactic or coordinate structures, whereas nucleus-satellite relations resemble hypotactic or subordinate relations.

There are different types of relations, based on the type of information or intention expressed: Condition, Contrast, Concession, Cause, Background, etc. Rhetorical relations can be represented in the form of trees, which have the following properties: completeness, uniqueness, connectedness and adjacency. Trees represent contiguous text, and the tree schemas can be applied recursively, to represent an entire text of arbitrary length.

The whole text in Example (1) above can be captured in a single relation: spans 1-7 are the satellite (i.e., the subordinate or less important part) to the nucleus presented in 8. The overall relation is one of Concession, as shown in Figure 1. The arrow pointing from 1-7 to 8 indicates that 8 is the nucleus, the most important part in the Concession relation. Spans 1-7 have further internal structure, which we could also analyze using RST.

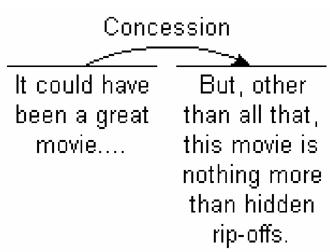

Figure 1. General structure for Example (1)

Unfortunately, a full discourse parser based on RST (or any other theory) does not exist yet. Soricut & Marcu (2003) created a sentence-level parser, trained on data from the RST Treebank (Carlson et al., 2002), a collection of articles from the Wall Street Journal. We have been testing this parser, which creates trees for individual sentences (but not for the full text). Our results are quite poor so far, probably due to the very different text genres. Current research aims to improve sentence-level parsing, and to create a corpus of manually-annotated reviews, in order to train a full whole-text parser.

The results of such parsing would help distinguish main from secondary parts of the text. There is a significant amount of research on how RST can be used to summarize text, exploiting the discourse structure to prune the less important parts (Marcu, 2000). Our plan is to use it for a dual purpose: (i) to pinpoint the most important parts of the text; and (ii) to calculate the aggregation of nuclei and satellites. In Example (1), that would mean, first of all, to identify spans 1-7 and span 8 as the main parts of the text, with span 8 as the nucleus of the relation between the two. In addition, the analysis tells us that the relation between those two spans of text is one of Concession. That means that there is a discrepancy in the situations, events or opinions expressed by each span. In the example, we see that the first part of the text contains a large number of positive words and phrases (*great, excellent, beautiful, some of the best, well done, likeable*), but the weight of those must be decreased in the final aggregation, because they are in the satellite of a Concession relation, and the most important part, what the author wanted to convey, is that the movie contains *hidden rip-offs*, a negative phrase.

RST classifies parts of a text according to a number of relations. The number and types of relations are often based on those proposed by Mann and Thompson (1988), but extensions and modifications are possible [2]. In addition, a higher-level classification could be imposed, dividing the text into stages, or parts, typically determined by the text genre (Eggins & Martin, 1997). For example, in present-day reviews of movies, there is usually a clear structure: introduction of the movie, plot, actors, director, background (e.g., other movies by the same director or cast), and evaluation. Segmenting each text into these stages would help identify the parts that contain an actual evaluation of the work, and not of the characters. RST has been integrated into genre analysis for other genres (Taboada, 2004a, 2004b), and could be easily integrated into the literary review genre and other genres in this project.

5. Processing Documents

The documents are first tagged with parts of speech (adjective, noun, verb). The words with subjective content are extracted and compared to a custom-built lexicon of words annotated with evaluation tags (i.e., positive for the word *excellent*, negative for the word *poor*). This electronic dictionary (or lexicon) assigns numeric values to words in the text (e.g., 5 for *outstanding*, -5 for *appalling*). The lexicon is being built partly automatically, based on the context of those words in documents found on the Internet (Turney & Littman, 2002). We are testing different methods of creating the dictionary (Taboada et al., 2006). We have already applied some of these methods to the problem of extracting sentiment from reviews about movies and consumer products (Taboada & Grieve, 2004). Our current dictionary contains 3,314 adjectives, whose semantic orientation was calculated using AND searches on Google. As described in our previous work, the values in the dictionary are normalized, so that 0 is the median value for the entire dictionary.

The final step in the process is to devise an algorithm to aggregate the negative and positive words in the document. We are currently using a weighted average of the adjectives in the text. Weights are assigned according to whether the adjective appears in the first, second, or last third of the text, as shown in Figure 2 (Taboada & Grieve, 2004). The intuition behind these weights is that authors tend to summarize or repeat their opinions

[2] Each relation in RST has a formal definition. Definitions and examples for the most common relations can be found on the RST website (Mann, 2005).

towards the end of the text. We also take negation into account, changing the sign of an adjective in the scope of a negating word (e.g., *not, no, nor, neither*). Negating words are considered within scope if they are found up to five words to the left of the adjective.

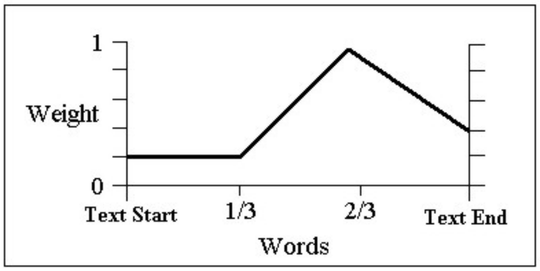

Figure 2: Weights given to adjectives

Future work involves a discourse analysis of the texts, to examine the types of patterns that signal the presence of subjective content; and a method to determine the contribution of different rhetorical relations to a text's sentiment. Words other than adjectives will also be considered, as long as they convey sentiment. 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.

6. Two Examples

Since we are describing work in progress, we do not yet have large-scale quantitative results. In this section, we show a detailed analysis of two documents, one for each author, explaining what processing was carried out, and the current results.

The documents are reviews of (at the time) recently published works by the two authors. The review of John Galsworthy's plays (*A Bit o' Love, The Foundations* and *The Skin Game*) was published June 26, 1920, in the *Saturday Review* (Anonymous, 1920). The second document is a review of D.H. Lawrence's *The White Peacock*, published March 18, 1911, in *The Academyand Literature* (Anonymous, 1911). The Galsworthy text comments on the work of an established artist, and issues quite a damning criticism of his work. The text on Lawrence, on the other hand, is about an up-and-coming artist, who, up to that point, had been known only as a poet. The first one is 1,018 words long, whereas the D.H. Lawrence text contains 629 words.

6.1. Semantic Orientation for Adjectives

Space precludes a full examination of the entire texts. We will concentrate on some passages. The Galsworthy text starts with a simple statement: "For many years Mr. Galsworthy has been consistently overpraised." It ends with a summary of that opinion: "Mr. Galsworthy, in fact, remains the second-rate artist he always was." The entire text is organized around those two statements, with a lengthy elaboration of the first by way of a general criticism of Galsworthy's work (lack of creativity; he is ridden by ideas, but creates no real characters; his views are too present), and a specific example of how this is evident in the play *A Bit o' Love*.

The first process applied to the text (apart from normalization of punctuation and spacing) is the part of speech tagging. Each word is assigned a part of speech (noun, verb, adjective, determiner, etc.). Tagging is performed automatically, using Eric Brill's freely available tagger (Brill, 1995). After tagging, all words tagged as adjectives are extracted and their semantic orientation extracted from our dictionary. Example (2) shows in bold type the words that were tagged as adjectives in the first few sentences of the text, with the SO values according to the dictionary in square brackets.

(2) For many years Mr. Galsworthy has been consistently overpraised. His admirers, detecting in his **imaginative** [2.13] work—and particularly in his plays—the quality of **moral** [-2.06] earnestness, have taken him to their **susceptible** [0.03] hearts as one of the **supreme** [-0.41] artists of our time; but it is as a **creative** [4.001] artist, **pure** [-0.35] and **simple** [1.01], that he fails. He has many gifts, many qualities—**technical** [4.57] ability, imaginativeness, sympathy, experience of life, ideas, ideals; but the one **supreme** [-0.41], **essential** [2.95] gift—the ability to create living men and women working out their destinies in the grip of fate—is not his. He is ridden by his ideas, harried by his ideals; he has no spaciousness, no ease, no geniality; and his characters are invariably irritatingly **true** [0.65] to type and the instruments for their author's views on sociology, politics and what not.

One could disagree with some of the adjective values. They were calculated automatically, and according to their context in web pages indexed by Google (Taboada et al., 2006). What we would like to point out here is that many other words convey opinion: *earnestness, ability, imaginativeness* (all nouns), or *fails* (a verb). Note also that one of the most important words, *overpraised*, is not tagged as an adjective. The tagger interpreted it as a verb (a past participle), which is, strictly speaking, correct in this case. It is also clear from the example that the context, and the person or object being evaluated, are quite relevant. For instance, the word *susceptible* is applied to Galsworthy's admirers; it does not necessarily reflect upon him or his work; *pure* and *simple* are used to emphasize a statement and do not refer to any entity in the text. Finally, the word *creative* (one of the most positive in this fragment) is negated through the verb *fails.* All of those aspects (words beyond adjectives, context and sentence topic) are part of our future work.

Applying this same method to the entire text, we extracted all the adjectives, and produced a weighted average, with the final number of 0.19. This is a positive number, but quite close to the 0 level, reflecting the fact that many of the statements in the text were negative in nature.

The same procedure was carried out on the Lawrence text, of which we show a portion in (3). This text starts with a contrast between Lawrence's previous work as a poet, and what the reviewer sees as a promising novelist career. It describes *The White Peacock* in detail, and concludes by saying that "…he has given us a book of considerable achievement and infinite promise." Example (3) shows the adjectives detected by the tagger. As with the

Galsworthy text, some crucial words are missing, such as the verbs *surprises* and *charms*, and *disillusioned*, which was tagged as a past participle. The final number for the entire text was 0.25, a slightly more positive value than for the Galsworthy text.

(3) Hitherto we have only known Mr. D. H. Lawrence as being one of the many **interesting** [1.411] poets discovered by the English Review. Henceforth we shall certainly know him as the author of "The White Peacock," for it is beyond all argument an **admirable** [0.58] and **astonishing** [0.38] piece of work. We use the word "**astonishing**" [0.38] advisedly, for, like most **new** [3.59] books of **uncommon** [0.49] merit, "The White Peacock" surprises even while it charms. There are pages in it that made the **present** [1.01] reviewer, a **sophisticated** [1.62] and disillusioned reader of novels, lay down the book and rub his eyes in wonder at the author's individuality and courage.

6.2. Rhetorical parsing

The texts are next processed through a rhetorical, or discourse parser. As explained in Section 4, there is no available parser for entire texts. The only existing parser (Soricut & Marcu, 2003) is one that analyzes individual sentences, classifying their parts (main and subordinate clauses, clausal adjuncts and other clausal components) into nuclei and satellites, and then defining the type of relation between those. The parser was designed for newspaper articles, and does not work as well for these texts. Future work involves adapting it to our purposes. Let us examine, however, its current output.

The sentences in Example (2) were segmented. The first one is a simple sentence, and did not undergo further segmentation. The second sentence is quite complex, and was divided into 6 spans, as shown in Example (4), with span numbers in square brackets. The structure of the text, according to the parser, is displayed in Figure 3.

(4) [1] His admirers, [2] detecting in his imaginative work [3] —and particularly in his plays—the quality of moral earnestness, [4] have taken him to their susceptible hearts as one of the supreme artists of our time; [5] but it is as a creative artist, pure and simple, [6] that he fails.

There are quite a few problems with the analysis. Its main failure is that the relation between the two main parts is too abstractly captured as an Elaboration relation, whereas a Contrast relation would be more appropriate, rephrased as: "his admirers think of him as creative; he fails as a creative author." The segmentation itself is problematic, especially around the parenthetical remarks between dashes.

We hope that the example is sufficient to illustrate the type of analysis that we want to perform, even though the results are far from perfect at this point. The important aspect of this analysis is that it identifies nuclei and satellites in the text. As we mentioned in Section 4, we plan to use this analysis for two purposes: to extract nuclei, and to aggregate the semantic orientation of individual spans according to the relation that joins them. In the example, there are quite a few elaboration relations. The semantic orientation of the words in each of those spans (1-4) can be simply added, since they are all contributing to the same idea. However, the contrast between 1-4 and 5-6 cannot be simply aggregated.

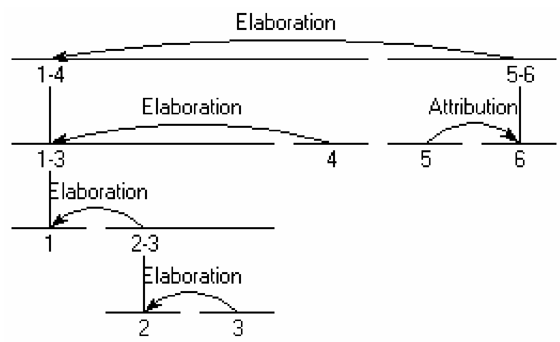

Figure 3. Rhetorical structure of one sentence

At the present time, we are not using relations to aggregate (given the fact that the parser does not yet capture them accurately). We are extracting the nuclei in the text, and calculating semantic orientation for those. For the text in Example (4), the nuclei are the fragments show in (5).

(5) [1] His admirers
[4] have taken him to their susceptible hearts as one of the supreme artists of our time;
[6] that he fails.

Nuclei for the entire text are extracted, and then the semantic orientation calculation is performed again, this time using adjectives found only in the nuclei. The Galsworthy text goes down in overall semantic orientation to -0.01. This probably reflects the fact that many of the positive adjectives are found in the satellites, or les important parts of the text. However, the same method applied to the Lawrence text yields an overall semantic orientation of 0.14, lower than the original 0.25. Such number is not an accurate reflection of the semantic orientation in the Lawrence text, since it is a generally positive review.

As is obvious from these two examples, our current system requires much further development. We are in the process of error-checking and improving each of the components, from the tagger to the adjective list (including other words than adjectives). The rhetorical parser is a very important part of that effort. We believe it can be made more efficient by improving the segmentation, and training it on examples drawn from our corpus.

7. Conclusions

This paper describes the initial stages of a project tracking the literary reputations of six authors between 1900 and 1950, and the applicability of existing techniques for extracting sentiment from texts that discuss and criticize these authors.

One of the techniques for calculating sentiment and semantic orientation that has been developed is the analysis of adjectives from the text. This can give useful results but is limited by the size and accuracy of the list of adjectives used, the accuracy of the algorithm used to identify adjectives, the ability of the algorithm to recognize the context in which the adjective appears (including the presence of negating elements and where the adjective appears in the text), the contribution to the sentiment of the text by words of other parts of speech, and the overall discourse structure of the text. Each of these limitations suggests fruitful avenues of research.

We are engaged in developing algorithms for automatically developing adjective dictionaries. Future research will expand this effort to include semantic orientation dictionaries for nouns and verbs as well. As these are developed, algorithms for integrating their contribution to the orientation of the text as a whole can be investigated.

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 from those that are satellites so that their respective contributions can be appropriately calculated.

Finally, since the overall structure of a text is often correlated with the genre of the text, we must often be sensitive to the bias that machine learning techniques can inadvertently bring. Freely available texts such as newspapers that often provide the corpus for machine learning algorithms have a consistent structure that is different from the critical reviews that we are analyzing.

8. Acknowledgments

This project is funded through a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada, and an institutional SSHRC grant from Simon Fraser University. Thanks to Katia Dilkina, Jack Grieve, Caroline Anthony and Kimberly Voll for creating some of the tools used here, and Robert Outtrim and Lee Simons for collecting, scanning and preparing the documents.

9. References

Anonymous. (1911). Review of D. H. Lawrence's 'The White Peacock'. The Academy and Literature, March 11, 328.

Anonymous. (1920). Ridden by ideas. Review of John Galsworthy's 'Plays: Fourth Series'. Saturday Review, 129, 590.

Bai, X., Padman, R. & Airoldi, E. (2004). Sentiment Extraction from Unstructured Text Using Tabu Search-Enhanced Markov Blanket (Technical Report CMU-ISRI-04-127). Pittsburgh: Carnegie Mellon University.

Brill, E. (1995). Transformation-based error-driven learning and Natural Language Processing. Computational Linguistics, 21(4), 543-565.

Carlson, L., Marcu, D. & Okurowski, M.E. (2002). RST Discourse Treebank [Corpus]. Philadelphia, PA: Linguistic Data Consortium.

Eggins, S. & Martin, J.R. (1997). Genres and registers of discourse. In T.A.v. Dijk (Ed.), Discourse as Structure and Process. Discourse Studies: A Multidisciplinary Introduction (pp. 230-256). London: Sage.

Esuli, A. & Sebastiani, F. (2005). Determining the semantic orientation of terms through gloss classification. Proceedings of the 14th ACM International Conference on Information and Knowledge Management. Bremen, Germany.

Fellbaum, C. (Ed.). (1998). WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Fromm, H. (1991). Academic Capitalism and Literary Value. Athens: University of Georgia Press.

Gamon, M. (2004). Sentiment classification on customer feedback data: Noisy data, large feature vectors, and the role of linguistic analysis. Proceedings of COLING 2004 (pp. 841-847). Geneva, Switzerland.

Gordon, L. (1977). Eliot's Early Years. Oxford: Oxford University Press.

Guillory, J. (1993). Cultural Capital: The Problem of Literary Canon Formation. Chicago: University of Chicago Press.

Hatzivassiloglou, V. & McKeown, K. (1997). Predicting the semantic orientation of adjectives. Proceedings of 35th Meeting of the Association for Computational Linguistics (pp. 174-181). Madrid, Spain.

Herrnstein Smith, B. (1988). Contingencies of Value. Harvard: Harvard University Press.

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

Kamps, J., Marx, M., Mokken, R.J. & de Rijke, M. (2004). Using WordNet to measure semantic orientation of adjectives. Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004) (pp. 1115-1118). Lisbon, Portugal.

Landauer, T.K. & Dumais, S.T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104(2), 211-240.

Lecker, R. (Ed.). (1991). Canadian Canons: Essays in Literary Value. Toronto: University of Toronto Press.

Mann, W.C. (2005). RST Web Site, from http://www.sfu.ca/rst

Mann, W.C. & Thompson, S.A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. Text, 8(3), 243-281.

Marcu, D. (2000). The Theory and Practice of Discourse Parsing and Summarization. Cambridge, Mass: MIT Press.

Martin, J.R. & White, P. (2005). The Language of

Evaluation. New York: Palgrave.

Pang, B., Lee, L. & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using Machine Learning techniques. Proceedings of Conference on Empirical Methods in NLP (pp. 79-86).

Rapp, R. (2004). A freely available automatically generated thesaurus of related words. Proceedings of 4th International Conference on Language Resources and Evaluation (LREC 2004). Lisbon, Portugal.

Remplin, C. (1995). Feminism and the Politics of Literary Reputation. Lawrence: University of Kansas Press.

Soricut, R. & Marcu, D. (2003). Sentence level discourse parsing using syntactic and lexical information. Proceedings of Human Language Technology and North American Association for Computational Linguistics Conference (HLT-NAACL'03). Edmonton, Canada.

Stone, P.J. (1997). Thematic text analysis: New agendas for analyzing text content. In C. Roberts (Ed.), Text Analysis for the Social Sciences. Mahwah, NJ: Lawrence Erlbaum.

Stone, P.J., Dunphy, D.C., Smith, M.S. & Ogilvie, D.M. (1966). The General Inquirer: A Computer Approach to Content Analysis. Cambridge, MA: MIT Press.

Taboada, M. (2004a). Building Coherence and Cohesion: Task-Oriented Dialogue in English and Spanish. Amsterdam and Philadelphia: John Benjamins.

Taboada, M. (2004b). The genre structure of bulletin board messages. Text Technology, 13(2), 55-82.

Taboada, M., Anthony, C. & Voll, K. (2006). Creating semantic orientation dictionaries. Proceedings of 5th International Conference on Language Resources and Evaluation (LREC) (to appear). Genoa, Italy.

Taboada, M. & Grieve, J. (2004). Analyzing appraisal automatically. Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text (AAAI Technical Report SS-04-07) (pp. 158-161). Stanford University, CA.

Tuchman, G. & Fortin, N.E. (1989). Edging Women Out: Victorian Novelists, Publishers, and Social Change. New Haven: Yale University Press.

Turney, P. (2001). Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. Proceedings of the 12th European Conference on Machine Learning (ECML-2001). Freiburg, Germany.

Turney, P. (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).

Turney, P. & Littman, M. (2002). Unsupervised learning of semantic orientation from a hundred-billion-word corpus (No. ERB-1094, NRC #44929): National Research Council of Canada.

Turney, P. & Littman, M. (2003). Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems, 21(4), 315-346.

White, P.R.R. (2003). An Introductory Course in Appraisal Analysis, from http://www.grammatics.com/appraisal

Whitelaw, C., Garg, N. & Argamon, S. (2005). Using Appraisal groups for sentiment analysis. Proceedings of ACM SIGIR Conference on Information and Knowledge Management (CIKM 2005) (pp. 625-631). Bremen, Germany.

Wiebe, J., Wilson, T., Bruce, R., Bell, M. & Martin, M. (2004). Learning subjective language. Computational Linguistics, 30(3), 277-308.