Tracking Literary Reputation with Text Analysis Tools

Maite Taboada, Dept. of Linguistics
Mary Ann Gillies, Dept. of English
Paul McFetridge, Dept. of Linguistics
Robert Outtrim, Independent Scholar

Simon Fraser University

Introduction

This project marries two different research tracks

- Literary reputation
  - How is reputation made or lost?
- Sentiment extraction
  - How can computational tools calculate the sentiment expressed in a document?

Literary reputation

- "Why does some literature supposedly transcend the ages and so constitute 'culture' while other once-popular books languish in disuse?" (Tuchman & Fortin 1989: 1)
- Can we correlate what is written about an author and his/her work to the author's reputation and subsequent canonicity?

Goals of the project

- Examine the critical reviews of six authors writing in the first half of the 20th century
  - Three are no longer part of the canon, although they were once considered important

John Galsworthy
(1867-1933)

Marie Corelli
(1855-1924)

Arnold Bennett
(1867-1931)

  - Three have an upward trajectory to their careers

D. H. Lawrence
(1885–1930)

Virginia Woolf
(1882-1941)

T. S. Eliot
(1888-1965)

- Map information contained in the critical texts to the authors' reputation

Sentiment extraction

- Discover whether a text is expressing positive or negative sentiment about its topic
- Employ information retrieval and text categorization methods
- Current state of the art
  - Text is treated as a bag of words
  - No consideration is given to
    - where positive and negative words occur
    - structural information within the text (e.g., introduction, conclusion)
- Proposed improvement: Make full use of the structure of the text by developing a discourse parsing tool

Materials and process

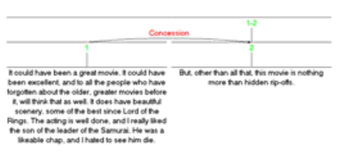
- Collect published material about the authors between 1900 and 1950
  - Literary reviews
  - Press notes
  - Magazine or periodical press articles (critical or scholarly)
  - Letters to the editor (including by the authors themselves)
- Process materials: scan, clean up scanning errors and tag
- Tags
  - Not just for a general search (TEI), but also as factors in the calculation of sentiment
  - Tag the critical author as well as the primary author
  - Publication type, audience numbers and profile, political affiliation
- Currently, pilot project with Galsworthy and Lawrence
  - 330 documents scanned (480,000 words)

Methods

- Tag documents with parts of speech (Brill 1995)
  - Develop a dictionary for literary discourse
  - Adaptation of taggers developed for present-day text to early 20th century British and American texts
- Extract relevant words (positive and negative)
- Aggregate words' semantic orientation
  - Naïve or basic method, using keywords
  - Need to take into account intensifiers (*very good*) and negation (*not very good*)
- Performance of similar methods on present-day movie reviews is about 68% accurate
- Taking text structure into account will enhance performance
- Use discourse parsing to determine
  - Subjective and objective sentences
  - Topic sentences
  - Relevance

Discourse parsing

- In this project, based on Rhetorical Structure Theory (Mann & Thompson 1988, Taboada and Mann 2006)
  - Rhetorical relations as the building blocks of text
  - They help explain coherence
  - Examples: Cause, Concession, Condition, Elaboration, Summary
- Review texts tend to have a typical rhetorical structure
  - List of pros and cons (performance reviews)
  - Opinions usually summarized at the end
  - Frequent use of concessive relations
  - Elaborations sometimes tangential
- Automated discourse parsing
  - Some preliminary work (Schilder 2002, Soricut and Marcu 2003)
  - We are developing a parsing method for literary reviews, based on our own data



It could have been a great movie. 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. It does have beautiful scenery, some of the best since Lord of the Rings. The acting is well done, and I really liked the son of the leader of the Samurai. He was a likeable chap, and I hated to see him die.

But, other than all that, this movie is nothing more than hidden rip-offs.

Fig. 1: Rhetorical structure in a present-day movie review

Example: Using keywords

- Final two paragraphs of a review of John Galsworthy's *The Freelands*, published in *The Athenaeum* (1915)
- Green: positive; red: negative

Sections highlighted by a human
(overall SO: +1)

We must not, however, discuss that aspect of the problem further, but hasten to acknowledge the worth of Mr. Galsworthy's character-drawing. His women are as good as his men, and we cannot single out any one of them for special praise. His editor and journalist help to sweeten callings which have a tendency to embitter men nowadays. His rebels show hardly a trace of the arrogant self-sufficiency which makes that class of person objectionable; and his Philistines only act according to their lights, though they may be credited with a certain amount of wilful blindness. The old lady who insists on putting a good face on everything is wholly delightful.

The author begins in a jerky' style, but happily drops it before the reader has had time to become exasperated.

Sections highlighted by our system
(overall SO: +0.28)

We must not, however, discuss that aspect of the problem further, but hasten to acknowledge the worth of Mr. Galsworthy's character-drawing. His women are as good as his men, and we cannot single out any one of them for special praise. His editor and journalist help to sweeten callings which have a tendency to embitter men nowadays. His rebels show hardly a trace of the arrogant self-sufficiency which makes that class of person objectionable; and his Philistines only act according to their lights, though they may be credited with a certain amount of wilful blindness. The old lady who insists on putting a good face on everything is wholly delightful.

The author begins in a jerky' style, but happily drops it before the reader has had time to become exasperated.

- The system picks up the right sections, but it also includes many other words and phrases that are not central to the point → noise
- To get rid of noise, we need to focus on the rhetorical structure of the text

Example: After discourse parsing

- Existing sentence-based parser (Soricut and Marcu 2003) that extracts the most important parts in a relation (e.g., result in a cause-result relation)
- Run our semantic orientation calculator on rhetorically important parts
  - SO after extracting main parts: 1.04

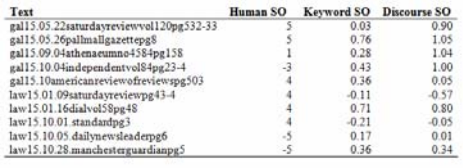
Main parts extracted by the discourse parser (in blue)

We must not, however, discuss that aspect of the problem further, but hasten to acknowledge the worth of Mr. Galsworthy's character-drawing. His women are as good as his men, and we cannot single out any one of them for special praise. His editor and journalist help to sweeten callings which have a tendency to embitter men nowadays. His rebels show hardly a trace of the arrogant self-sufficiency which makes that class of person objectionable; and his Philistines only act according to their lights, though they may be credited with a certain amount of wilful blindness. The old lady who insists on putting a good face on everything is wholly delightful.

The author begins in a jerky' style, but happily drops it before the reader has had time to become exasperated.

Evaluation and results

- Preliminary results based on 10 texts; qualitative evaluation of individual tools
  - Using the discourse parser improves some of the results in the right direction
  - Differences between keyword- and context-based methods are not significant yet



Text	Human SO	Keyword SO	Discourse SO
gal15.05.22saturdayreviewvol120pg532-33	5	0.03	0.90
gal15.05.26pallmallgazettepg8	5	0.76	1.05
gal15.09.04athenaeumno4584pg158	1	0.28	1.04
gal15.10.04independentvol84pg23-4	-3	0.43	1.00
gal15.10americanreviewofreviewspg503	4	0.36	0.05
law15.01.09saturdayreviewpg43-4	4	-0.11	-0.57
law15.01.16dialvol58pg48	4	0.71	0.80
law15.10.01.standardpg3	4	-0.21	-0.05
law15.10.05.dailynewsleaderpg6	-5	0.17	0.01
law15.10.28.manchesterguardianpg5	-5	0.36	0.34

Table 1. Keyword and discourse results for 10 texts

- Next challenge: comparative evaluation
  - How do we validate evaluations of overall semantic orientation?
    - Human annotators assign SO for texts that they read
    - Reliability comparisons with results of automated assignation
  - How do we map SO to reputation?
    - Develop reputation algorithms to produce reputation trajectories with variable weight given to economic and cultural factors

Contribution

- A large body of data about six authors
  - Will be coded in XML and made available
- A set of tools for text analysis, reusable for other tasks
- Parallel project on extracting semantic orientation from present-day movie and book reviews and consumer products

References & Acknowledgements

- Brill, E. (1995) Transformation-based error-driven learning and Natural Language Processing. *Computational Linguistics, 21* (4), 543-565.
- Mann, W.C. & S.A. Thompson (1988) Rhetorical Structure Theory: Toward a functional theory of text organization. *Text, 8* (3), 243-281.
- Schilder, F. (2002) Robust discourse parsing via discourse markers, topicality and position. *Natural Language Engineering, 8* (2/3), 235-255.
- Soricut, R. & D. Marcu (2003) Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of Human Language Technology and North American Association for Computational Linguistics Conference (HLT-NAACL'03)*. Edmonton, Canada.
- Taboada, M. & W.C. Mann (2006) Rhetorical Structure Theory: Looking back and moving ahead. *Discourse Studies, 8* (3), 423-459.
- Tuchman, G. & N.E. Fortin (1989) *Edging Women Out: Victorian Novelists, Publishers, and Social Change*. New Haven: Yale University Press.
- Voll, K. & M. Taboada (2007) Not all words are created equal: Extracting semantic orientation as a function of adjective relevance. In *Proceedings of the 20th Australian Joint Conference on Artificial Intelligence* (pp. 337-346). Gold Coast, Australia.

This project is supported by the Natural Sciences and Engineering Research Council of Canada, and by Simon Fraser University, under a SSHCR Institutional grant. Thanks to Julian Brooke and Kimberly Voll for their work on the semantic orientation calculator.

SFU SIMON FRASER UNIVERSITY
THINKING OF THE WORLD

Contact: Maite Taboada – mtaboada@sfu.ca – http://www.sfu.ca/~mtaboada

SDH/SEMI Conference – University of British Columbia – June 3, 2008

sdh semi