Using Rhetorical Structure Theory for Detection of Fake Online Reviews

Olu Popoola, Aston University, UK.
essien.popoola@gmail.com

Abstract

Fake online book reviews, where authors and 'review factories' secretly pay writers to review products and services, are an increasing concern for consumer protection regulators worldwide. This study uses Rhetorical Structure Theory to analyze a forensic collection of authentic and fake Amazon book reviews drawn from a Deceptive Review corpus to test the potential for the application of discourse coherence analysis to the specific task of developing linguistic heuristics for spotting fake reviews and to the general area of linguistic deception detection. The study introduces a theory of genre violation to explain deception in reviews, highlights the deceptive pragmatics and discourse strategies of paid review writers and confirms the utility of RST in forensic linguistic contexts.

1 Introduction

Consumer protection laws and regulations in most 'free market' jurisdictions prohibit fake online reviews, undisclosed paid-for editorial content and misleading actions and omissions that (may) deceive the average consumer. Agents (i.e. paid writers) as well as businesses can be prosecuted. Since consumer education is key to fraud prevention, regulatory discourse routinely includes warnings and heuristics for detecting different kinds of fraud and deception. Many of these are based on noticing visual language features such as spelling mistakes and overly positive language that makes a product out to be 'the best thing ever' (Competition Bureau Canada, 2015).

The value and utility of these fake review detection heuristics could be improved by a systematic method of incorporating discourse-level features. These may be easier to interpret and more amenable to regulatory heuristics development than stylometric measures (e.g. unigrams and syntax). This study deploys the analysis of discourse coherence relations to unlock linguistic information useful for heuristic development from within the structure, sequence and sections of a text.

Previous uses of RST for deception detection have had mixed results. Rubin et al. (2015) used RST to compare authentic news stories with fictional news stories written as competition entries for a 'Bluff the Listener' radio show. RST relations were found to have limited discriminatory power (63% accuracy), due to the latent influence of humour on linguistic profiles of both truths and lies. Feng (2015) tested an automated RST parser on a corpus of authentic TripAdvisor reviews and deceptive reviews written under experimental conditions. The parser underperformed (50% i.e. at chance level) compared to unigram (87%) and syntax (88%) measures; it was unable to identify a sufficiently diverse set of relations likely due to an absence of explicit discourse markers in the linguistic data which may be typical of product reviews.

This study addresses the limitations of this previous research. A manual RST analysis was conducted on a forensic corpus (i.e. real review data with established ground-truth) of 25 known fake and 25 authentic Amazon book reviews drawn from the Deceptive Review (DeRev) corpus (Fornaciari and Poesio, 2014). Previous deception detection research on this dataset has built machine learning models utilizing stylometric measures with relatively high accuracy levels of 75%-85% (Fornaciari and Poesio, 2014; Hernández-

Proceedings of the 6th Workshop Recent Advances in RST and Related Formalisms, pages 58–63,
Santiago de Compostela, Spain, September 4 2017.

Castaneda et al, 2016); this study hypothesized that those observed stylometric differences reviews would manifest as significant variation in the coherence relational structure of fake compared to authentic reviews and qualitative differences in the pragmatic strategies of fake and true review writers.

2 Method and Data

The DeRev corpus is a collection of 6,819 Amazon book reviews of 68 books written by 4811 different reviewers. This study focused on the 118 'gold standard' fake reviews. Ground-truth for these fake reviews was obtained through following up the journalistic research of David Streitfield, who interviewed review writers that admitted to being paid $10 to $15 dollars per review ('offending writers'), 'offending authors' who admitted paying for bulk reviews (e.g. $999 for 50 reviews) and the owner of a review production factory who had been making over $20000 per month before being exposed (Streitfield, 2012).

Fornaciari and Poesio used Streitfield's investigative journalism to collect known fake reviews by searching Amazon for 1) reviews of books written by 'offending authors', and 2) reviews written by 'offending writers'. From those collected reviews, only those that matched the following set of meta-linguistic deceptive review heuristics were selected: a) be part of a review cluster i.e. one of at least two reviews posted for the same book within 72 hours. b) be written by an author that used a nickname rather than real name, and c): be assigned an 'Unknown' rather than Verified Amazon purchase status. The gold-standard corpus was completed with a matching number of reviews whose authenticity was established by the fact that the books authors were either dead (e.g. Ernest Hemmingway) or highly successful (e.g. Stephen King), making 236 reviews in total.

Manual RST coding was conducted by the author on 50 gold-standard reviews (25 true, 25 fake) all between 50 and 250 words in length (see Figure 1 above). Controlling for length minimized the effect of this variable on predicting deception with RST; this length was chosen as convenient and sufficient for manual RST coding.

DeRev-RST Corpus (Popoola, 2016)

	All reviews (50)	True (25)	Fake (25)
No. Words	4931	2222	2709
Average number of words per review / stdev	98.6 / 40.7	88.9 / 43.2	108.4 / 36.3
No. of RST coherence relations annotated	490	239	251
Average number of relations per review / stdev	9.8/5.0	9.6 / 5.6	10.0 / 4.4
Average 'words per relation' / stdev	10.7 / 2.6	10.0 / 2.9	11.3 / 2.1

Figure 1: DeRev-RST corpus statistics.

Macro-relation	RST relations	Macro-relation	RST relations
ATTRIBUTE	Attribute	EVALUATION	Comment
BACKGROUND	Background		Conclusion
	Circumstance		Evaluation
CAUSE	Cause	EXPLANATION	Justify
	Consequence		Reason
	Result		Evidence
COMPARISON	Analogy	JOINT	Disjunction
	Comparison		Joint
	Preference		List
CONTRAST	Antithesis	MANNER-MEANS	Manner
	Concession		Means
	Contrast	SUMMARY	Summary
ELABORATION	Elaboration		Restatement
ENABLEMENT	Enablement	TEMPORAL	Sequence
	Purpose		Temporal

Figure 2: RST macro-relations used and their definitions.

Carlson and Marcu's (2001) extended set of RST relations was used for initial coding but only the macro-relations (summary groupings of relations; see Figure 2 above) were used in the predictive analysis model to minimize the impact of ambiguous relations on coding consistency. Additionally, an external party collated the 50 reviews according to the sample specification (and renamed the files) so that the author could code the reviews blind to truth value.

3 Results

In the analysis of the corpus, the fake reviews have more *Elaboration, Joint* and *Background* macro-relations; the true reviews have more *Evaluation*, *Contrast and Explanation* macro-relations. Only True reviews contain *Comparison* relations.

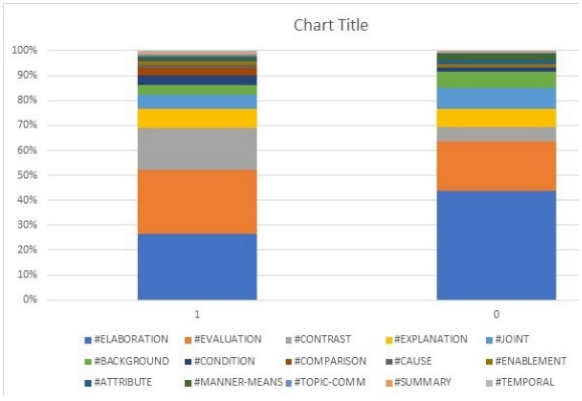

Figure 3: Comparative frequency of RST macro-relations. 1=True reviews (239 relations); 0=Fake reviews (251 relations)

The boxplots below (Figures 4 and 5) suggest that the use of *Elaboration* relation distinguishes true from fake reviews discourse. Although overall use of *Evaluation* macro-relations does not substantially differ between true and fake reviews, the relative proportion of *Evaluation* vs. *Elaboration* is much lower in deceptive reviews.

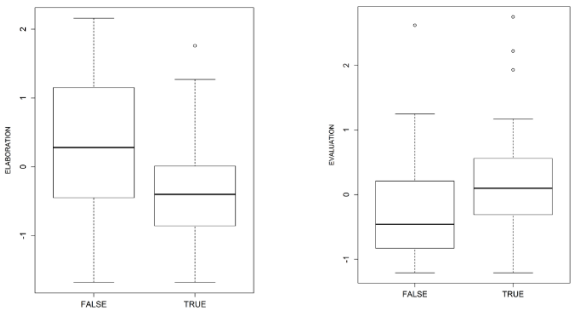

Figure 4: Boxplot comparison of Elaboration and Evaluation macro-relation frequencies in fake and true reviews.

The range of relation frequencies indicate significant effects for *Contrast* relations as a feature of authentic reviews. Specifically, 14 *Contrast_Concession* relations were only found in the true sample. 31 out of 37 *Contrast_Antithesis* relations were found in the true sample. Both authentic and deceptive reviews contain *Background* relations, although fake reviews use them more frequently. A logistic regression model that fit all 12 macro-relations (*R square = 0.68)* indicates that the differences for *Elaboration and Contrast* are significant (Figure 5a.)

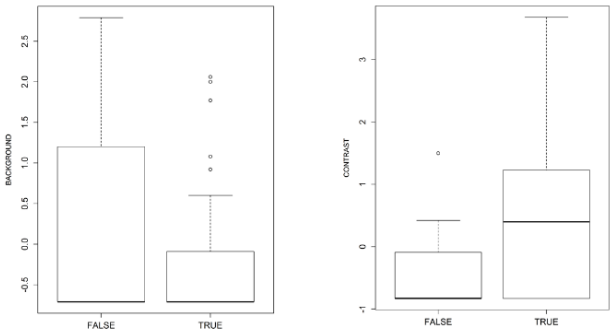

Figure 5: Boxplot comparison of Contrast and Background macro-relation frequencies in fake and true reviews

LOGISTIC REGRESSION

Relations (Total)	P score	Exp(B)
Contrast (86)	.02	.11
Elaboration (263)	.05	3.79
Background (38)	.17	2.08
Joint (54)	.23	2.32
Explanation (56)	.60	1.44
Evaluation (174)	.67	.77

Hosmer+Lemshow = 0.77; Nagelkerke R Square=0.68]

Figure 5a: Logistic regression results for six most frequent relations in DeRev-RST corpus.

4 Discussion

4.1 Elaboration

While the high frequency of *Elaboration* relations is generally to be expected in RST analysis, the fact that paid-for reviews use significantly more *Elaboration* relations than authentic ones reflects the deceptive context of communication. In fake reviews, there is more synopsis and description of topics; the plot elaboration in Figure 6 takes up half of the total review. This is likely due to paid review writers, who at most only superficially read the books they are reviewing, using information that is easily gleaned from book PR materials e.g. back cover synopsis.

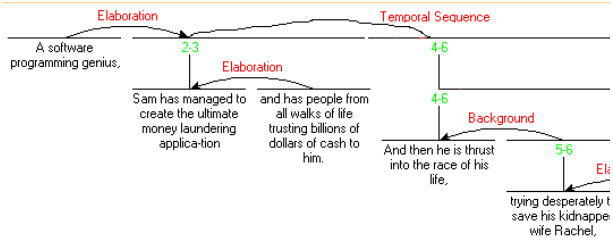

Figure 6: Plot elaboration in a fake review

Being paid £5 to £10 per review means that for the activity to be profitable, time must be spent on writing multiple reviews rather than reading many books. This inevitably affects the quality of evaluation and appraisal of the books.

4.2 Evaluation

While the frequency of *Evaluation* relations does not clearly discriminate between fake and authentic reviews, a lower proportion of evaluative text is a feature of the deceptive reviews; where true reviews have on average equal amounts *Elaboration and Evaluation*, fake reviews have a 2:1 ration (see Figure 4 above). Paid writers often use generic appraisal, simply adding phrases such as *"...a must read..."* or "*I would recommend…"* to the end of a descriptive review. In contrast, *Evaluation* in the genuine reviews is longer and more subjective i.e. explaining why the reviewer liked the book rather than why the reader would like the book.

FAKE: This is a must read for anyone considering taking the Hobet examination and is looking for a sure-fire way to succeed.

TRUE: This book made me think and made me remember that it is okay to dream. Who can argue with that?

Figure 7: Comparative examples of *Evaluation*

4.3 Contrast

A significant feature of authentic reviews was the use of *Contrast* relations with an evaluative function. The true reviews are far more likely to mention potentially negative aspects of a book in the context of an overall positive appraisal; *Contrast* relations (which include *Concession* and *Antithesis*) are the discourse mechanism for this (e.g. Fig 8 below).

This strategy of expressing 'caveats' has been noted as a feature of negative English language movie reviews (Taboada et al, 2014). Hedged positive evaluation has also been found to feature in Japanese academic book reviews (Itakura, 2013). Mitigated evaluation is a feature of the review genre (at least in certain languages/cultures).

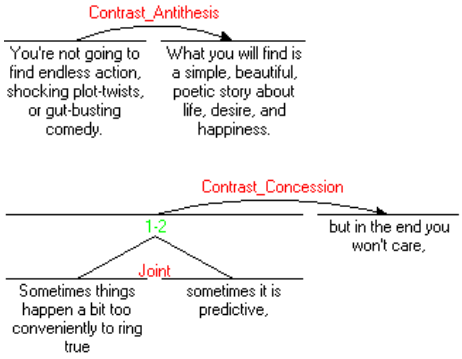

Figure 8: Examples of *Contrast* relations in true reviews.

This sets up the hypothesis that deceptive reviews are a *genre (*or *register) violation*. The situational context of the deception – individuals producing multiple reviews, under time constraints that prohibit proper reading, to maximize income – impacts on the pragmatic and discourse strategies of paid writers and affect the language choices made. Under these conditions providing the nuanced opinion typical of the review genre is both challenging and inefficient.

4.4 *Background*

The *Background* relation did not show a significant effect (see Figures 5 and 5a above) but its use in fake reviews present examples of *deceptive pragmatics.* Deceptive use of *Background* relations deploys persuasion to affect reader perceptions of the review rather than of the book – as if the reader needs convincing of the veracity of the review. One example,
in Figure 9 below, has a *Background* relation with a *Reason* relation in the satellite presenting a motivation for purchase.

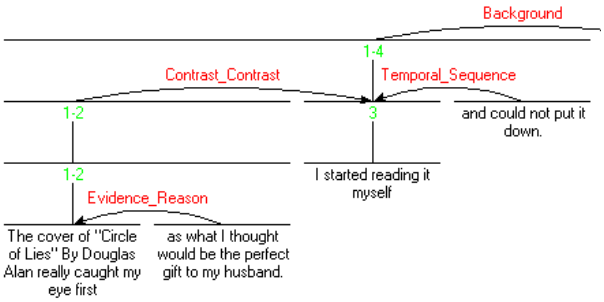

Figure 9: Examples of *Background* in a fake review.

4.5 *Nuclearity*

The qualitative content and location of the "most nuclear" discourse unit (Stede, 2008) is a predictor of deception in these reviews. Mann and Thompson's (1987) deletion test and Marcu's (2000) Strong Nuclearity Hypothesis were used to locate the 'nucleus discourse unit'(NDU) for each review.

	Fake (25)	True (25)
NDU in 1st sentence of review	17	9
NDU mentions *Title*	18	3
NDU mentions *Author*	8	5
NDU describes content/plot	8	4
NDU contains appraisal/evaluation	18	22

Figure 10: Comparative analysis of NDUs

Figure 10 illustrates marked differences in the NDUs of the fake and real reviews. The fake review NDUs were mainly located in the opening sentence, typically mentioned the book title and often provided author name with a brief plot/content description (e.g. Fig 11 below). Authentic review NDUs contained a key evaluation/opinion of the book without (or with minimal) content or plot description and were more likely to occur within the body of the review (e.g. Fig 12 below).

This unexpected finding suggests that techniques for identifying salient discourse such as automatic summarization may be useful for computer-aided deception detection and further supports the use of RST and related formalisms in the development of a linguistic theory of deception.

5. Conclusion

This pilot study has revealed that paid review writers deploy deceptive pragmatics i.e. a coherent set of linguistic strategies deployed to support the intent to deceive. Deceptive reviews contain violations of genre conventions related to evaluation, and contamination from related genres such as synopsis or press release. RST analysis has provided rich qualitative data for the generation of a set of regulatory heuristics that might include consumer warnings such as: 1) fake reviews are *more* likely to mention book titles, authors and give details of a book's contents; 2) fake 5-star reviews tend to be all positive, whereas genuine 5-star reviews usually contain caveats. Future research will address the challenge of replicating RST analysis on big linguistic data sets by identifying relations signals to assist automated analysis, testing the potential of 'textual coherence ratios' such as *Elaboration/Evaluation* as explanatory 'discourse metrics' and investigating whether models of discourse salience and summarization tool can be used in deception detection.

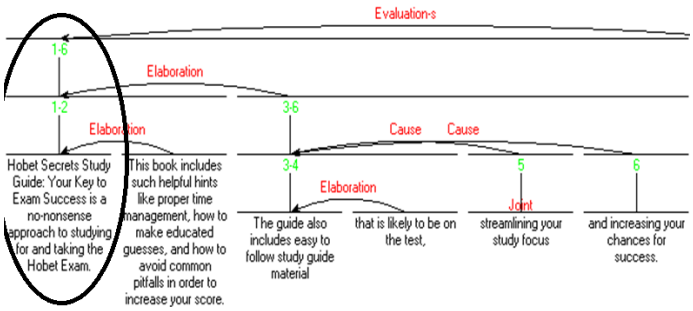

Figure 11: Fake review NDU located in opening sentence.

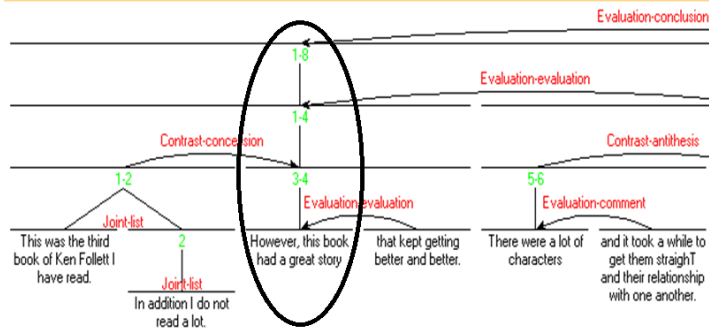

Figure 12: True review NDU located in body of text.

References

Carlson, Lynn., and Marcu, Daniel (2001) *Discourse tagging reference manual.* ISI Technical Report ISI-TR-545, 54, 56

Competition Bureau Canada (2015) *Don't buy into fake online endorsements.* Retrieved from http://www.competitionbureau.gc.ca/eic/site/cb-bc.nsf/eng/03782.html

Feng, Vanessa W. (2015) *RST-style discourse parsing and its applications in discourse analysis.* Doctoral Dissertation, University of Toronto.

Fornaciari, Tommaso, Poesio, Massimo. (2014). *Identifying fake Amazon reviews as learning from crowds.* Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 279–287

Itakura, Hiroko. (2013). *Hedging praise in English and Japanese book reviews.* Journal of Pragmatics, 45(1), 131-148.

Mann, William C., and Christian MIM Matthiessen. "Functions of language in two frameworks." *Word* 42, no. 3 (1991): 231-249.

Mann, William C., and Sandra A. Thompson. *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute, 1987.

Marcu, Daniel. *The theory and practice of discourse parsing and summarization*. MIT press, 2000.

Rubin, Victoria L., Conroy, Niall J., and Chen, Yimin. (2015) *Towards news Verification: Deception detection methods for news discourse.* In Proceedings of HICSS48 Symposium on Rapid Screening Technologies, Deception Detection and Credibility Assessment Symposium

Stede, Manfred. "RST revisited: Disentangling nuclearity." *Subordination 'versus' Coordination 'in Sentence and Text* (2008): 33-59.

Streitfield, David (2012) *The best book reviews money can buy.* Retrieved from http://www.nytimes.com/2012/08/26/business/book-reviewers-for-hire-meet-a-demand-for-online-raves.html

Taboada, Maite, Carretero, Marta and Hinnell, Jennifer (2014) *Loving and hating the movies in English, German and Spanish.* Languages in Contrast, 14(1), 127-161.